UNIVERSITÁ DEGLI STUDI DI VERONA

DEPARTMENT OF BIOTECHNOLOGY

GRADUATE SCHOOL OF NATURAL AND ENGINEERING SCIENCES

DOCTORAL PROGRAM IN BIOTECHNOLOGY

WITH THE FINANCIAL CONTRIBUTION OF UNIVERSITÁ DEGLI STUDI DI VERONA

CYCLE XXXV

TITLE OF THE DOCTORAL THESIS

Benchmarking of differential abundance methods and development of bioinformatics and statistical tools for metagenomics data analysis

S.S.D. BIO/11

Coordinator:	Professor Matteo Ballottari
Tutor:	Professor Nicola Vitulo

Doctoral Student: Doctor Matteo Calgaro

Benchmarking of differential abundance methods and development of bioinformatics and statistical tools for metagenomics data analysis - Matteo Calgaro
Doctoral Thesis
Verona, 03/2023
ISBN: 9788869251962

No book can ever be truly accomplished.

As we work around it, we learn enough to find it immature

by the time we step away from it.

-Karl Raimund Popper

Ringraziamenti

Completare questo dottorato di ricerca è stata una delle esperienze più impegnative, ma gratificanti della mia vita. Senza il sostegno, la guida e l'incoraggiamento di così tante persone, non sarei arrivato a questo punto.

In primo luogo, desidero esprimere la mia profonda gratitudine al mio supervisore, il Prof. Nicola Vitulo. La tua competenza, la tua guida e il tuo sostegno sono stati preziosi lungo questo percorso. Il tuo esempio sarà sempre una fonte di ispirazione per me.

Inoltre, desidero ringraziare tutti i co-autori con cui ho avuto la possibilità di collaborare in questi anni. Tra loro, un ringraziamento speciale va alla Prof.ssa Chiara Romualdi e al Prof. Davide Risso per i preziosi feedback e le critiche costruttive che mi hanno aiutato a migliorare il mio lavoro. Sono grato anche alla Prof. Giovanna Felis e alla Dott.ssa Sonia Facchin. Avete mostrato pazienza, curiosità e interesse per il mio lavoro. La passione per la ricerca di tutti voi mi ha ispirato e motivato a fare sempre del mio meglio.

Esprimo la mia gratitudine anche alla Prof.ssa Silvia Storti per il corso in cui ho avuto l'opportunità di essere tutor. Grazie per la tua guida e il tuo sostegno.

Ai miei compagni di dottorato, colleghi e amici. Grazie per aver creato una comunità di supporto quando ne avevo bisogno.

La mia sincera gratitudine va anche alla mia famiglia, in particolare ai miei genitori e alla famiglia del mio partner. La vostra fiducia incondizionata nelle mie capacità mi ha dato la carica anche quando le cose erano difficili.

Infine, al mio partner Enrico, grazie per aver fatto il tifo per me, per essere stato la mia roccia e la mia fonte di equilibrio durante questa corsa sulle montagne russe. La tua fiducia in me, le tue parole di incoraggiamento e la tua comprensione durante le mie ore di lavoro hanno significato tanto. Ti amo.

A tutti coloro che hanno contribuito al mio percorso, in modi grandi e piccoli, estendo la mia più profonda gratitudine e apprezzamento. Grazie.

Matteo Calgaro, Padova, 03/2023

Acknowledgements

Completing this PhD has been one of the most challenging, yet rewarding experiences of my life. Without the support, guidance, and encouragement of so many people, I would not have made it to this point.

First and foremost, I want to express my deepest gratitude to my advisor, Prof. Nicola Vitulo. Your expertise, mentorship, and support have been invaluable to me throughout this journey. Your example will always be a source of inspiration for me.

I would also like to thank all the co-authors I had the chance to collaborate with during these years. Among them, a special thanks goes to Prof. Chiara Romualdi and Prof. Davide Risso for the insightful feedback and constructive criticism that helped me improve my work. I'm grateful to Prof. Giovanna Felis and Dr. Sonia Facchin. You have shown patience, curiosity, and interest in my work. The passion for research, of you all, inspired and motivated me to always do my best.

I express my gratitude towards Prof. Silvia Storti for the course I had the opportunity to tutor. Thanks for your guidance and support.

To my fellow graduate students, coworkers, and friends. Thank you for providing a supportive community when I needed it.

My heartfelt gratitude goes also to my family, especially my parents, and my partner's family. Your unconditional belief in my abilities have given me the charge even when things were challenging.

Finally, to my partner Enrico, thank you for being my rock, my cheerleader, and my source of sanity during this rollercoaster ride. Your belief in me, your words of encouragement, and your understanding during my hours of work have meant the world to me. I love you.

To all who have contributed to my journey, in big and small ways, I extend my deepest gratitude and appreciation. Thank you.

Matteo Calgaro, Padova, 03/2023

Sommario

L'analisi di dati nell'ambito del microbioma e della metagenomica è stato il tema principale del mio dottorato. L'obiettivo primario di questa tesi si muove attorno all'osservazione dei limiti dei metodi per lo studio dell'abbondanza differenziale e culmina con la creazione di un framework analitico che permette la loro misurazione e comparazione. Come obiettivo secondario, inoltre, la tesi vuole enfatizzare la necessità di una solida analisi statistica esplorativa ed inferenziale nei dati di metabarcoding, tramite la presentazione di alcuni casi studio.

Inizio presentando 2 studi strettamente collegati in cui i metodi per l'analisi di abbondanza differenziale sono i protagonisti. L'analisi di abbondanza differenziale è lo strumento principale per individuare differenze nelle composizioni delle comunità microbiche in gruppi di campioni di diversa provenienza. Rappresenta quindi il primo passo per la comprensione delle comunità microbiche, delle relazioni tra i loro membri e di questi con l'ambiente. Il primo studio riguarda un lavoro di confronto tra metodi. A partire da una collezione di dataset metagenomici, l'obiettivo era di valutare le performance di metodi per l'analisi dell'abbondanza differenziale, anche nati in altri ambiti di ricerca (*e.g.*, RNA-Seq e single-cell RNA-Seq). Invece, con il secondo studio presento un software che ho sviluppato grazie ai risultati ottenuti dalla precedente ricerca. Attualmente, il pacchetto software, in linguaggio R, è disponibile su Bioconductor (*i.e.*, una piattaforma open-source per l'analisi e la visualizzazione di dati biologici). Esso consente agli utenti di replicare sui propri dataset il confronto tra metodi per lo studio dell'abbondanza differenziale e la conseguente analisi delle performance.

Infine, mostro alcune delle sfide che ho incontrato nell'analisi di questo tipo di dato attraverso 2 casi studio riguardanti il microbioma umano, la sua composizione e dinamica, sia in stato di salute che malattia. Nel primo studio, dei soggetti sani sono stati trattati con una mistura di probiotici per valutare variazioni del microbiota intestinale ed eventuali associazioni con alcuni aspetti psicologici. Un'attenta analisi esplorativa, l'impiego di tecniche di clustering e l'utilizzo di modelli di regressione lineare ad effetti misti hanno consentito di svelare un forte effetto soggetto-specifico e la presenza di diversi batteriotipi di partenza che mascheravano l'effetto complessivo del trattamento probiotico. Invece, nel secondo studio mostro come, a partire da campioni salivari, sono stati individuati dei biomarcatori associati all'esofagite eosinofila (*i.e.*, una malattia cronica immuno-mediata a carico dell'esofago che causa disfagia, occlusioni e stenosi esofagee). Nonostante la bassa numerosità campionaria è stato possibile costruire un modello per discriminare tra casi e controlli con una buona accuratezza. Anche se ancora prematuro, questo risultato rappresenta un passo promettente verso la diagnosi non invasiva di questa malattia che per il momento viene fatta solo tramite biopsia esofagea.

Keywords: Microbioma, Analisi di Abbondanza Differenziale, Confronto tra metodi, Analisi Esplorativa, Analisi dei Dati, Probiotici, Esofagite Eosinofila.

Abstract

Microbiome and metagenomics data analysis has been the main theme of my PhD programme. As a main goal, the thesis moves from the observed limitations of the differential abundance analysis tools to a benchmark and a framework against which they could be measured and compared. Furthermore, as a secondary goal, the presentation of some case studies wants to emphasise the need for a sound exploratory and inferential statistical analysis in metabarcoding data.

Firstly, I present two closely related studies in which differential abundance analysis methods play the main role. The differential abundance analysis is the principal approach to detect differences in microbial community compositions between different sample groups, and hence, for understanding microbial community structures and the relationships between microbial compositions and the environment. I start by introducing a benchmarking study in which differential abundance analysis methods, even from different domains (*e.g.*, RNA-Seq and single-cell RNA-Seq), were used in a collection of microbiome datasets to evaluate their performance. Then, I continue with the presentation of software package that I developed from the results obtained in the previous research. The software package, in R language, is currently available on Bioconductor (*i.e.*, an open-source software platform for analysing and visualising biological data). It allows users to replicate the benchmarking of differential abundance analysis methods and evalute their performances on their own datasets.

Secondly, I highlight the microbiome data analysis challenges presenting two case studies about the human microbiome and its composition and dynamics in both disease and healthy states. In the first study, healthy volunteers were treated with a probiotic mixture and the changes in the gut microbiome were studied in conjunction with some psychological aspects. A careful data exploration, clustering, and mixed-effects regression models, unveiled subject-specific effects and the presence of different bacteriotypes which masked the probiotic effect. Instead, in the second study

I show how to identify disease-related microbial biomarkers for eosinophilic oesophagitis (*i.e.*, a chronic immune-mediated inflammatory disease of the oesophagus that causes dysphagia, food impaction of the oesophagus, and esophageal strictures) from saliva. Despite the low sample size it was possible to train a model to discriminate between case and control states with a decent accuracy. While still premature, this represents a promising step for the non-invasive diagnosis of eosinophilic oesophagitis which is now possible only through esophageal biopsy.

Keywords: Microbiome, Differential Abundance Analysis, Benchmarking, Data Exploration, Data Analysis, Probiotics, Eosinophilic Oesophagitis.

Contents

List of Figures

List of Tables

Acronyms

ALR Additive Log-Ratio
AR AutoRegressive
ASV Amplicon Sequence Variants

BH Benjamini-Hochberg
BMC Between Methods Concordance

CAT Concordance At the Top
CCA Canonical Correspondence Analysis
CE Classification Error
CLR Centred Log-Ratio
CPM Counts Per Million
CSS Cumulative Sum Scaling
CV Cross Validation

DA Differentially Abundant/Differential Abundance
DAA Differential Abundance Analysis
DCA Detrended Correspondence Analysis
DM Dirichlet-Multinomial
DNA DeoxyriboNucleic Acid

EoE Eosinophilic oEsophagitis
EoEHSS Eosinophilic oEsophagitis Histologic Scoring System
eos/HPF eosinophils per High-Power Field
EREFS oEdema, Rings, Exudates, Furrows, and Strictures

FP False Positive
FPR False Positive Rate

GI Gastro-Intestinal
GMPR Geometric Mean of Pairwise Ratios
GOF Goodness Of Fit

HMP Human Microbiome Project
HTS High Throughput Sequencing

IBS Irritable Bowel Syndrome
IQR InterQuartile Range
ITS Internal Transciber Spacer

KS Kolmogorov-Smirnov

LRT Likelihood Ratio Test

MD Mean-Difference

MDD Major Depressive Disorder

MDS Multi-Dimensional Scaling

mRNA messenger RNA

NB Negative Binomial

NGS Next Generation Sequencing

NMDS Non-metric MultiDimensional Scaling

OGD OesophagoGastroDuodenoscopy

OTU Operational Taxonomic Unit

PCA Principal Component Analysis

PCoA Principal Coordinates Analysis

PCR Polymerase Chain Reaction

PERMANOVA PERMutational ANalysis Of VAriance

PLS Partial Least Squares

PLS-DA Partial Least Squares Discriminant Analysis

POMS Profile Of Mood State

PPI Proton Pump Inhibitor

PSQI Pittsburgh Sleep Quality Index

RDA ReDundancy Analysis

RF Random Forests

RLE Relative Log Expression

RMSE Root Mean Squared Errors

RNA RiboNucleic Acid

RNA-Seq RNA Sequencing

rRNA ribosomal RNA

RSID Random Sample IDentifier

SCFA Short Chain Fatty Acid

scRNA-Seq single cell RNA-Seq

sPLS-DA sparse Partial Least Squares Discriminant Analysis

SS-ANOVA Smoothing Splines ANanalysis Of VAriance

SV Sequence Variant

SVM Support Vector Machines

TMM Trimmed Mean of M-values

TP True Positive

TSS Total Sum Scaling

WMC Within Method Concordance

WMS Whole Metagenome shotgun Sequencing

ZIG Zero-Inflated Gaussian

ZINB Zero-Inflated Negative Binomial

ZPD Zero Probability-Difference

Chapter 1

General Introduction

1.1 The Microbiome

1.1.1 Definition

According to the 1988 definition given by Whipps *et. al* [1], the microbiome may be defined as a characteristic microbial community occupying a reasonably well-defined habitat which has distinct physio-chemical properties. The term thus not only refers to the microorganisms involved but also encompasses their theatre of activity.

In 2019, leading microbiome researchers from academic, governmental, and industry groups representing diverse areas of expertise, considered the definition still valid and extended it by adding two explanatory sentences to distinguish the terms microbiome and microbiota and emphasise its dynamic character [2]. The living microorganisms populating the microbiome (Prokaryotes [Bacteria, Archaea], Eukaryotes [*e.g.*, Protozoa, Fungi, and Algae]) compose the microbiota, while their "theatre of activity" includes microbial structures, metabolites, mobile genetic elements (*e.g.*, transposons, phages, and viruses), and relic DNA (extracellular DNA derived from dead cells) embedded in the environmental conditions of the habitat. The microbiome, which forms a dynamic and interactive micro-ecosystem prone to change in time and scale, is integrated in macro-ecosystems including eukaryotic hosts, and is crucial for their functioning and health (Fig. 1.1).

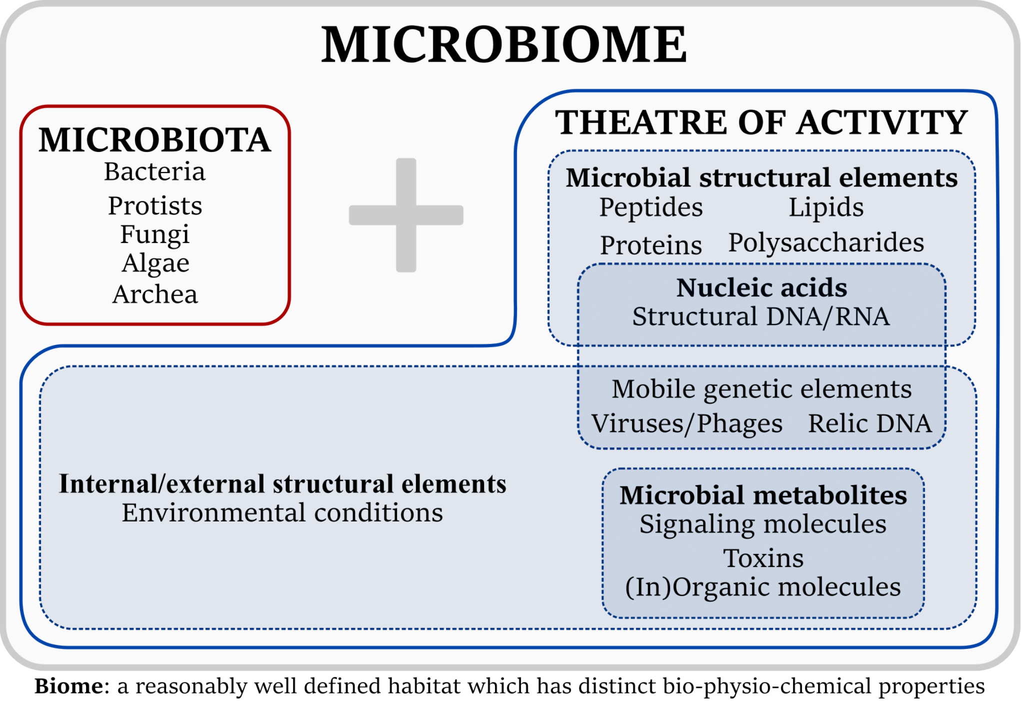


Figure 1.1: A schematic, rearranged from [2], highlighting the composition of the term microbiome containing both the microbiota (community of microorganisms) and their "theatre of activity" (structural elements, metabolites/signal molecules, and the surrounding environmental conditions).

1.1.2 Microbiome research

Microbiome research started back in the seventeenth century originating from microbiology. Progress in this field has often been driven by the development of new equipment, techniques, and technological inventions. Starting from microscopy and cultivation based approaches, passing through electron and scanning transmission microscopy and the discovery of the DNA, to date we have sequencing technologies, PCR, and cloning techniques that enable the investigation of microbial communities using cultivation independent, DNA and RNA-based approaches. To this regard, the advent of the Next Generation Sequencing (NGS) technologies coupled with bioinformatics development, reduced the underlying costs associated with different methods and strategies for sequencing genomes. As a result, the scientific community could enlarge the scope and scale of almost all genomics research projects. Over the past few decades several large-scale projects and initiatives such as the Human Microbiome Project [3], the Earth Microbiome Project [4], and many others [5–7], began to investigate the microbes that inhabit the human body, soils, oceans, and elsewhere.

The main scope of microbiome research is certainly the improvement of health for humans, animals, plants, and the ecosystem as a whole. It is not a coincidence that microbes are the predominant and the first form of life on the planet. Their ability to inhabit hostile environments incompatible with most forms of life demonstrates a spectrum of evolutionary, functional and metabolic diversity that vastly exceeds that of all other organisms in the tree of life [8]. Moreover, they cover the surfaces of all other organisms (occupying internal and even intracellular niches) and influence diverse physiological activities of their hosts, including nutrition, health–disease status and hence well-being [8]. Apparently, microbes provide ecosystems' services that are crucial to local and global sustainability, whether we are talking about a human body, a plant, a cultivated field, a farming facility, a food industry, or a wastewater treatment system.

Just to cite a glaring example, studies suggest that the microbiome of a newborn is widely stimulated when first exposed to microorganisms during neonatal life and the type of delivery plays a role in his/her immune system maturation [9–11], microbiome research in this field shows the need for the development of strategies for minimising or limiting the impact of caesarean on the microbiome development, favouring future health [9]. Other emerging applications of microbiome research in human health are presented by Cullen and colleagues [12]. Some of them are related to diet and its effect on gut microbiome composition and function. Indeed, gut microbiota interactions are related to alteration of immune response, susceptibility to or protection against inflammatory diseases such as irritable bowel syndrome, irritable bowel disease, and colorectal cancer [13–15]. Diet itself, but also probiotics and prebiotics supplementation can be used as an interventional tool to prevent or ameliorate the symptoms of a growing list of neurological disorders including autism, schizophrenia, Parkinson's disease, multiple sclerosis, bipolar mood disorders, anxiety, and depression which are associated with the gut-brain axis [16] (*i.e.*, the bidirectional communication network that links the enteric and central nervous systems related to the neurologic, endocrine, humoral/metabolic, and immune pathways [17]). In addition, microbiome can also be used as a possible diagnosis/prognosis tool for a wide range of pathologies, *e.g.*, irritable bowel disease, progression of diabetes, and others [18, 19]. An example, also detailed in Chapter 6, is related to the diagnostic power of microbiome in Eosinophilic oEsophagitis (EoE), a chronic immune-mediated inflammatory disease of the oesophagus that causes dysphagia, food impaction of the oesophagus, and oesophageal strictures [20]. Diagnosis is possible through oesophageal biopsy but salivary microbiome analysis in combination with machine learning approaches could become a valid, cheap, non-invasive test to segregate between EoE and non-EoE patients [21]. Many other success stories based on microbiome research in the fields of plant health, feed products and livestock health, food production and human health are presented by Rocìo and colleagues [22]. Among these, we find the boosting of sustainable crop productivity through nitrogen-fixing microorganisms inoculation in soybean seeds [23], the reduction of antibiotics use in livestock by improving the animal gut microbiome through prebiotics and feed additives [24], the identification of sources and microbial transmission routes for the improvement of food security and hygiene through the study of microbiome composition and distribution in a food-processing plant [25].

In summary, microbiome research has the potential to contribute substantially on many levels to global efforts to achieve sustainability [8]. Nevertheless, microbiome research continues to be prevalently performed one ecosystem at-a-time, leading to fragmentation of the landscape. Such fragmentation shadows new biological concepts to be discovered: patterns in microbiome interaction, diversity of functions and roles may not be seen [26]. As a consequence, international efforts are oriented towards a systems approach needed to connect research between scientific fields to create a holistic understanding on how microbiomes can be modulated for desirable functions.

1.1.3 Microbiome data

As already mentioned, cultivation independent approaches to assess microbial communities and their metagenomes were enabled by technological advancements. From a biological sample, DNA, RNA, small molecules, proteins, and other information can be extracted as summarised in Fig. 1.2 from the work of Weinstock [27].

To better introduce the terminology, the term "metagenomics" was born in 1998 by Handelsman et. al [28], and it refers to the study of the theoretical collection of all genomes from members in a microbial community from a specific environment. A decade later, according to Gilbert and Dupont [7], metagenomics was most appropriately divided into two research areas driven by technological application: single-gene surveys on one side, and random shotgun studies of all environmental genes on the other. The first can be seen as a directed, focused metagenomic study. Briefly, single targets are amplified using Polymerase Chain Reaction (PCR), and then the products are sequenced, providing an analysis of the range of different orthologs for that target within a given community. This approach is also called metabarcoding as it relies on the use of several taxonomically informative amplicon barcodes such as the 16S (*e.g.*, one or more hypervariable regions of the bacterial small subunit of the 16S ribosomal RNA gene, used for prokaryotic DNA), 18S (for eukaryotic DNA), and Internal Transciber Spacer (ITS) (*i.e.*, non coding DNA between genes). In order to investigate biodiversity paired reads are aggregated into sequences. They were usually organised into Operational Taxonomic Units (OTUs), *i.e.*, clusters of sequencing reads that differed by less than a fixed dissimilarity threshold, and more recently, into Amplicon Sequence Variantss (ASVs) with a single nucleotide resolution, thanks to new methods that control errors based on the quality of the sequencing run [29, 30]. Metabarcoding data can be used to take a community census and create tables of taxa abundances, compute ecological metrics, perform competition and symbiosis analysis, assess microbial differential abundance between groups of samples, and so forth. In the second approach, Whole Metagenome shotgun Sequencing (WMS), total DNA is isolated from a sample and then sequenced resulting in a profile of all genes within the community. WMS data can be used to perform genome assembly and gene predictions, identify gene variants, study population genetics, build pathways, and reconstruct the capabilities of a community.

Nowadays, other authors prefer to divide metagenomics according to the research aspect it pursues [31]. On one hand, a structural approach to study the structure of the uncultivated microbial population and the reconstruction of the complex metabolic network established between community members [32, 33]. On the other hand, a functional approach to identify genes that code for a function of interest [34, 35].

Given the latter division, 16S rRNA gene surveys are excluded from the metagenomics definition. Indeed, in 16S rRNA gene analysis, the study is focused on a single gene (often a portion of it) used as a taxonomic marker. Nevertheless, targeted sequencing is a cheap and the most common method for profiling bacterial communities. Moreover, it is also possible to partly overcome the limited functional and genetic information. Some recently developed bioinformatics tools infers the genes and functional capabilities of the community by leveraging the genome sequences of known microorganisms in databases [27, 36].

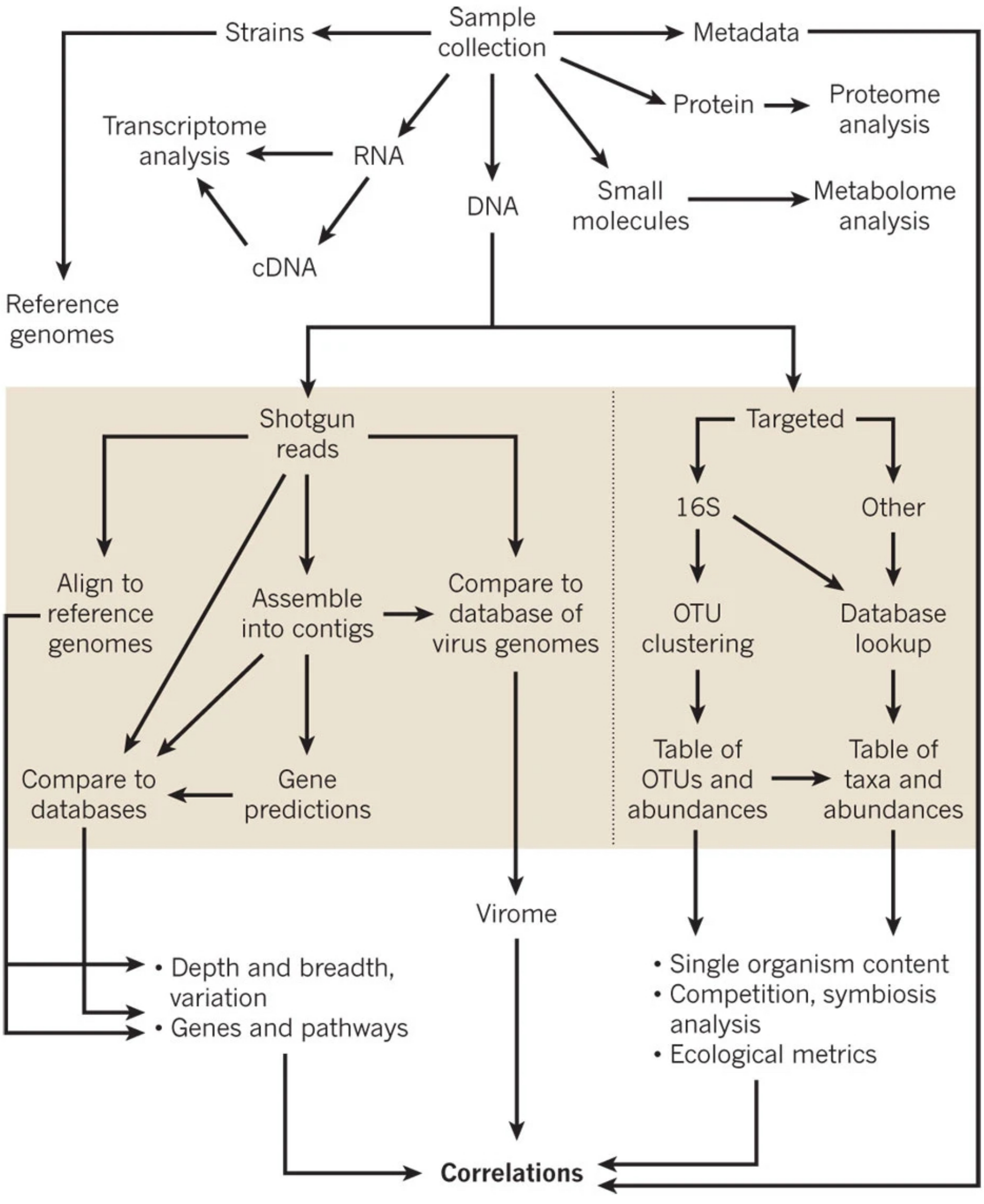


Figure 1.2: Derived from the work of Weinstock [27], this graph depicts the analysis potential of a microbiome sample. The highlighted section represents the two main analysis approaches: the Whole Metagenome shotgun Sequencing (WMS) and the targeted gene survey. The combination of results from both approaches with the annotated sample metadata allows correlating microbial information with phenotypes.

1.2 Microbiome data analysis

1.2.1 Experimental design

Data analysis of a microbiome research should start before the research itself to help in the design of it. Decades of experience and insights on microbiome research produced a long list of best practices for carrying out a microbiome study correctly [37]. Indeed, whether we are considering High Throughput Sequencing (HTS) of DNA (metagenomics), RNA (metatranscriptomics), analysis of secreted bioactive compounds (metabolomics), or the analysis of specific marker genes (metabarcoding), microbiome data are often noisy, sparse, compositional, and high-dimensional.

Consequently, a good thought out experimental design is mandatory to facilitate the analysis and obtain accurate and meaningful results as extensively explained by Knight *et al.* [37] and summarised in Fig. 1.3. For example, cross-sectional studies are useful for finding differences in microbial communities between different groups of samples. However, stratification by potential confounders such as age, sex, diet, lifestyle factors, medications, or by sampling depth, atmospheric agents exposure, soil cultivation, depending on the sample type (*e.g.*, human or environmental), is crucial to unmask spurious associations during the analysis. Differently, longitudinal studies are well suited to control for confounders and permit to assess the microbial community stability over time. However, one of the many things to take into account when designing a longitudinal study is the correct choice of the sampling times to maximise information and minimise the costs.

Other aspects to consider, especially when biotechnologies are involved, are the technical factors and sample processing standardisation [37]. Sources of variability could be introduced in every step of the process, these includes: kit reagents, primers, sample storage, and other factors.

Finally, a complete and clean sample metadata curation and collection is crucial for a faster data analysis and the consequent data interpretation. Even with all these best practices in mind, the analysis of NGS data remains challenging due to both the overall complexity of microbial communities and the intrinsic characteristics of sequencing-generated data.

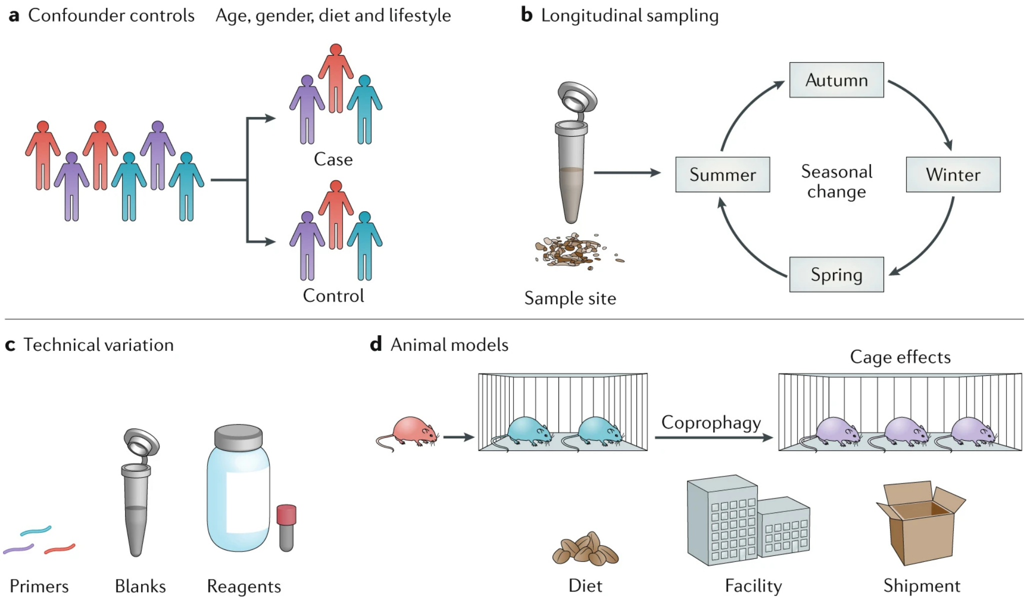


Figure 1.3: From the work of Knight et al. [37] this figure summarises some of the factors to take into account when conducting a microbiome experiment. **a** Stratification by confounding variables in case-control studies. **b** Sample site and collection timing for longitudinal studies. **c** Sources of technical variation during the sample processing. **d** Diet, facility, shipment, cage effect and coprophagy on animal studies.

1.2.2 Heterogeneity, over-dispersion, compositionality, intra-dependency, and sparsity in microbiome data

Different people can differ greatly from one another in terms of their microbiota and, to make things even more complex, the diversity spanned in body subsites is comparable with the diversity spanned by completely different kinds of environments.

While, to some extent, the heterogeneity arising from known biological and technical factors could be taken into account thanks to well curated metadata, unknown variability is difficult to handle and must be assessed during the exploratory data analysis [38]. Related to high variability, microbiome data, like other sequencing data, are characterised by over-dispersion. Indeed, microbial counts can vary from a few units to several thousands for the same taxon across specimens.

Each sample is also characterised by its own library size due to different sequencing efficiency across samples [39]. As a consequence, the number of reads eligible for quantification are considerably influenced by the different sequencing depths making the comparison across samples harder. Normalisation and transformation methods have been implemented to overcome these computational challenges [40–42].

Moreover, sequencing instruments can deliver reads only up to a fixed capacity. Thus, it is proper to think of these data as compositional to highlight that they represent a random sample of the relative abundance of the microbes in the ecosystem being studied (see Fig. 1.4) [43]. In general, compositionality refers to a statistical framework that deals with data representing relative proportions or percentages of different parts within a whole, where the sum of the parts is constant. While NGS and metabarcoding data, in their raw form, are counts rather than proportions, the term "compositional" can be used in the context of microbiome analysis to capture the relative abundance or proportions of different taxa within a sample or ecosystem. Although the raw NGS data itself is not inherently compositional, the parameter of interest in the analysis is indeed the proportion or relative abundance of taxa, which can be considered compositional. It is important to recognize that the compositionality of the data is distinct from the nature of the data itself. This compositional aspect of the data provides valuable information about the relationships between the parts, *i.e.*, the microbes in the ecosystem, and should be analysed after the proper transformations [44]. Compositionality of data is not the only reason for considering the existence of some relationships between the mapped microorganisms in the dataset. Indeed, the microbiota is a cooperative system, whereby microorganisms interact in a biological or biochemical relationship, including mutualistic or antagonistic relationships [45]. This inter-variable dependency should be appropriately modelled by either including into the statistical analyses a dependency structure or using multivariate approaches

[45, 46].

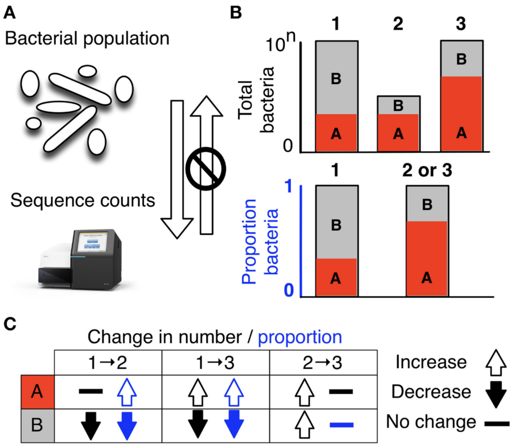
Together with the high microbiota variability among samples and the compositionality of the data, we find another fundamental characteristic of microbiome datasets: sparsity, *i.e.*, a high number of null values. Zero counts could have originated from microbes that are not present in the sample for biological reasons (structural or biological zeroes), by relatively rare microbes compared to others in the specimen, for which the sequencing depth was not sufficient (sampling zeroes), or by technical bias that inhibits the measurement of specific transcripts (technical zeroes) [47]. Sparsity can be handled simply by discarding rare features, with the inevitable loss of information, or by using more sophisticated statistical methodologies and computational techniques which models the zero inflation or that differently manage zeroes.

1.2.3 Quality control and ecological exploration

According to the state of the art guidelines, also detailed in the online book "Orchestrating Microbiome Analysis" [48], the quality control and exploration of microbiome data is the first step for any further analysis and model building.

Technical biases should not affect the dataset and standard summaries and graphical representations should raise awareness towards the presence of outliers, patterns, batches, contaminants, and so forth. Abundance and prevalence metrics, *i.e.*, the frequency of samples where certain microbes were detected, can be graphically assessed (Fig. 1.5 a, b). This to focus on changes which pertain to the majority of the samples, or identify rare microbes, which may be conditionally abundant in a small number of samples. Library size comparisons allow to detect outliers or, in combination with the number of distinct sequences, depicts whether the sampling depth was sufficient to estimate microbial diversity (Fig. 1.5 c, d).

Indeed, microbial diversity estimation is a central topic in microbiome data analysis. α-diversity is used to describe the within-sample diversity and



Change in number / proportion

	1→2	1→3	2→3
A	— ⇧	⇧ ⇧	⇧ —
B	⬇ ⬇	⬇ ⬇	⇧ —

Figure 1.4: From the work of Gloor et al. [43] this figure elucidates the compositional nature of HTS data. **a** Illustrates that the data observed after sequencing a set of nucleic acids from a bacterial population cannot inform on the absolute abundance of molecules. The number of counts in a HTS dataset reflect the proportion of counts per feature (OTU, gene, etc.) per sample, multiplied by the sequencing depth. Therefore, only the relative abundances are available. **b** The bar plots show the difference between the count of molecules and the proportion of molecules for two features, A (red) and B (gray) in three samples. The top bar graphs show the total counts for three samples, and the height of the color illustrates the total count of the feature. When the three samples are sequenced we lose the absolute count information and only have relative abundances, proportions, or “normalised counts” as shown in the bottom bar graph. Note that features A and B in samples 2 and 3 appear with the same relative abundances, even though the counts in the environment are different. **c** The table shows real and perceived changes for each sample if we transition from one sample to another.

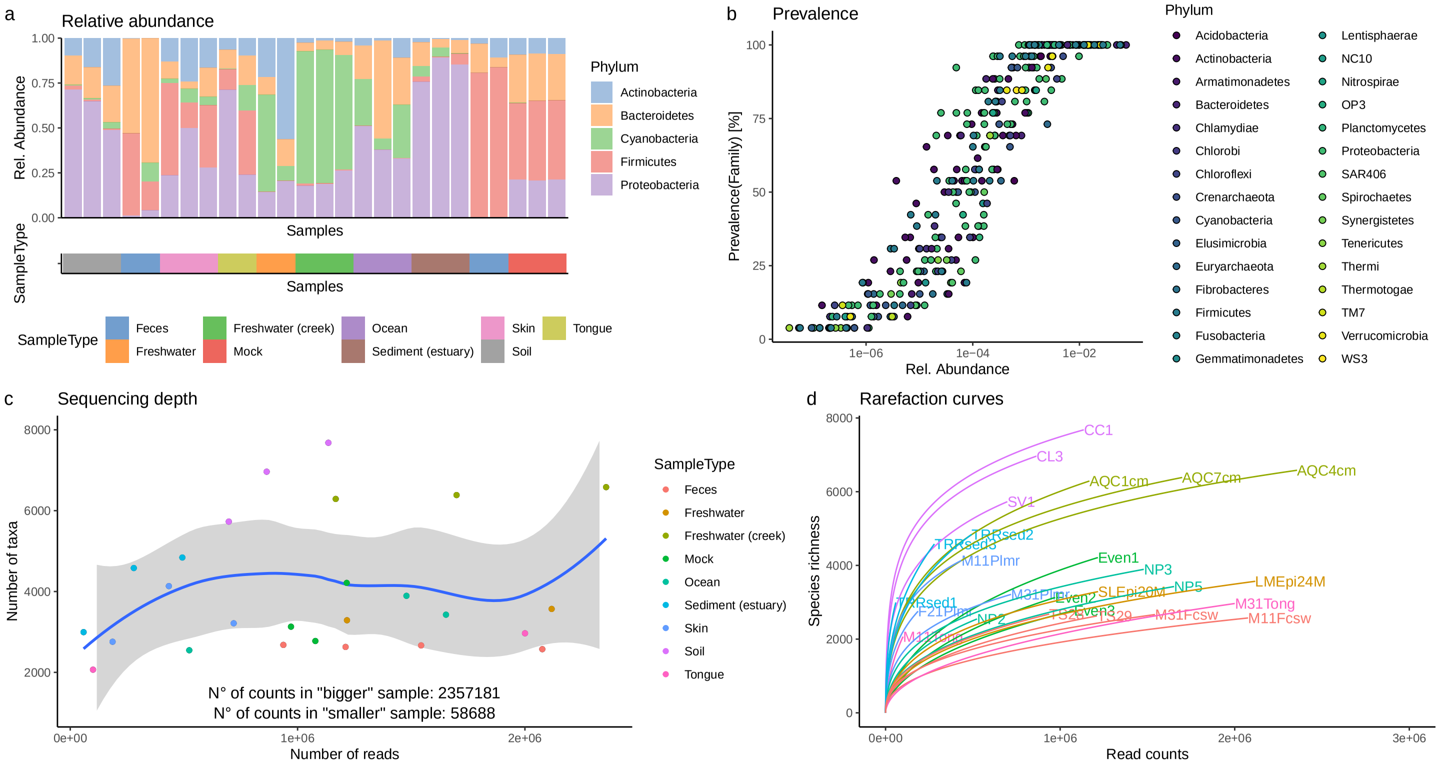


Figure 1.5: Some examples of the quality control and exploration of a microbiome dataset. This specific dataset is taken from the work of Caporaso *et al.* [49] where the microbial communities of environmental samples and known “mock communities” are available. **a** Exploration may involve questions such as how microorganisms are distributed across samples (abundance) and **b** what microorganisms are present in most of the samples (prevalence). **c** Quality control, instead, may involve the relationships between the total number of reads of a sample and its corresponding number of features. **d** More sophisticated methods such as the rarefaction curves are created by randomly re-sampling the pool of samples several times and then plotting the average number of species found on each sample. An ascending graph implies insufficient sampling depth to capture the overall microbial richness, while a curve indicates saturation.

several metrics are available. Many α-diversity indexes are based on a combination of sample richness and evenness such as the Shannon, Simpson, and Inverse Simpson indices. Sample richness is the simplest measure, it consists in counting up the observed number of different taxa in a sample. Differently, the evenness is the extent to which species are evenly distributed. Other α-diversity measures are based on phylogenetic information instead. While α-diversity focuses on community variation within a sample, β-diversity describes the between-sample dissimilarities. Bray-Curtis index (for compositional data), Jaccard index (for presence/absence data), Aitchison distance (*i.e.*, the euclidean distance for CLR transformed abundances), and the UniFrac distances (based on the phylogenetic information) are some of the most common beta diversity measures. To evaluate and visualise these dissimilarities, ordination methods are used. They summarise community data by producing a low-dimensional ordination space in which similar species and samples are plotted close together, and dissimilar species and samples are placed far apart [50]. The most common ordination methods used in microbiome research belong to the exploratory multivariate methods [51]. They include Principal Component Analysis (PCA), Non-metric MultiDimensional Scaling (NMDS), and Principal Coordinates Analysis (PCoA). Alternatively, when the aim of the ordination is to study the association between community composition and variables of interest, interpretive multivariate methods such as ReDundancy Analysis (RDA) and Canonical Correspondence Analysis (CCA) can be used [51].

When the actual statistical significance values for community differences between groups of samples are of interest, the PERMutational ANalysis Of VAriance (PERMANOVA) [52] is a widely used non-parametric multivariate approach. Not only does it assess the group's centroids closeness in the ordinated space, but it also allows the study of group dispersions.

Introducing machine learning approaches, supervised or unsupervised clustering techniques can be used to find groups of samples which share similar community profiles. Finally, discriminative multivariate methods such

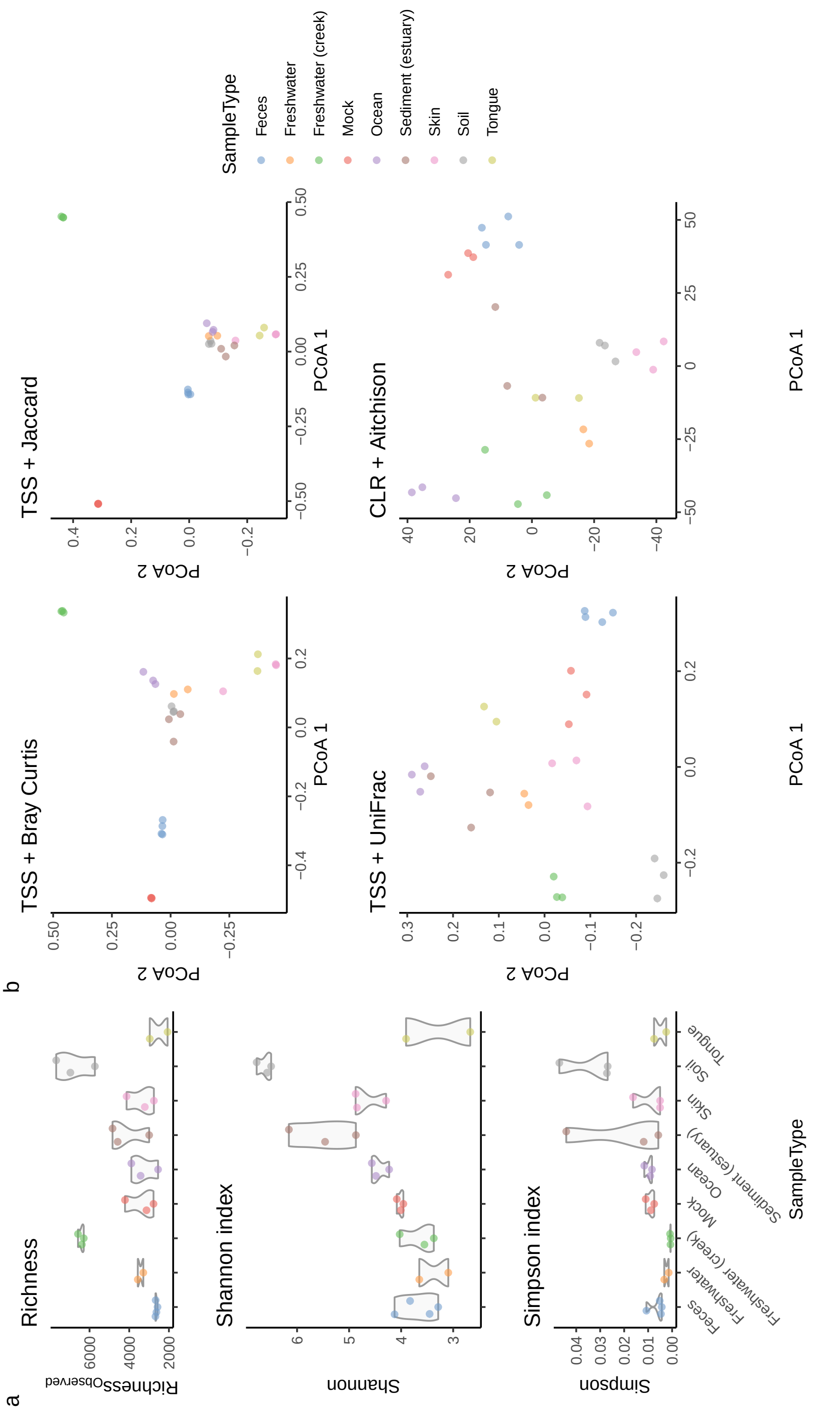


Figure 1.6: Some examples of the ecological analysis for the dataset introduced in Fig. 1.5. **a** α-diversity indexes. The observed richness simply counts the number of distinct taxa in the samples. Shannon and Simpson indexes instead, combine the number of diverse species and their proportions in the whole community. **b** β-diversity indexes based on Total Sum Scaling (TSS) normalised counts or Centred Log-Ratio (CLR) transformed counts. PCoA ordination based on Bray-Curtis and Jaccard dissimilarities, UniFrac phylogenetic distance, and Aitchison distance.

as Support Vector Machines (SVM), Random Forests (RF), sparse Partial Least Squares Discriminant Analysis (sPLS-DA), and others are an extension of the interpretive multivariate techniques used to maximise the separation of samples among different classes. The so-called loadings, *i.e.*, the coefficients computed by these methods, measure the relative contribution of each member of the community (or a subset of it) to the separation. These kinds of methods also permit the class prediction of new samples based on the microbial community composition, opening them to a potential diagnostic use.

While newcomers in microbiome analysis tend to employ simpler exploratory techniques, interpretive and discriminatory ordination approaches with a more rigorous multivariate hypothesis testing are taking place. The latter enable the linking of ecological and functional measures of microbial communities with environmental gradients, host (patient) information, and time and space variables [51]. Nevertheless, great caution must be taken when using complex multivariate methods. Indeed, the increased complexity may solve spurious results due to otherwise simplistic method's assumptions but, in return, they could also result in a lower interpretability [53].

1.2.4 Differential abundance analysis

The identification of microbial taxa whose abundance is different across groups of samples is one of the main goals in microbiome data analysis. Just to cite a common application, the identified microbial taxa could offer biological insights into disease mechanisms and potentially be further explored as biomarkers for disease prevention, diagnosis, and treatment.

Numerous Differential Abundance Analysis (DAA) methods have been proposed in the past decades, *e.g.*, [40, 54–59], (and keep being proposed, *e.g.*, [60, 61]), trying to address the challenging characteristics of microbiome data. See Table 1.1 for a list of DAA tools included in the latest reviews in the field [62–64]. According to Yang and Chen [62], one way to differentiate between DAA tools is based on how they address zero inflation, *i.e.*, sparsity, and compositional effect.

In an over-dispersed count model, an overdispersion parameter is estimated to model the variability of the data as well as the level of sparsity. In this context, sampling and biological zeroes (*i.e.*, those due to insufficient sampling depth and those genuinely not present in the sample, respectively) are treated as any other count value within the count distribution and it is generally not possible to distinguish between them. While for the majority of low-abundance taxa it is reasonable to assume all zeroes as a combination of both sampling and biological zeroes, this may not hold for more abundant ones [47]. Indeed, for more abundant taxa, it becomes less likely that all zero counts can be solely attributed to sampling limitations or insufficient sequencing depth. In these instances, there is a higher probability that some of the zeroes are primarily due to biological absences rather than sampling artefacts. To this regard, mixture models are more flexible. A mixture component in zero specifically handles the structural zeroes, *i.e.*, those which are truly absent from a group of samples due to biological reasons, while the non-zero mixture component acts like the previously described over-dispersed count models. The extra parameter for the structural zero component significantly increases the modelling capability for zero-inflated counts. However, the increased complexity translates to more computational burden, potential overfitting, and the consequent power loss and estimates' instability. Finally, hurdle models refer to those that divide the modelling stage into two parts to correct for excess zeros. The first part determines whether the response outcome is a zero (of any type) or a positive count via a binary model (*e.g.*, by using logistic regression). Then conditioning on it being positive, the second stage models the level of the outcome which is a truncated-at-zero count outcome (*e.g.*, truncated Gaussian, truncated Negative Binomial, truncated Poisson). From the assessment of Xu and colleagues [65], hurdle models have similar goodness of fit and parameter estimation for the count component as their corresponding Zero-Inflated models. However, the estimation and interpretation of the parameters for the zero components differs, and hurdle models are more stable when structural zeroes are absent.

Another way to handle sparsity is through zero replacement. Many zero imputation methods are available in the literature [66] and rely on pseudo-count addition (*e.g.*, MaAsLin2 [67], ANCOM [59], ANCOM-II [68], and ANCOM-BC [60]) or a combination of pseudo-counts/uninformative or informative priors assumptions in a Bayesian fashion (*e.g.*, ALDEx2 assumes an uninformative prior Dirichlet distribution on the taxa proportions and a multinomial sampling process for the observed counts [57], eBay uses an Empirical Bayes approach with an informative prior estimated from the data to improve estimation efficiency [69]). Finally, zeroes may also be left untreated (*e.g.*, LDM [70] and DACOMP [71]).

Regarding the compositional structure of microbiome data, DAA tools leverage four main strategies to address it:

- The robust normalisation approach calculates normalisation factors or size factors to be used as offsets in count-based models or as divider to obtain normalised counts. Their aim is to capture the invariant part of the count distribution and be robust to outliers and differential features. These methods mostly rely on the assumption that the dataset to be normalised has a large invariant part and the majority of features do not change with respect to the condition under study [72]. Example of these are summarised in Table 1.2 and they are the Total Sum Scaling (TSS), Trimmed Mean of M-values (TMM) (used by edgeR [56]), Relative Log Expression (RLE) and its variant for sparse data (used by DESeq2 [73]), Cumulative Sum Scaling (CSS) (used by metagenomeSeq [40]), Centred Log-Ratio (CLR) (used by ALDEx2 [57]), Geometric Mean of Pairwise Ratios (GMPR) (used by the Omnibus test [72, 74]), and Wrench [54].
- To extend the previous, the reference taxa approach aims to find one taxon or a set of taxa that are relatively invariant with respect to the condition of interest and then use it or them to construct the

size factors. This is pursued through network based normalisation to find invariant taxa (*e.g.*, in RioNorm2 [75]), inspecting the median standard deviations of all the pairwise comparisons and choosing taxa below a critical threshold as a reference (*e.g.*, DACOMP [71]), or finding a taxon or a group of them which makes the least discoveries in DAA (*e.g.*, in RAIDA [76]).

- The pairwise log ratios approach (*e.g.*, used by ANCOM [59] and by DACOMP [71] prior to selecting the reference set of taxa) relies on the fact that the log ratios between non-DA taxa are considered constant and independent from the grouping variable. Therefore, it is expected that a high number of rejections per taxon will be a DA indication.
- The latest developed bias-correction approach (*e.g.*, as implemented in ANCOM-BC [60] or linDA [61]) takes into account that each sample is an unknown fraction of a unit volume of the ecosystem, and the sampling fraction varies from sample to sample. Exploiting the fact that a large number of taxa on each specimen is available, the information across taxa can be borrowed to estimate the sampling fraction bias [60]. Then it is included in linear regression models as an offset term to address the bias.

Given the DAA methods' variety, both in terms of theoretical assumptions and results, several attempts have been made to systematically benchmark DAA tools against one another [62, 64, 77–81]. They agree that none of the existing DAA methods can be applied blindly to any real microbiome dataset. According to Yang and Chen [62], an ideal DAA method should be scalable, to permit the analysis of large scale studies, flexible, to allow covariates adjustments and adapt to several experimental designs, robust, and powerful, being able to control false discoveries without sacrificing its ability to identify true findings. Robustness and power are critically needed to yield reliable microbiome biomarkers, increase the reproducibility across microbiome studies, and ultimately reduce the development cost. Indeed, the authors conclude that the applicability of an existing DAA method depends on specific settings, which are usually unknown a priori. Thus, they propose zicoSeq, a new tool drawing the strengths of other good performing methods [62]. Alternatively, the latest recommendations in DAA tool's choice provide for the use of more than one method simultaneously to retrieve a general consensus [48, 80, 81] or direct assessment of tools' performances on simulated or user's datasets to obtain the best performing method [64, 82, 83] (one of the benchmarking research and its application are proposed in Chapter 3 and 4).

Method	Short description	Citation
ALDEx2[1]	It uses a Bayesian estimation of true relative abundance by Monte Carlo sampling. CLR or similarly transformed data are tested and effects' p-values and q-values are provided after statistical testing.	[57]
ANCOM[1]	For a given taxon, the number of Additive Log-Ratio (ALR) transformed models where the taxon is differentially abundant with regard to the variable of interest is used to determine DA features.	[59]
ANCOM-II[1]	Log-ratios between the features of the samples belonging to the same experimental group are described with an ANOVA model. Structural, outlier, or sampling zeroes are detected and differently handled.	[68]
ANCOM-BC[1]	It estimates the unknown sampling fractions for each sample, corrects the bias induced by their differences through a log linear regression model including the estimated sampling fraction as an offset term.	[60]
corncob[1]	It studies differential abundance and differential variability by fitting a beta-binomial model. Two different testing procedures are implemented: Wald test or Likelihood Ratio Test (LRT).	[84]
DACOMP[1]	It is a non-parametric approach that uses a set of reference taxa that are non-differentially abundant, which can be estimated from the data or from outside information.	[71]
dearseq[2]	It uses log2-transformed CPMs to perform a variance component score test to detect differences between groups of samples accounting for data heteroscedasticity through precision weights.	[85]
DESeq2[2]	A Negative Binomial distribution is exploited to model the observed counts. RLE normalisation is used as an offset in a generalised linear model. Wald test or LRT are used to evaluate DA taxa.	[73][58]
eBay[1]	It uses an empirical Bayes approach. An informative prior and a Dirichlet-Tree Multinomial distribution are used to include phylogenetic information into the analysis.	[69]
edgeR[2]	A Negative Binomial distribution is used to model the observed counts. The TMM normalisation is used to correct the library sizes. Empirical Bayes estimation is used to detect DA features.	[56]
fastANCOM[1]	Uses a fast inference algorithm by exploiting a connection between linear models for all pairwise log-ratios of counts and linear models for log transformed counts.	[86]
LDM[1]	LDM uses a linear model on abundances or transformed counts. A decomposition of the sum of squares is used to test global and taxon-specific DA.	[70]
limma[2]	A linear model of the log2-transformed CPMs is fitted and residuals used to compute weights. An empirical Bayesian approach is then used for parameter estimation.	[58, 87]
linDA[1]	It fits linear regression models on the CLR transformed data, and corrects the bias due to compositional effects by using the mode of the regression coefficients across different taxa.	[61]
MaAsLin[1]	It performs association testing through a multi-model framework with arbitrary coefficients and contrasts of interest. Several models, normalisations, and transformations are available for different data types.	[67]
maSigPro[2]	It performs a two steps regression strategy to find longitudinal differential abundant features.	[88]
MAST[3]	A Truncated Gaussian hurdle model is used to describe the log2-transformed CPMs.	[89]
MetaDprof[1]	It consists in a two-stages spline-based method assuming heterogeneous error for detecting differentially abundant features across time.	[90]
metagenomeSeq[1]	Observed absolute abundances are modelled through Zero-Inflated Log-Normal mixture model using the built-in CSS normalisation.	[40]
metaLonDA[1]	It applies a semi-parametric method known as Smoothing Splines ANanalysis Of VAriance (SS-ANOVA) to detect longitudinal differential abundance. It models the counts using a negative binomial distribution.	[91]
metaSplines[1]	It applies SS-ANOVA to detect longitudinal differential abundance (available with metagenomeSeq).	[92]
mixMC[1]	It studies the association between abundances and variables of interest using a sparse Partial Least Squares Discriminant Analysis (sPLS-DA) on CLR transformed data.	[46]
NBMM[2]	It uses Negative Binomial mixed effects models to perform longitudinal differential abundance analysis through penalised likelihood. Kullback-Leibler distance ratio is used to detect significant differences.	[93]
NBZIMM[1]	It allows the implementation of mixed effects models assuming a Negative Binomial, Zero-Inflated Negative Binomial, or Gaussian distribution.	[94]
NOISeq[2]	It is a comprehensive resource that includes several tools for quality control, normalisation, and filtering. DAA is based on a non-parametric approach for the comparison of taxa statistics and a noise distribution.	[95, 96]
Omnibus[1]	It uses a Zero-Inflated Negative Binomial regression model and winsorized count data are used. The abundance, prevalence, and dispersion can be tested.	[74]
RAIDA[1]	It finds one reference taxon that makes the least discoveries in DAA. It utilises the ratio between features in a modified Zero-Inflated Log-Normal model.	[76]
RioNorm2[1]	It relies on a network based normalisation to find the relatively invariant taxa. It then uses a data driven approach to choose between Zero-Inflated Poisson and Zero-Inflated Negative Binomial to perform regression.	[75]
Seurat[3]	It transforms, normalises, and/or scales counts to assess differences between groups. Several tests could be used to obtain p-values and q-values.	[97]
ZIBBSeq[1]	It uses a regression model based on a Zero-Inflated Beta Binomial distribution. A constrained approach to model the over-dispersion as a polynomial function is used.	[98]
ZIBseq[1]	It uses a Zero-Inflated Beta distribution to model a square root or a cubic root transformation of the relative abundances.	[99]
ZicoSeq[1]	It uses a reference set of taxa adjusted for covariates and a linear model-based Smith permutation test to conduct association testing. Zero imputation is performed through a beta mixture prior.	[62]
ZIGDM[1]	It uses a Zero-Inflated Generalised Dirichlet Multinomial distribution to model taxon counts. Tests are for both differential mean and dispersion.	[100]
ZINQ[1]	It consists of a Zero-Inflated Quantile approach where quantile rank-score based tests on multiple quantiles of the non-zero part with adjustment for the zero inflation are used.	[101]

[1] Developed for microbiome data

[2] Developed for RNA-Seq data or Microarrays

[3] Developed for Single Cell RNA-Seq data

Table 1.1: Rearranged from the latest DAA reviews [62–64]. A list of DAA tools and a short description.

Normalisation	Short description	Citation
Centred Log-Ratio (CLR)	For each sample, the counts are divided by their geometric mean, followed by log transformation. Thus, the CLR size factor is the geometric mean of the counts in a sample.	[44]
Cumulative Sum Scaling (CSS)	The CSS size factor is the cumulative sum of counts up to a percentile determined by a data-driven approach.	[40]
Geometric Mean of Pairwise Ratios (GMPR)	For each sample, the GMPR method calculates the pairwise ratios to all other samples for each feature. The size factor is then the geometric mean of the median ratios for all features.	[72]
Rarefaction	Random subsampling of sequences from the initial sample library to a selected library size (usually the lowest library size among all the samples is used).	[102]
Relative Log Expression (RLE)	The RLE method calculates the geometric means (only for the positive counts in presence of zeroes) of all features as a "reference," and all samples are compared to the "reference" to produce ratios for all features. The median ratio is then taken to be the RLE size factor.	[55]
Total Sum Scaling (TSS)	The TSS size factor is simply the total number of reads in the sample.	
Trimmed Mean of M-values (TMM)	The TMM method first selects a reference sample, then all other samples are compared to the reference sample. The weighted trimmed mean of log-ratios between each pair of samples is then calculated as the TMM size factor.	[103]
Wrench	Wrench models feature-wise proportion ratios to a reference sample using a hurdle log-normal model, where a compositional scale factor is included so that the log fold changes on the absolute abundance level is centred at 0 (the majority of the taxa do not change).	[54]

Table 1.2: From the work of Yang and Chen [62], a list of the most used normalisation methods for microbiome data.

1.2.5 Integrative analysis and networks

As already widely discussed, microbes do not work in isolation. They coordinate with one another to form highly organised functional modules, competing for representation within ecological niches. Moreover, they operate in their "theatre of activity" interacting with microbial and environmental structural elements [1]. It is then straightforward to think about the many associations between the microbiome and covariates including metabolites, antibiotic usage, environmental factors, and host genetics that can influence host health [104]. To go beyond the reductionism of DAA, methods that know how to address this biological complexity are necessary [105].

Integrative analysis and correlation networks can be used. These analyses are reviewed by an extensive survey by Lutz and colleagues [63]. The goal of integrative analysis is to identify and quantify associations between the microbiome and covariates. The common biological motivation of each integrative method is to determine if associations exist between any of microbial features and the available covariates while controlling for the phenotypic response. Lutz and colleagues [63], discuss four methods for integrative analysis including Dirichlet-Multinomial Regression [106], Dirichlet-Multinomial Bayesian Variable Selection [107], Bayesian Zero-Inflated Negative Binomial [108], and a Dirichlet-Multinomial Linear Model with Bayesian variable selection [109]. Differently, the goal of network analysis is to build microbiome networks that describe microbial ecological associations (*i.e.*, taxa-taxa dependencies). These are useful to help in discovering fundamental properties and mechanisms of microbial ecosystems. To make inferences about microbial interactions several tools are available. They can be divided into two groups: the correlation-based and the partial correlation-based methods. Correlation-based methods include SparCC (Sparse Correlations for Compositional data) [110], CCLasso (Correlation inference for Compositional data through Lasso) [111], and REBACCA (Regularized Estimation of the BAsis Covariance based on Compositional dAta) [112]. Partial correlation-based methods include SpiecEasi (SParse InversE Covariance Estimation for Ecological Association Inference) [113] and HARMONIES (Hybrid Approach foR MicrobiOme Network Inferences via Exploiting Sparsity) [114]. Instead, SPRING (SemiParametric Rank-based approach for INference in Graphical model) [115] employs both correlation and partial correlation methods under a semi-parametric setting.

1.3 Conclusions

Microbiome research progress, as for many other research areas, is directly driven by technological advancements, the development of new techniques, and the continuous cross-talk between the two.

As an example, the technological advancements of the last decades allowed the study of microbiomes through culture-free approaches. The big amount of produced data required the creation and use of statistical tools for their analysis to get the first biological insights. With them, the instrument limits were uncovered and new analysis tools were developed to overcome or address them, where possible, and get more reliable, robust, unbiased, and hence reproducible results. Some of the analysis tools were too computational intensive to be used and reality's simplifications were necessary to permit the analysis. However, with the assessment of instruments' limits, adjustments and improvement of them were made possible, restarting the cycle.

As a result, different research fields often borrowed statistical tools to analyse similar data structures, despite differences in how the data was generated. However, this practice could lead to model misspecification and assumptions being violated. While some of the theoretical assumptions made by the statistical tools were biologically justifiable, others became obsolete due to the advent of more powerful computing machines, new technologies, or a change in perspective. For instance, in RNA sequencing experiments, it is typically assumed that only a minority of transcripts are differentially expressed, whereas in metabarcoding experiments, the differentially abundant microbes may not be a minority. Nevertheless, many of the current normalisation methods used in microbiome data analysis are based on the assumption that only a minority of features are differentially abundant.

Some delay is needed to let the community update their analysis framework once a new method has been developed or a new biological insight has found its way to be included in a tool. One reason is the distance in time between the theoretical conceptualisation of new biological information to leverage and its actual implementation in a tool that makes it actually available to the majority of the data scientists working in the field. Another reason which could present after the release of the first tool, is the increasing number of available tools that are developed introducing variants and little improvements and the impossibility of trying them all.

The statistical tools needed to handle the increasing complexity of the data will continue to evolve. The biggest challenge in complexity comprehension is however, interpretability. It is difficult to discern the meaning of a complex model used to describe an equally complex system. Despite this, researchers must strive to develop statistical tools that are not only capable of handling complexity but also provide meaningful interpretations of the obtained results. This requires a deep understanding of the underlying biological processes and a willingness to explore new approaches to data analysis.

Microbiome research is addressing many challenges, and it will keep doing it. Systematic reviews of the available tools and up-to-date guidelines are essential to handle the challenges in a collaborative and focused effort. This puts the basis for enhancing the development of new approaches beginning from a solid starting point where the criticisms and the limits of current methodology are clear. Contamination of competencies of experts in statistics, science, and technology will help build on the ideas to offer reliable and interpretable solutions to the many important quests of microbiome research.

References

1. Whipps, J., Lewis, K. & Cooker, R. in *Fungi in Biological Control Systems Roman Catholic Studies* 4, 161–187 (Manchester University Press, 1988).
2. Berg, G. *et al.* Microbiome definition re-visited: old concepts and new challenges. *Microbiome* **8,** 103 (2020).
3. Turnbaugh, P. J. *et al.* The Human Microbiome Project. *Nature* **449,** 804–810 (2007).
4. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551,** 457–463 (2017).
5. McDonald, D. *et al.* American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* **3,** e00031–18 (2018).
6. Vogel, T. M. *et al.* TerraGenome: a consortium for the sequencing of a soil metagenome. *Nature Reviews Microbiology* **7,** 252–252 (2009).
7. Gilbert, J. A. & Dupont, C. L. Microbial Metagenomics: Beyond the Genome. *Annual Review of Marine Science* **3,** 347–371 (2011).
8. Timmis, K. *et al.* The contribution of microbial biotechnology to sustainable development goals. *Microbial Biotechnology* **10,** 984–987 (2017).
9. Arboleya, S. *et al.* C-section and the Neonatal Gut Microbiome Acquisition: Consequences for Future Health. *Annals of Nutrition and Metabolism* **73,** 17–23 (2018).
10. Coelho, G. D. P. *et al.* Acquisition of microbiota according to the type of birth: an integrative review. *Revista Latino-Americana de Enfermagem* **29,** e3446 (2021).
11. Dominguez-Bello, M. G. *et al.* Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. *Nature medicine* **22,** 250–253 (2016).
12. Cullen, C. M. *et al.* Emerging Priorities for Microbiome Research. *Frontiers in Microbiology* **11** (2020).

13. Sommer, F. & Bäckhed, F. The gut microbiota — masters of host development and physiology. *Nature Reviews Microbiology* **11,** 227–238 (2013).
14. Round, J. L. & Mazmanian, S. K. The gut microbiota shapes intestinal immune responses during health and disease. *Nature Reviews Immunology* **9,** 313–323 (2009).
15. Rubinstein, M. R. *et al.* Fusobacterium nucleatum Promotes Colorectal Carcinogenesis by Modulating E-Cadherin/β-Catenin Signaling via its FadA Adhesin. *Cell Host & Microbe* **14,** 195–206 (2013).
16. Martin, C. R., Osadchiy, V., Kalani, A. & Mayer, E. A. The Brain-Gut-Microbiome Axis. *Cellular and Molecular Gastroenterology and Hepatology* **6,** 133–148 (2018).
17. Appleton, J. The Gut-Brain Axis: Influence of Microbiota on Mood and Mental Health. *Integrative Medicine: A Clinician's Journal* **17,** 28–32 (2018).
18. Gevers, D. *et al.* The Treatment-Naive Microbiome in New-Onset Crohn's Disease. *Cell Host & Microbe* **15,** 382–392 (2014).
19. Leustean, A. M. *et al.* Implications of the Intestinal Microbiota in Diagnosing the Progression of Diabetes and the Presence of Cardiovascular Complications. *Journal of Diabetes Research* **2018,** e5205126 (2018).
20. Muir, A. & Falk, G. W. Eosinophilic Esophagitis: A Review. *JAMA* **326,** 1310–1318 (2021).
21. Facchin, S. *et al.* Salivary microbiota composition may discriminate between patients with eosinophilic oesophagitis (EoE) and non-EoE subjects. *Alimentary Pharmacology & Therapeutics* **56,** 450–462 (2022).
22. Olmo, R. *et al.* Microbiome Research as an Effective Driver of Success Stories in Agrifood Systems – A Selection of Case Studies. *Frontiers in Microbiology* **13,** 834622 (2022).
23. Hungria, M., Nogueira, M. A. & Araujo, R. S. Soybean Seed Co-Inoculation with Bradyrhizobium spp. and Azospirillum brasilense:

A New Biotechnological Tool to Improve Yield and Sustainability. *American Journal of Plant Sciences* **6,** 811–817 (2015).

24. Dotsenko, G. *et al.* Enzymatic production of wheat and ryegrass derived xylooligosaccharides and evaluation of their in vitro effect on pig gut microbiota. *Biomass Conversion and Biorefinery* **8,** 497–507 (2018).
25. Zwirzitz, B. *et al.* The sources and transmission routes of microbial populations throughout a meat processing facility. *npj Biofilms and Microbiomes* **6,** 1–12 (2020).
26. Meisner, A. *et al.* Calling for a systems approach in microbiome research and innovation. *Current Opinion in Biotechnology* **73,** 171–178 (2022).
27. Weinstock, G. M. Genomic approaches to studying the human microbiota. *Nature* **489,** 250–256 (2012).
28. Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology* **5,** R245–R249 (1998).
29. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* **13,** 581–583 (2016).
30. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal* **11,** 2639–2643 (2017).
31. Alves, L. d. F. *et al.* Metagenomic Approaches for Understanding New Concepts in Microbial Science. *International Journal of Genomics* **2018,** 1–15 (2018).
32. Handelsman, J. Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews* **68,** 669–685 (2004).
33. Tringe, S. G. *et al.* Comparative Metagenomics of Microbial Communities. *Science* **308,** 554–557 (2005).

34. Schmeisser, C., Steele, H. & Streit, W. R. Metagenomics, biotechnology with non-culturable microbes. *Applied Microbiology and Biotechnology* **75,** 955–962 (2007).

35. Guazzaroni, M.-E., Silva-Rocha, R. & Ward, R. J. Synthetic biology approaches to improve biocatalyst identification in metagenomic library screening. *Microbial Biotechnology* **8,** 52–64 (2015).

36. Douglas, G. M. *et al.* PICRUSt2 for prediction of metagenome functions. *Nature Biotechnology* **38,** 685–688 (2020).

37. Knight, R. *et al.* Best practices for analysing microbiomes. *Nature Reviews Microbiology* **16,** 410–422 (2018).

38. Wang, Y. & LêCao, K.-A. Managing batch effects in microbiome data. *Briefings in Bioinformatics* **21,** 1954–1970 (2020).

39. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics* **15,** 121–132 (2014).

40. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods* **10,** 1200–1202 (2013).

41. Hong, J., Karaoz, U., de Valpine, P. & Fithian, W. To rarefy or not to rarefy: robustness and efficiency trade-offs of rarefying microbiome data. *Bioinformatics* **38,** 2389–2396 (2022).

42. McMurdie, P. J. & Holmes, S. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **8** (2013).

43. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology* **8** (2017).

44. Aitchison, J. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society: Series B (Methodological)* **44,** 139–160 (1982).

45. Weiss, S. *et al.* Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME Journal* **10,** 1669–1681 (2016).
46. Lê Cao, K.-A. *et al.* MixMC: A Multivariate Statistical Framework to Gain Insight into Microbial Communities. *PLoS ONE* **11** (2016).
47. Silverman, J. D., Roche, K., Mukherjee, S. & David, L. A. Naught all zeros in sequence count data are the same. *Computational and Structural Biotechnology Journal* **18,** 2789–2798 (2020).
48. Lahti, L., Shetty, S., Borman, T. & Ernst, F. G. *Orchestrating Microbiome Analysis with Bioconductor* (2021).
49. Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences* **108,** 4516–4522 (supplement_1 2011).
50. Reft, R. K. Ordination as a tool for analyzing complex data sets. *Vegetatio* **42,** 171–174 (1980).
51. Paliy, O. & Shankar, V. Application of multivariate statistical techniques in microbial ecology. *Molecular ecology* **25,** 1032–1057 (2016).
52. Anderson, M. J. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26,** 32–46 (2001).
53. Jiang, D. *et al.* Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities. *Frontiers in Genetics* **10** (2019).
54. Kumar, M. S. *et al.* Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genomics* **19,** 799 (2018).
55. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biology* **11,** R106 (2010).
56. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).

57. Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G. & Gloor, G. B. ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. *PLoS ONE* **8,** e67019 (2013).
58. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15,** R29 (2014).
59. Mandal, S. *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease* **26** (2015).
60. Lin, H. & Peddada, S. D. Analysis of compositions of microbiomes with bias correction. *Nature Communications* **11,** 3514 (2020).
61. Zhou, H., He, K., Chen, J. & Zhang, X. LinDA: linear models for differential abundance analysis of microbiome compositional data. *Genome Biology* **23,** 95 (2022).
62. Yang, L. & Chen, J. A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions. *Microbiome* **10,** 130 (2022).
63. Lutz, K. C. *et al.* A Survey of Statistical Methods for Microbiome Data Analysis. *Frontiers in Applied Mathematics and Statistics* **8** (2022).
64. Cappellato, M., Baruzzo, G. & Di Camillo, B. Investigating differential abundance methods in microbiome data: A benchmark study. *PLoS Computational Biology* **18,** e1010467 (2022).
65. Xu, L., Paterson, A. D., Turpin, W. & Xu, W. Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data. *PLoS ONE* **10,** e0129606 (2015).
66. Lubbe, S., Filzmoser, P. & Templ, M. Comparison of zero replacement strategies for compositional data with large numbers of zeros. *Chemometrics and Intelligent Laboratory Systems* **210,** 104248 (2021).

67. Mallick, H. *et al.* Multivariable association discovery in population-scale meta-omics studies. *PLoS Computational Biology* **17,** e1009442 (2021).

68. Kaul, A., Mandal, S., Davidov, O. & Peddada, S. D. Analysis of Microbiome Data in the Presence of Excess Zeros. *Frontiers in Microbiology* **8,** 2114 (2017).

69. Liu, T., Zhao, H. & Wang, T. An empirical Bayes approach to normalization and differential abundance testing for microbiome data. *BMC Bioinformatics* **21,** 225 (2020).

70. Hu, Y.-J. & Satten, G. A. Testing hypotheses about the microbiome using the linear decomposition model (LDM). *Bioinformatics* **36,** 4106–4115 (2020).

71. Brill, B., Amir, A. & Heller, R. Testing for differential abundance in compositional counts data, with application to microbiome studies. *The Annals of Applied Statistics* **16,** 2648–2671 (2022).

72. Chen, L. *et al.* GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ* **6,** e4600 (2018).

73. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 550 (2014).

74. Chen, J. *et al.* An omnibus test for differential distribution analysis of microbiome sequencing data. *Bioinformatics* **34,** 643–651 (2018).

75. Ma, Y., Luo, Y. & Jiang, H. A novel normalization and differential abundance test framework for microbiome data. *Bioinformatics* **36,** 3959–3965 (2020).

76. Sohn, M. B., Du, R. & An, L. A robust approach for identifying differentially abundant features in metagenomic samples. *Bioinformatics* **31,** 2269–2275 (2015).

77. Thorsen, J. *et al.* Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon

data analysis methods used in microbiome studies. *Microbiome* **4,** 62 (2016).

78. Hawinkel, S., Mattiello, F., Bijnens, L. & Thas, O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Briefings in Bioinformatics* **20,** 210–221 (2019).
79. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5,** 27 (2017).
80. Calgaro, M., Romualdi, C., Waldron, L., Risso, D. & Vitulo, N. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biology* **21,** 191 (2020).
81. Nearing, J. T. *et al.* Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications* **13,** 342 (2022).
82. Patuzzi, I., Baruzzo, G., Losasso, C., Ricci, A. & Di Camillo, B. metaSPARSim: a 16S rRNA gene sequencing count data simulator. *BMC Bioinformatics* **20,** 416 (2019).
83. Calgaro, M., Romualdi, C., Risso, D. & Vitulo, N. benchdamic: benchmarking of differential abundance methods for microbiome data. *Bioinformatics* **39,** btac778 (2023).
84. Martin, B. D., Witten, D. & Willis, A. D. Modeling microbial abundances and dysbiosis with beta-binomial regression. *The Annals of Applied Statistics* **14,** 94–115 (2020).
85. Gauthier, M., Agniel, D., Thiébaut, R. & Hejblum, B. P. dearseq: a variance component score test for RNA-seq differential analysis that effectively controls the false discovery rate. *NAR Genomics and Bioinformatics* **2,** lqaa093 (2020).
86. Zhou, C., Wang, H., Zhao, H. & Wang, T. fastANCOM: a fast method for analysis of compositions of microbiomes. *Bioinformatics* **38,** 2039–2041 (2022).

87. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43,** e47–e47 (2015).

88. Conesa, A., Nueda, M. J., Ferrer, A. & Talón, M. maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics* **22,** 1096–1102 (2006).

89. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16,** 278 (2015).

90. Luo, D., Ziebell, S. & An, L. An informative approach on differential abundance analysis for time-course metagenomic sequencing data. *Bioinformatics* **33,** 1286–1292 (2017).

91. Metwally, A. A. *et al.* MetaLonDA: a flexible R package for identifying time intervals of differentially abundant features in metagenomic longitudinal studies. *Microbiome* **6,** 32 (2018).

92. Paulson, J. N., Talukder, H. & Corrada Bravo, H. *Longitudinal differential abundance analysis of microbial marker-gene surveys using smoothing splines* 2017.

93. Sun, X. *et al.* Statistical inference for time course RNA-Seq data using a negative binomial mixed-effect model. *BMC Bioinformatics* **17,** 324 (2016).

94. Zhang, X. & Yi, N. NBZIMM: negative binomial and zero-inflated mixed models, with application to microbiome/metagenomics data analysis. *BMC Bioinformatics* **21,** 488 (2020).

95. Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: a matter of depth. *Genome Research* **21,** 2213–2223 (2011).

96. Tarazona, S. *et al.* Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research,* gkv711 (2015).

97. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36,** 411–420 (2018).
98. Hu, T., Gallins, P. & Zhou, Y.-H. A Zero-inflated Beta-binomial Model for Microbiome Data Analysis. *Stat (International Statistical Institute)* **7,** e185 (2018).
99. Peng, X., Li, G. & Liu, Z. Zero-Inflated Beta Regression for Differential Abundance Analysis with Metagenomics Data. *Journal of Computational Biology* **23,** 102–110 (2016).
100. Tang, Z.-Z. & Chen, G. Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics* **20,** 698–713 (2019).
101. Ling, W. *et al.* Powerful and robust non-parametric association testing for microbiome data via a zero-inflated quantile approach (ZINQ). *Microbiome* **9,** 181 (2021).
102. McMurdie, P. J. & Holmes, S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Computational Biology* **10,** e1003531 (2014).
103. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11,** R25 (2010).
104. Levy, M., Thaiss, C. A. & Elinav, E. Metabolites: messengers between the microbiota and the immune system. *Genes & Development* **30,** 1589–1597 (2016).
105. Quinn, T. P., Gordon-Rodriguez, E. & Erb, I. *A Critique of Differential Abundance Analysis, and Advocacy for an Alternative* 2021.
106. Chen, J. & Li, H. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *The Annals of Applied Statistics* **7,** 418–442 (2013).

107. Wadsworth, W. D. *et al.* An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinformatics* **18,** 94 (2017).

108. Jiang, S. *et al.* A Bayesian zero-inflated negative binomial regression model for the integrative analysis of microbiome data. *Biostatistics* **22,** 522–540 (2021).

109. Koslovsky, M. D., Hoffman, K. L., Daniel, C. R. & Vannucci, M. A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes. *The Annals of Applied Statistics* **14,** 1471–1492 (2020).

110. Friedman, J. & Alm, E. J. Inferring Correlation Networks from Genomic Survey Data. *PLoS Computational Biology* **8,** e1002687 (2012).

111. Fang, H., Huang, C., Zhao, H. & Deng, M. CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics* **31,** 3172–3180 (2015).

112. Ban, Y., An, L. & Jiang, H. Investigating microbial co-occurrence patterns based on metagenomic compositional data. *Bioinformatics* **31,** 3322–3329 (2015).

113. Kurtz, Z. D. *et al.* Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLoS Computational Biology* **11,** e1004226 (2015).

114. Jiang, S. *et al.* HARMONIES: A Hybrid Approach for Microbiome Networks Inference via Exploiting Sparsity. *Frontiers in Genetics* **11** (2020).

115. Yoon, G., Gaynanova, I. & Müller, C. L. Microbial Networks in SPRING - Semi-parametric Rank-Based Correlation and Partial Correlation Estimation for Quantitative Microbiome Data. *Frontiers in Genetics* **10** (2019).

Chapter 2

About this thesis

2.1 Thesis structure

After a master degree in Statistical Sciences, microbiome data analysis has been the main objective of my PhD programme. From my first approach to metabarcoding data I realised the overwhelming variety, and sometimes complexity, of statistical tools available to pursue the many research questions I was facing. Only through a process of gradually understanding the interplay between biological mechanisms and data-generating processes have I been able to understand the most appropriate statistical tools for the situation at hand.

The objective of this thesis is to describe my scientific research and its challenges through a collection of 4 research articles that I published during the last 3 years of PhD. As a main goal, the thesis moves from the observed limitations of the DAA tools to a benchmark and a framework against which they could be measured and compared. As a secondary goal, some case studies of metabarcoding analysis highlight the need for sound exploratory and inferential statistical analysis. The thesis is divided into two parts as follows.

In Part I - Differential Abundance Analysis, I present two closely related studies in which DAA methods play the main role. As explained in detail in Chapter 1, DAA consists of a variety of methods and is one of the most important approaches for detecting differences in microbial community composition between different sample groups, for understanding microbial community structures, and the relationships between microbial compositions and the environment. The aim was to illustrate how different DA methods lead to different results and how to choose the right methods for each dataset. First, in Chapter 3, I present a benchmarking study [1] in which DA methods from different domains were used in a collection of microbiome datasets to evaluate their performance. Here I performed the data analyses and curated the code repository. As a consequence of the results obtained in chapter 3, I realised that a benchmarking study alone, despite its comprehensiveness, was not sufficient to drive methodological decisions on an every-day basis, indiscriminately for all microbiome datasets, due to the peculiarities of each one. For this reason I developed benchdamic [2], presented in Chapter 4, a ready-to-use R application available on the Bioconductor platform (*i.e.*, an open-source software platform for analyzing and visualizing biological data) that allows users to replicate DAA's performance evaluation on their own datasets. Here my contribution involved the package development and maintenance, including the creation of an extensive manual.

An equally important aspect of my PhD programme is well represented by the many collaborations which allowed me to extensively interact with clinicians, microbiologists and other researchers. They allowed me to i) test exploratory and DAA methods and best practices on real data and ii) study the human microbiome and its composition and dynamics in both disease and healthy states. For example, in the research of Guidetti and colleagues [3] I performed the metabarcoding data analysis for a randomized double-blind crossover study, evaluating the effects of a probiotics mixture in a group of 61 subjects (2-16 years) with an Autism Spectrum Disorder diagnosis. Instead, in the research of Solito *et al.* [4], we assessed the impact of a probiotic supplementation in pediatric obesity on weight, metabolic alterations, selected gut microbial groups, and functionality. The study design was a cross-over, double-blind, randomized control trial involving 101 youths (6-18 years) with obesity and insulin-resistance on diet. Here, I used the mixed-effects models to perform statistical analyses and I was able to highlight the beneficial effects of the probiotics on insulin sensitivity. In another work, I performed the metabarcoding data analysis in a double-blind, placebo-controlled, pilot study of Facchin *et al.* [5]. 49 patients (19 with Crohn's disease and 30 with ulcerative colitis) were randomized to oral administration of microencapsulated-sodium-butyrate or placebo for 2 months. Here we showed that sodium-butyrate supplementation seems to be associated with the growth of bacteria able to produce Short Chain Fatty Acid (SCFA) with potentially anti-inflammatory action.

Thus, in part II - Case Studies, I present two other papers where I played a major role. They are an observational study of healthy individuals taking probiotics [6] and a clinical research to identify disease-related biomarkers [7]. In Chapter 5, healthy volunteers were treated with a probiotic mixture and the changes in the gut microbiome were studied in conjunction with some psychological aspects [6]. This study was extremely useful to highlight exploratory data analysis difficulties. In brief, an unsupervised clustering approach on ordinated samples allowed the separation of the individuals in several groups associated to different putative bacteriotypes (*i.e.*, stable clusters of bacterial communities that co-exist together). Apparently, each bacteriotype reacted differently to the probiotic intake and its global effect was confounded by their presence. Instead, chapter 6 explored the potential of DAA methods to identify disease-related microbial biomarkers for Eosinophilic oEsophagitis (EoE) (*i.e.*, a chronic immune-mediated inflammatory disease of the oesophagus that causes dysphagia, food impaction of the oesophagus, and esophageal strictures) from saliva [7]. A promising result was obtained for the non-invasive diagnosis of EoE which is now possible only through esophageal biopsy.

References

1. Calgaro, M., Romualdi, C., Waldron, L., Risso, D. & Vitulo, N. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biology* **21,** 191 (2020).
2. Calgaro, M., Romualdi, C., Risso, D. & Vitulo, N. benchdamic: benchmarking of differential abundance methods for microbiome data. *Bioinformatics* **39,** btac778 (2023).
3. Guidetti, C. *et al.* Randomized Double-Blind Crossover Study for Evaluating a Probiotic Mixture on Gastrointestinal and Behavioral Symptoms of Autistic Children. *Journal of Clinical Medicine* **11,** 5263 (2022).
4. Solito, A. *et al.* Supplementation with Bifidobacterium breve BR03 and B632 strains improved insulin sensitivity in children and adolescents with obesity in a cross-over, randomized double-blind placebo-controlled trial. *Clinical Nutrition* **40,** 4585–4594 (2021).
5. Facchin, S. *et al.* Microbiota changes induced by microencapsulated sodium butyrate in patients with inflammatory bowel disease. *Neurogastroenterology & Motility* **n/a,** e13914 (n/a 2020).
6. Calgaro, M. *et al.* Metabarcoding analysis of gut microbiota of healthy individuals reveals impact of probiotic and maltodextrin consumption. *Beneficial Microbes* **12,** 121–136 (2021).
7. Facchin, S. *et al.* Salivary microbiota composition may discriminate between patients with eosinophilic oesophagitis (EoE) and non-EoE subjects. *Alimentary Pharmacology & Therapeutics* **56,** 450–462 (2022).

Part I

Differential Abundance Analysis

Chapter 3

Assessment of statistical methods from single cell, bulk RNA-Seq, and metagenomics applied to microbiome data

The work described in this chapter is taken from: M. Calgaro, C. Romualdi, L. Waldron, D. Risso, N. Vitulo; Assessment of statistical methods from single cell, bulk RNA-Seq, and metagenomics applied to microbiome data. Genome Biol 21, 191 (2020). https://doi.org/10.1186/s13059-020-02104-1

Contributions

DR, CR, and NV conceived the project. LW co-developed the evaluation strategies. **MC** and DR drafted the manuscript. LW, CR, and NV reviewed and edited the manuscript. **MC** performed the data analyses and curated the code repository. All Authors read and approved the final manuscript.

Challenges and future perspectives

This article is my first attempt to explore the tools of Differential Abundance Analysis (DAA). It arose from a practical problem I faced in other projects when I had to select a DAA tool from a long list to detect Differential Abundant (DA) features between two groups of samples. Each tool gave different results, also depending on the choice of normalisation and filtering. Although there was some agreement between the methods, it was important for me to understand if there was one method that was superior to the others.

Given my background in statistical sciences, I first examined the assumptions of the tools based on parametric distributions and found that there was little assessment of the appropriateness of these assumptions for the real data. For this reason, I began to benchmark tools for DAA. I focused not only on whether the assumptions of the tools held, but also examined false discovery rates, concordance of results, and power. The main challenge of this project, which continues in the next chapter, was to move from the observed limitations of the DAA tools to a framework against which they could be measured and compared. This was made possible through a process of trial and error that allowed the other authors and I to select the most communicative analyses. Another aspect that required a lot of effort was the development of appropriate visualisation tools, looking for the proper graphical output to make the interpretation of the results easier, especially for non-expert data-scientists.

As mentioned in the first chapter, at the time of writing this thesis, several new DAA tools has been developed and published. Nevertheless, the findings described in the paper are still valid and are partially confirmed by independent studies.

Article location

`https://doi.org/10.1186/s13059-020-02104-1`

Supplementary material

The supplementary material for this article consists of 2 additional files available at `https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02104-1#additional-information`.

3.1 Abstract

Background

The correct identification of differentially abundant microbial taxa between experimental conditions is a methodological and computational challenge. Recent work has produced methods to deal with the high sparsity and compositionality characteristic of microbiome data, but independent benchmarks comparing these to alternatives developed for RNA-Seq data analysis are lacking.

Results

We compare methods developed for single-cell and bulk RNA-Seq, and specifically for microbiome data, in terms of suitability of distributional assumptions, ability to control false discoveries, concordance, power, and correct identification of differentially abundant genera. We benchmark these methods using 100 manually curated datasets from 16S and whole metagenome shotgun sequencing.

Conclusions

The multivariate and compositional methods developed specifically for microbiome analysis did not outperform univariate methods developed for differential expression analysis of RNA-Seq data. We recommend a careful exploratory data analysis prior to application of any inferential model and we present a framework to help scientists make an informed choice of analysis methods in a dataset-specific manner.

3.2 Background

Study of the microbiome, the uncultured collection of microbes present in most environments, is a novel application of high-throughput sequencing that shares certain similarities but important differences from other applications of DNA and RNA sequencing. Common approaches for microbiome studies are based on deep sequencing of amplicons of universal marker-genes, such as the 16S rRNA gene, or on Whole Metagenome shotgun Sequencing (WMS). Community taxonomic composition can be estimated from microbiome data by assigning each read to the most plausible microbial lineage using a reference annotated database, with a higher taxonomic resolution in WMS than in 16S [1, 2]. The final output of such analyses usually consists of a large, highly sparse, taxa per sample count table.

Differential Abundance Analysis (DAA) is one of the primary approaches to identify differences in the microbial community composition between samples and to understand the structures of microbial communities and the associations between microbial compositions and the environment. DA analysis has commonly been performed using methods adapted from RNA-Seq analysis; however, characteristics specific to microbiome data make differential abundance analysis challenging. Compared to other high-throughput sequencing techniques such as RNA-Seq, metagenomic data are sparse, *i.e.*, the taxa count matrix contains many zeros. This sparsity can be explained by both biological and technical reasons: some taxa are very rare and present only in a few samples, while others are very lowly represented and cannot be detected because of an insufficient sequencing depth or other technical reasons.

In recent years, single cell RNA-Seq (scRNA-Seq) has revolutionised the field of transcriptomics, providing new insight on the transcriptional program of individual cells, casting light on complex, heterogeneous tissues, and revealing rare cell populations with distinct gene expression profiles [3–6]. However, due to the relatively inefficient mRNA capture rate, scRNA-Seq data are characterised by dropout events, which leads to an excess of zero read counts compared to bulk RNA-Seq data [7, 8]. Thus, with the advent of this technology, new statistical models accounting for dropout events have been proposed. The similarities with respect to sparsity observed in both scRNA-Seq and metagenomics data led us to pose the question of whether statistical methods developed for the differential expression of scRNA-Seq data perform well on metagenomic DA analysis.

Some benchmarking efforts have compared the performance of methods [9–12] both adapted from bulk RNA-Seq and developed for microbiome DAA [13, 14]. While some tools exist to guide researchers [15], a general consensus on the best approach is still missing, especially regarding the methods' capability of controlling false discoveries. In this study, we benchmark several statistical models and methods developed for metagenomics [13, 14, 16–18], bulk RNA-Seq [19–21], and, for the first time, scRNA-Seq [7, 8, 22–24] on a collection of manually curated 16S and WMS [25, 26] real data as well as on a comprehensive set of simulations. We include in the comparison several tools that take into account the compositional nature of the data: they achieve this through the use of the DM distribution (*e.g.*, ALDEx2), Multinomial distribution with reference frames (Songbird), or the Centred Log-Ratio (CLR) transformation (*e.g.*, ALDEx2, mixMC). The novelty of our benchmarking efforts is twofold. First, we include in the comparison novel methods recently developed in the scRNA-Seq and metagenomics literatures; second, unlike previous efforts, our conclusions are based on several performance metrics on real data that range from type I error control and goodness of fit to replicability across datasets, concordance among methods, and enrichment for expected DA microbial taxa.

3.3 Results

We benchmarked a total of 18 approaches (Additional file 1: Supplementary Table S2) on 100 real datasets (Additional file 1: Supplementary Table S1), evaluating goodness of fit, type I error control, concordance, and power through (i) reliability of DA results in real data based on enrichment analysis and (ii) specificity and sensitivity using 28,800 simulated datasets (Fig. 3.1; Additional file 2: Supplementary Table S4).

The benchmarked methods include both DA methods specifically proposed in the metagenomics literature and methods proposed in the single-cell and bulk RNA-Seq fields. The manually curated real datasets span a variety of body sites and characteristics (*e.g.*, sequencing depth, alpha and beta diversity). The diversity of the data allowed us to test each method on a variety of circumstances, ranging from very sparse, very diverse datasets, to less sparse, less diverse ones.

We first analyzed 18 16S, 82 WMS, and 28 scRNA-Seq public datasets in order to assess whether scRNA-Seq and metagenomic data are comparable in terms of sparsity. We observed overlap in the fractions of zero counts between the scRNA-Seq, WMS, and 16S, but with scRNA-Seq datasets having a lower distribution of sparsities (ranging from 12 to 75%) as compared to 16S (ranging from 55 to 83%) and WMS datasets (ranging from 35 to 89%) whose distributions of zero frequencies were not significantly different from each other (Wilcoxon test, W=734, p-value=0.377, Additional file 1: Supplementary Fig. S1 a, b). To establish whether the difference between scRNA-Seq and metagenomic data was due to the different number of features and samples, which are intrinsically related to sparsity, we explored the role of library size and experimental protocol (Additional file 1: Sup-

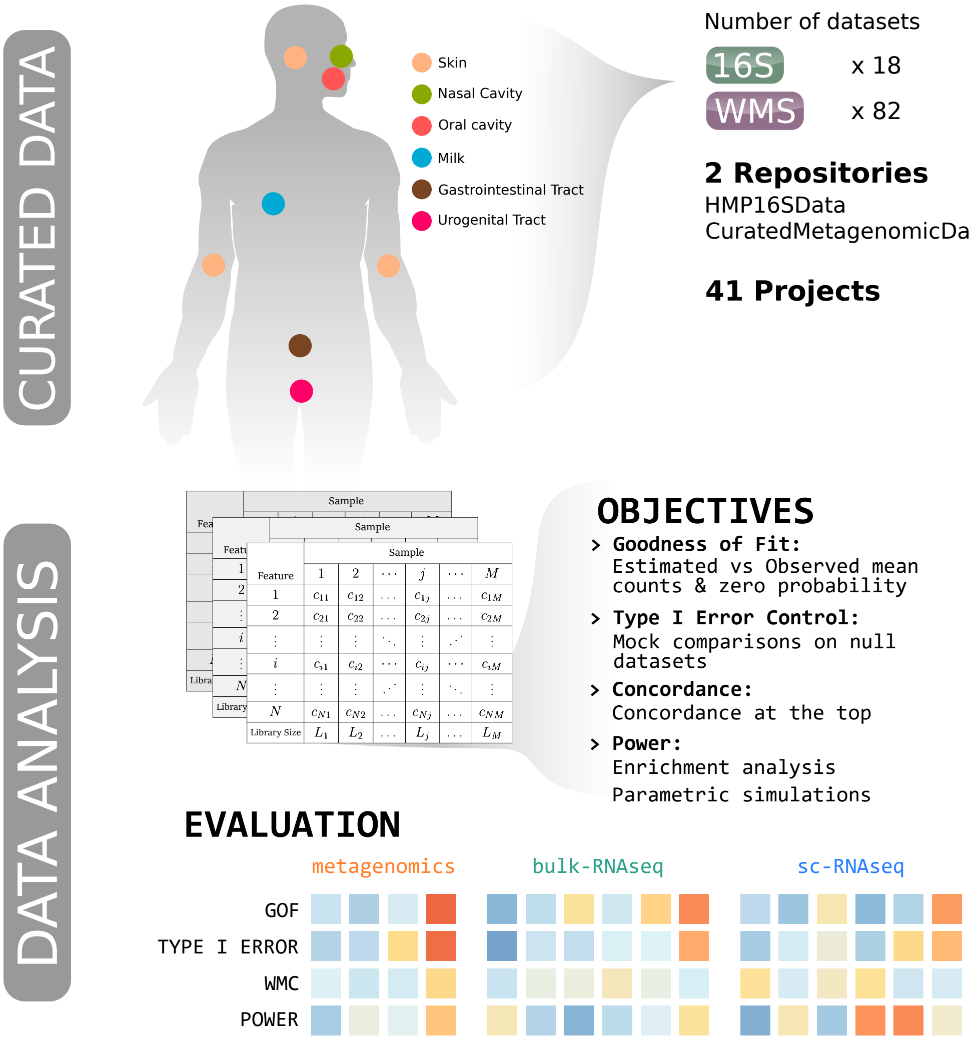


Figure 3.1: Starting from 41 Projects collected in 2 manually curated data repositories (HMP16SData and curatedMetagenomicData Bioconductor packages), 18 16S and 82 WMS datasets were downloaded. Biological samples belonged to several body sites (*e.g.*, oral cavity), body subsites (*e.g.*, tongue dorsum), and conditions (*e.g.*, healthy vs. disease). Feature per sample count tables were used in order to evaluate several objectives: goodness of fit for 5 parametric distributions, type I error control, concordance, and power for 18 differential abundance detection methods. Methods developed for metagenomics, bulk-RNA-Seq, or scRNA-Seq were ranked using empirical evaluations of the above-cited objectives.

plementary Fig. S1 c). scRNA-Seq datasets showed a marked difference in terms of the number of features and sparsity, as they are derived from different experimental protocols. Full-length data (*e.g.*, Smart-seq) are on average sparser than droplet-based data (*e.g.*, Drop-seq) but both are less sparse than 16S and WMS.

These results indicate that metagenomic data are even more sparse than scRNA-Seq, and thus that zero-inflated models designed for scRNA-Seq could, at least in principle, have good performance in a metagenomic context.

3.3.1 Goodness of fit

As different methods rely on different statistical distributions to perform DA analysis, we started our benchmark by assessing the Goodness Of Fit (GOF) of the statistical models underlying each method on the full set of 16S and WMS data. For each model, we evaluated its ability to correctly estimate the mean counts and the probability of observing a zero (Fig. 3.2). We evaluated five distributions: (1) the Negative Binomial (NB) used in edgeR [19] and DeSeq2 [20], (2) the Zero-Inflated Negative Binomial (ZINB) used in ZINB-WaVE [23], (3) the truncated Gaussian Hurdle model of MAST [7], (4) the Zero-Inflated Gaussian (ZIG) mixture model of metagenomeSeq [13], and (5) the Dirichlet-Multinomial (DM) distribution underlying ALDEx2 [14]. The truncated Gaussian Hurdle model was evaluated following two data transformations, the default logarithm of the counts per million (logCPM) and the logarithm of the counts rescaled by the median library size (see the Methods section). Similarly, the ZIG distribution was evaluated considering the scaling factors rescaled by either one thousand (as implemented in the metagenomeSeq Bioconductor package) and by the median scaling factor (as suggested in the original paper). We assessed the goodness of fit for each of these models using the stool samples from the Human Microbiome Project (HMP) as a representative dataset (Fig. 3.2 a–d); all other datasets gave similar results (Additional file

1: Supplementary Fig. S2). A useful feature of this dataset is that a subset of samples was processed both with 16S and WMS and hence can be used to compare the distributional differences of the two data types. Furthermore, this dataset includes only healthy subjects in a narrow age range, providing a good testing ground for covariate-free models.

The NB distribution showed the lowest root mean square error (RMSE, see the Methods section) for the mean count estimation, followed by the ZINB distribution (Fig. 3.2 a, b). This was true for both 16S and WMS data, in most of the considered datasets (Additional file 1: Supplementary Fig. S2). Moreover, for both distributions, the difference between the estimated and observed means was symmetrically distributed around zero, indicating that the models did not systematically under- or overestimate the mean abundances (Fig. 3.2 a, b; Additional file 1: Supplementary Fig. S2). Conversely, the ZIG distribution consistently underestimated the observed means, both for 16S and WMS and independently on the scaling factors (Fig. 3.2 a, b). The Hurdle model was sensitive to the choice of the transformation: rescaling by the median library size rather than by one million reduced the RMSE in both 16S and WMS data (Fig. 3.2 a, b). This was particularly evident in 16S data (Fig. 3.2 a), in which the default logCPM values resulted in a substantial overestimation of the mean count, while the median library size scaling led to underestimation. Given the clear problems with logCPM, we only used the median library size for MAST and the median scaling factor for metagenomeSeq in all subsequent analyses. The DM distribution overestimated observed means for low-mean count features and underestimated observed values for high-mean count features. This overestimation effect was more evident in WMS than in 16S.

Concerning the ability of models to estimate the probability of observing a zero (referred to as zero probability difference, ZPD), we found that Hurdle models provided good estimates of the observed zero proportion for 16S (Fig. 3.2 c) and WMS datasets (Fig. 3.2 d). The NB and ZINB distributions, on the other hand, tended to overestimate the zero probability for features with

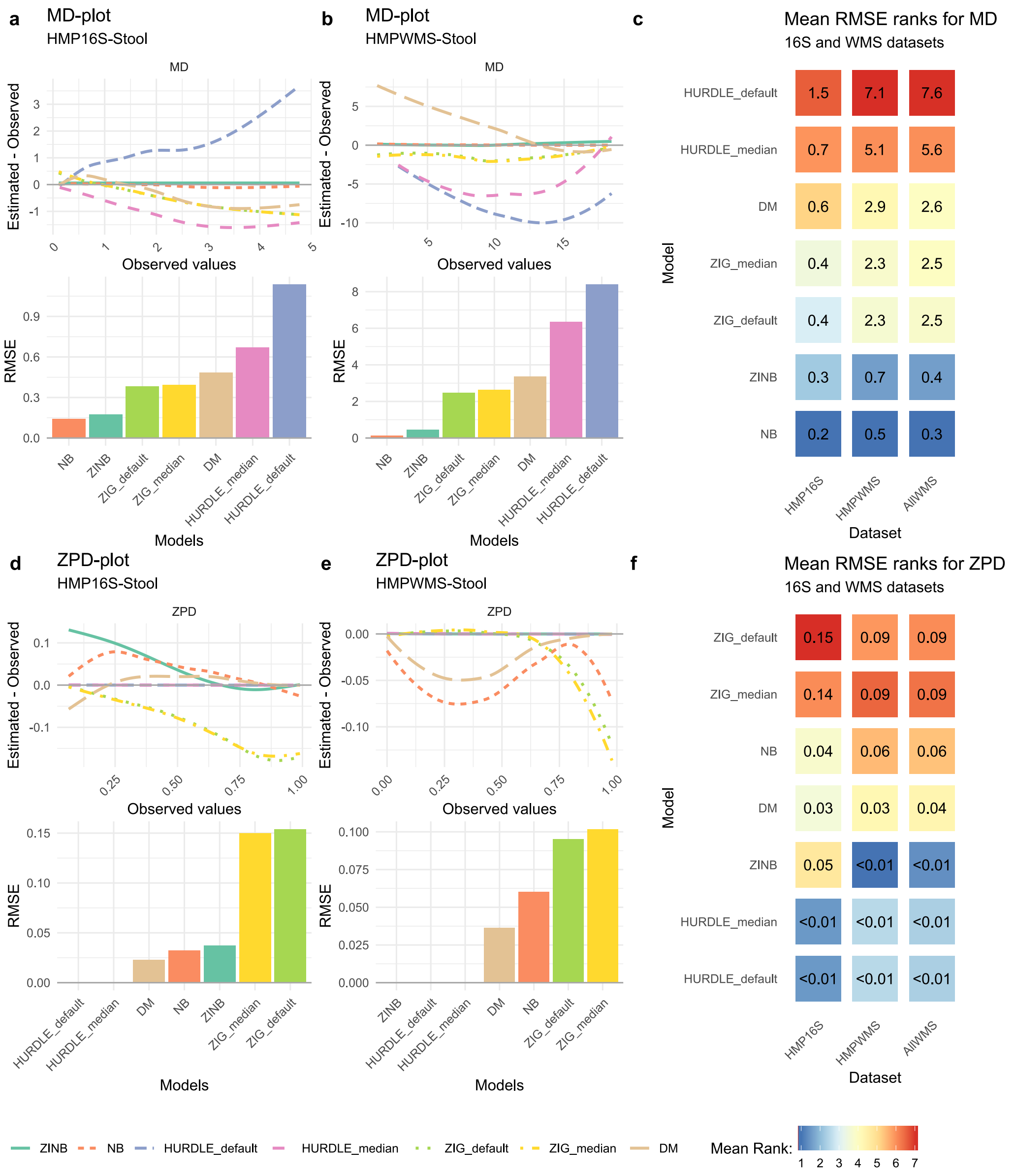


Figure 3.2: **a** Mean-Difference (MD) plot and Root Mean Squared Errors (RMSE) for HMP 16S Stool samples. **b** MD plot and RMSE for HMP WMS Stool samples. **c** Average rank heatmap for MD performances in HMP 16S datasets, HMP WMS datasets and all other WMS datasets. The value inside each tile refers to the average RMSE value on which ranks are computed. **d** Zero Probability-Difference (ZPD) (see the Methods section) plot and RMSE for HMP 16S Stool samples. **e** ZPD plot and RMSE for HMP WMS Stool samples. **f** Average rank heatmap for ZPD performances in HMP 16S datasets, HMP WMS datasets, and all other WMS datasets. The value inside each tile refers to the average RMSE value on which ranks are computed.

a low observed proportion of zero counts in 16S (Fig. 3.2 c). In WMS data, the ZINB distribution perfectly fitted the observed proportion of zeros, while the NB and DM models tended to underestimate it (Fig. 3.2 d). Finally, the ZIG distribution always underestimated the observed proportion of zeros, especially for highly sparse features (Fig. 3.2 c, d).

In summary, across all datasets, the best fitting distributions were the NB and ZINB: the NB distribution seemed to be particularly well-suited for 16S datasets, while the ZINB distribution seemed to better fit WMS data (Fig. 3.2 e). We hypothesise that this is due to the different sequencing depths of the two platforms. In fact, while our 16S datasets have an average of 4891 reads per sample, in WMS, the mean depth is 3.6×10^8 (3×10^8 for HMP). To confirm this observation, we carried out a simulation experiment by down-sampling reads from deep-sequenced WMS samples (rarefaction): while the need for zero inflation seemed to diminish as we got closer to the number of reads typical of the corresponding 16S experiments, the profile did not completely match between approaches (Fig. Additional file 1: Supplementary Fig. S4 b). This suggests that, while sequencing depth is an important contributing factor, it is not enough to completely explain the distributional differences between the two platforms.

3.3.2 Type I error control

We next sought to evaluate type I error rate control of each method, *i.e.*, the probability of the statistical test to call a feature DA when it is not. To do so, we considered mock comparisons between the same biological Stool HMP samples (using the same Random Sample IDentifier (RSID) in both 16S and WMS), in which no true DA is present. Briefly, we randomly assigned each sample to one of two experimental groups and performed DA analysis between these groups, repeating the process 1000 times (see the Methods section for additional details). In this setting, the p-values of a perfect test should be uniformly distributed between 0 and 1 [27] and the False Positive Rate (FPR or observed α), which is the observed proportion of significant tests, should match the nominal value (*e.g.*, $\alpha = 0.05$).

To evaluate the impact of both the normalisation step and the estimation and testing step in bulk RNA-Seq inspired methods, we included in the comparison both edgeR with its default normalisation (TMM), as well as with DESeq2 recommended normalisation ("poscounts," *i.e.*, the geometric mean of the positive counts) and vice versa (see Additional file 1: Supplementary Table S2). Similarly, because the zinbwave observational weights can be used to apply several bulk RNA-Seq methods to single-cell data [24], we have included in the comparison edgeR, DESeq2, and limma-voom with zinbwave weights.

The qq-plots and Kolmogorov-Smirnov (KS) statistics in Fig. 3.3 show that most methods achieved a p-value distribution reasonably close to the expected uniform. Notable exceptions in the 16S experiment were edgeR with TMM normalisation and robust dispersion estimation (edgeR_TMM_robustDisp), metagenomeSeq, and ALDEx2 (Fig. 3.3 a, b). While the former two appeared to employ liberal tests, the latter was conservative in the range of p-values that are typically of interest (0–0.1). In the WMS data, departure from uniformity was observed for metagenomeSeq and edgeR_TMM_robustDisp, and limma_voom_TMM_zinbwave, which employed liberal tests, as well as corncob_LRT, ALDEx2, and scde, which were conservative in the range of interest (Fig. 3.3 c, d). We note that in the context of DA, liberal tests will lead to many false discoveries, while conservative tests will control the type I error at a cost of reduced power, potentially hindering true discoveries.

We next recorded the FPR by each method (by definition all discoveries are false positives in this experiment) and compared it to its expected nominal value. This analysis confirmed the tendencies observed in Fig. 3.3 a, b and c, d. In particular, edgeR_TMM_robustDisp and metagenomeSeq were very liberal in both 16S (Fig. 3.3 e) and WMS data (Fig. 3.3 f); in the case of metagenomeSeq, as much as 30% of the features were deemed DA in the 16S datasets when claiming a nominal FPR of 5% (Fig. 3.3 e).

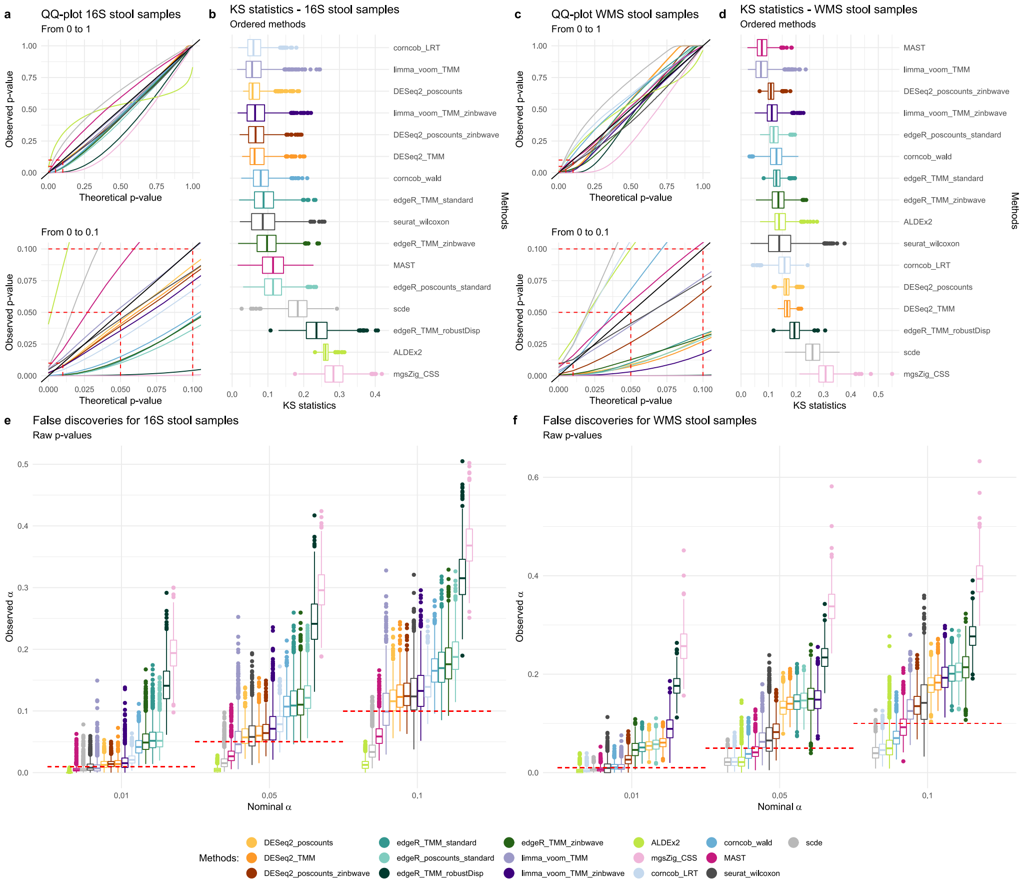


Figure 3.3: **a** Quantile-quantile plot from 0 to 1 and 0 to 0.1 zoom for DA methods in 41 16S HMP stool samples. Average curves for mock comparisons are reported. **b** Kolmogorov-Smirnov statistic boxplots for DA methods in 41 16S HMP stool samples. **c** Quantile-quantile plot from 0 to 1 and 0 to 0.1 zoom for DA methods in 41 WMS HMP stool samples. Average curves for mock comparisons are reported. **d** Kolmogorov-Smirnov statistic boxplots for DA methods in 41 WMS HMP stool samples. **e** Boxplots of the proportion of raw p-values lower than the commonly used thresholds for the nominal α (0.01, 0.05, and 0.1) for 41 16S stool samples. **f** Boxplots of the proportion of raw p-values lower than the commonly used thresholds for the nominal α (0.01, 0.05, and 0.1) for 41 WMS stool samples.

ALDEx2, scde, and MAST, albeit conservative, were able to control type I error. In between these two extremes, edgeR, DESeq2, and limma showed an observed FPR slightly higher than its nominal value. In particular, DESeq2-based methods, limma-voom, and MAST were very close to the nominal FPR for 16S (Fig. 3.3 e), while limma-voom, MAST, and corncob (with Wald test) were the closest in WMS data (Fig. 3.3 f). Of note, corncob seemed slightly conservative in WMS data and slightly liberal in 16S data, with LRT being closer than Wald to the nominal value in 16S (Fig. 3.3 e) and vice versa in WMS data (Fig. 3.3 f). The zinbwave weights showed mixed results: DESeq2 with zinbwave weights was better than the unweighted versions in WMS, while the weights did not help edgeR and limma in controlling the type I error rate. Taken together, these results suggest that the majority of the methods do not control the type I error rate, both in 16S and WMS data, confirming previous findings [10, 12]. However, for most approaches, the observed FPR is only slightly higher than its nominal value, making the practical impact of this result unclear.

3.3.3 Between-method concordance

We measured the ability of each method to produce replicable results in independent data in six datasets [25, 26, 28–30] (Additional file 1: Supplementary Table S3) that showed different α- and β-diversity, as well as different amounts of DA between two experimental conditions (Additional file 1: Supplementary Fig. S5). Each dataset was randomly split in two equally sized subsets and each method was separately applied to each subset. The process was repeated 100 times (see the Methods section for details). To assess the ability of methods to return concordant results from independent samples, we employed the Concordance At the Top (CAT) [31]measure to assess Between Methods Concordance (BMC) by comparing the list of DA features across methods in the subset (ranked by p-value when available or by importance in the case of the songbird and mixMC; see Methods). We used BMC to (i) group methods based on their degree of agreement and (ii) identify those methods sharing the largest amount of discoveries with the majority of the other methods. Although concordance is not a guarantee of validity, it is a requirement of validity, so methods sharing the largest amount of discoveries with the majority of other methods may be more likely to also be producing valid results.

Concordance analysis performed on 16S Tongue Dorsum vs. Stool dataset (Fig. 3.4 a) showed that the methods clustered within two distinct groups: the first comprising all methods that include a TMM normalisation step, songbird, and scde, the second containing all the other approaches (Fig. 3.4 a). Even within the second group, methods segregated by normalisation, as can be seen by the tight clustering of all the methods that include a poscount normalisation step (Fig. 3.4 a). This indicates that, in 16S data, the choice of the normalisation has a pronounced effect on inferential results, even more so than the choice of the statistical test. A similar result was previously observed in bulk RNA-Seq data [32]. The use of observational weights to account for zero inflation did not seem to matter in these data, and in general, scRNA-Seq methods did not agree with each other (Fig. 3.4 a). Similarly, the clustering did not separate compositional and non-compositional methods (Fig. 3.4 a). We noted that metagenomeSeq was not concordant with any other method and that the two corncob approaches formed a tight group, confirming that modelling strategies have more impact than the choice of the test statistics in these data.

A different picture emerged from the analysis of the WMS data (Fig. 3.4 b). Here, methods are clustered by the testing approach. The bottom cluster comprised the bulk RNA-Seq methods with the inclusion of the Wilcoxon nonparametric approach, metagenomeSeq, and mixMC. The middle cluster consisted of the zinbwave methods and ALDEx2. The top cluster comprised MAST, corncob, scde, and songbird. Overall, mixMC and the methods based on NB generalised linear models showed the highest BMC values. When observational weights were added to those models, the BMC decreased, but still a good level of concordance was observed with their

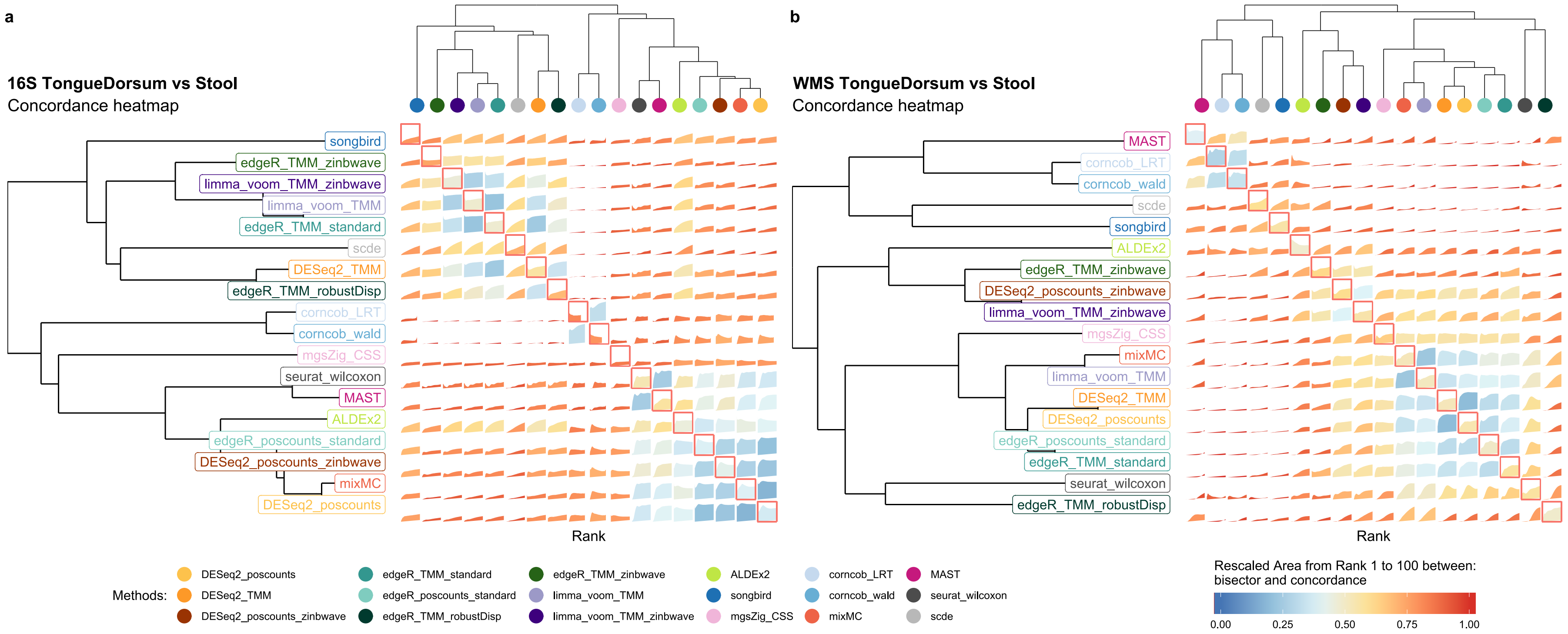


Figure 3.4: **a** Between Methods Concordance (BMC) and Within Method Concordance (WMC) (main diagonal) averaged values from rank 1 to 100 for DA methods evaluated in replicated 16S Tongue Dorsum vs. Stool comparisons. **b** BMC and WMC (main diagonal) averaged values from rank 1 to 100 for DA methods evaluated in replicated WMS Tongue Dorsum vs. Stool comparisons.

respective unweighted version.

We noted that the BMC is highly dataset-specific and depends on the amount of DA between the compared groups. Indeed, BMC decreased with decreased beta diversity of the dataset, and the role of normalisation became less clear (Additional file 1: Supplementary Fig. S6).

3.3.4 Within-method concordance

The CAT metric was used again for assessing the Within Method Concordance (WMC), *i.e.*, the amount of concordance of the results of each method on the two random subsets.

WMC was clearly dataset-dependent, showing high levels of concordance in datasets with a high differential signal (*e.g.*, tongue vs. stool, Fig. 3.5 a) and low concordance in datasets with a low differential signal (*e.g.*, supragingival vs. subgingival, Fig. 3.5 e). Overall, the replicability of results in WMS studies was slightly higher than that of 16S datasets.

In terms of method comparison, corncob showed high levels of concordance in WMS datasets but lower concordance in all 16S datasets (Fig. 3.5). Similarly, songbird showed the highest concordance in mid (Fig. 3.5 d) and low (Fig. 3.5 f) diversity WMS datasets but did not perform well in 16S (especially for the highly diverse TongueDorsum vs. Stool comparison; Fig. 3.5 a).

The addition of zinbwave weights to edgeR, DESeq2, and limma-voom did not always help: it was sometimes detrimental, *e.g.*, for edgeR in the schizophrenia dataset (Fig. 3.5 d) and sometimes led to an improvement in replicability, *e.g.*, for limma-voom in the Tongue Dorsum vs. Stool dataset (Fig. 3.5 a). The schizophrenia dataset had the lowest sample size among all the datasets evaluated, suggesting that sample size may play an important role in estimating zinbwave weights.

While this analysis confirmed the unsatisfactory performance of metagenomeSeq (Fig. 3.5 a, b, and f), ALDEx2, which was very conservative in terms of type I error control (Fig. 3.3), showed overall good performance, with the notable exception of the high-diversity WMS dataset (Fig. 3.5 b), for which it was the worst performing method. To sum up, the highest concordance was measured, in all WMS datasets, by the corncob-based and songbird methods, while RNA-Seq methods performed better in 16S datasets, confirming that the two platforms yield substantially different data. mixMC was the only method that never showed poor concordance regardless of the technology or of the diversity of the compared groups.

Taken together, these analyses suggest that both BMC and WMC are highly dependent on the amount of DA observed in the dataset: higher DA leads to a higher concordance. Moreover, WMC was similar among the compared methods, indicating that the replicability of the DA results depends more on the strength of DA than on the choice of the method (Fig. 3.5).

3.3.5 Enrichment analysis

While mock comparisons and random splits allowed us to evaluate model fit and concordance, these analyses do not assess the correctness of the discoveries. Even the method with the highest WMC could nonetheless consistently identify false positive DA taxa.

While the lack of ground truth makes it challenging to assess the validity of DA results in real data, enrichment analysis [33] can provide an alternative solution to rank methods in terms of their ability to identify as significant taxa that are known to be differentially abundant between two groups.

Here, we leveraged the peculiar environment of the gingival site: the supragingival biofilm is directly exposed to the open atmosphere of the oral cavity, favoring the growth of aerobic species. In the subgingival biofilm, however, the atmospheric conditions gradually become strict anaerobic, favoring the growth of anaerobic species [34]. From the comparison of the two sites, we thus expected to find an abundance of aerobic microbes in the supragingival plaque and of anaerobic bacteria in the subgingival plaque. DA analysis should reflect this difference by finding an enrichment of aerobic (anaerobic) bacteria among the DA taxa with a positive (negative) log-fold-change.

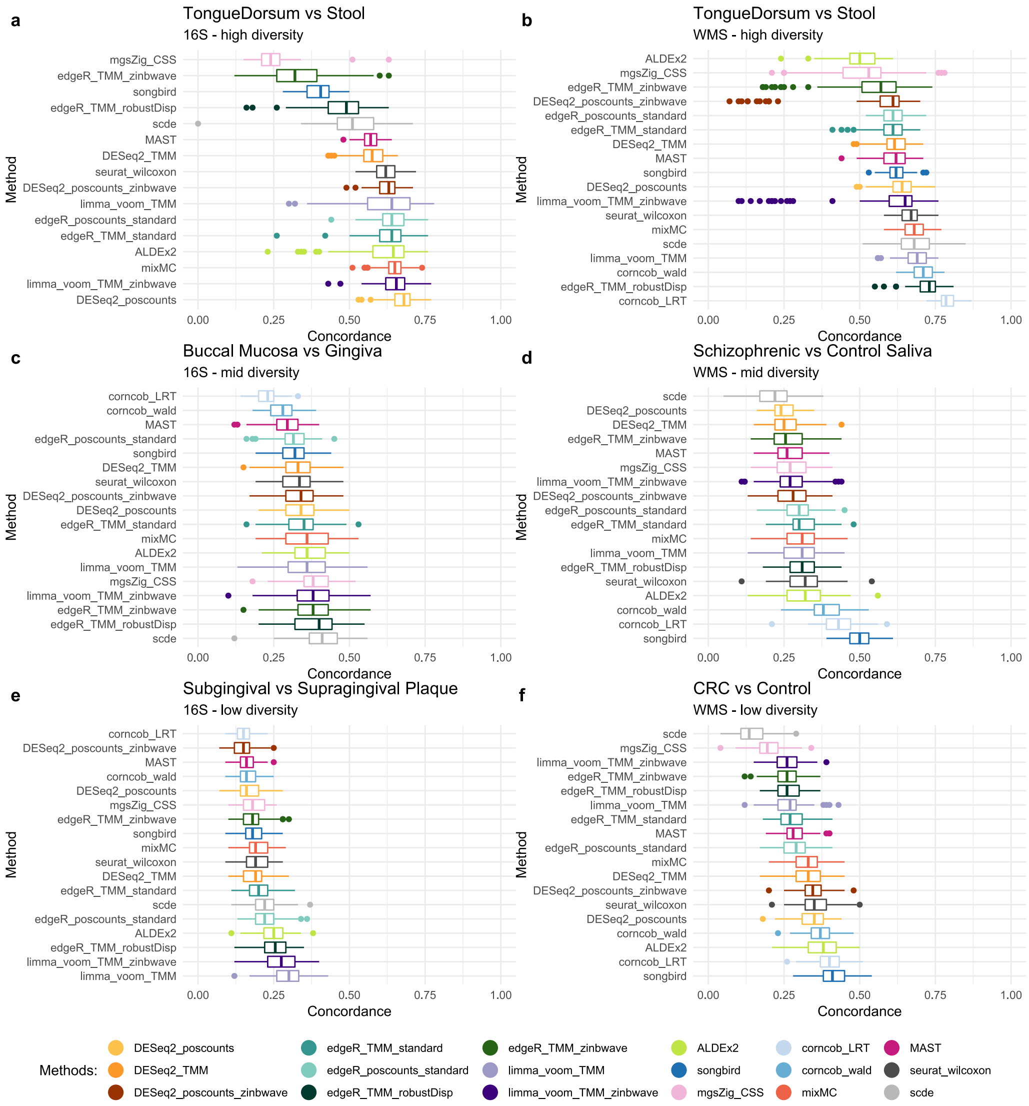


Figure 3.5: a Boxplot of WMC on high diversity 16S datasets: Tongue Dorsum vs. Stool. Due to the high sparsity and low sample size of the dataset, the Concordance At the Top (CAT) at rank 100 was not computable for corncob methods: it was possible to estimate the model only for a few features. **b** Boxplot of WMC on high diversity WMS datasets: Tongue Dorsum vs. Stool. **c** Boxplot of WMC on mid diversity 16S datasets: Buccal Mucosa vs. Attached Keratinised Gingiva. **d** Boxplot of WMC on mid diversity WMS datasets: Schizophrenic vs. Healthy Control saliva samples. **e** Boxplot of WMC on low diversity 16S datasets: Supragingival vs. Subgingival plaque. **f** Boxplot of WMC on low diversity WMS datasets: Colon Rectal Cancer patient vs. Healthy Control stool samples.

We tested this hypothesis by comparing 38 16S supragingival and subgingival samples (for a total of 76 samples) from the HMP (see the Methods section for details). The DA methods showed a wide range of power, identifying 2 (ALDEx2) through 305 (metagenomeSeq) significantly DA taxa (Fig. 3.6 a). However, almost all methods correctly found an enrichment of anaerobic microbes among the taxa under-abundant in supragingival and an enrichment of aerobic microbes among the over-abundant ones (Fig. 3.6 a; Fig. Additional file 1: Supplementary Fig. S7). Furthermore, as expected, no enrichment was found for facultative anaerobic microbes, which are able to switch between aerobic and anaerobic respiration (Fig. 3.6 a).

Although most methods performed well, scde, ALDEx2, and MAST had too low power to detect any enrichment (at 0.05 significance level), as their number of identified DA taxa was very low (Fig. 3.6 a). This analysis confirmed the conservative behavior of these methods in 16S data (Fig. 3.3 e). Finally, metagenomeSeq and edgeR with robust dispersion estimation found the correct enrichments, but they also identified many anaerobic taxa with a positive log-fold-change (Fig. 3.6 a), confirming their liberal tendencies (Fig. 3.3 e). Overall, these results were confirmed by the same comparison in WMS data (Fig. Additional file 1: Supplementary Fig. S8), but the reduced sample size of our WMS dataset resulted in a reduced power to detect DA for all methods (see the Methods section).

To explore the ability of each method to correctly rank the DA taxa independently of its power, we tested whether over-abundant aerobic taxa and under-abundant anaerobic taxa were more likely to be ranked at the top when ranking taxa by each method's test statistics. To do so, we considered the top K taxa (with K from 1 to 20%; see the Methods section) and computed the difference between putative true positives (TP; over-abundant aerobic taxa and under-abundant anaerobic taxa) and putative false positives (FP; under-abundant aerobic taxa and over-abundant anaerobic taxa; Fig. 3.6 b). Reassuringly, increasing the threshold resulted in a larger difference between TP and FP for most methods (Fig. 3.6 b), indicating that

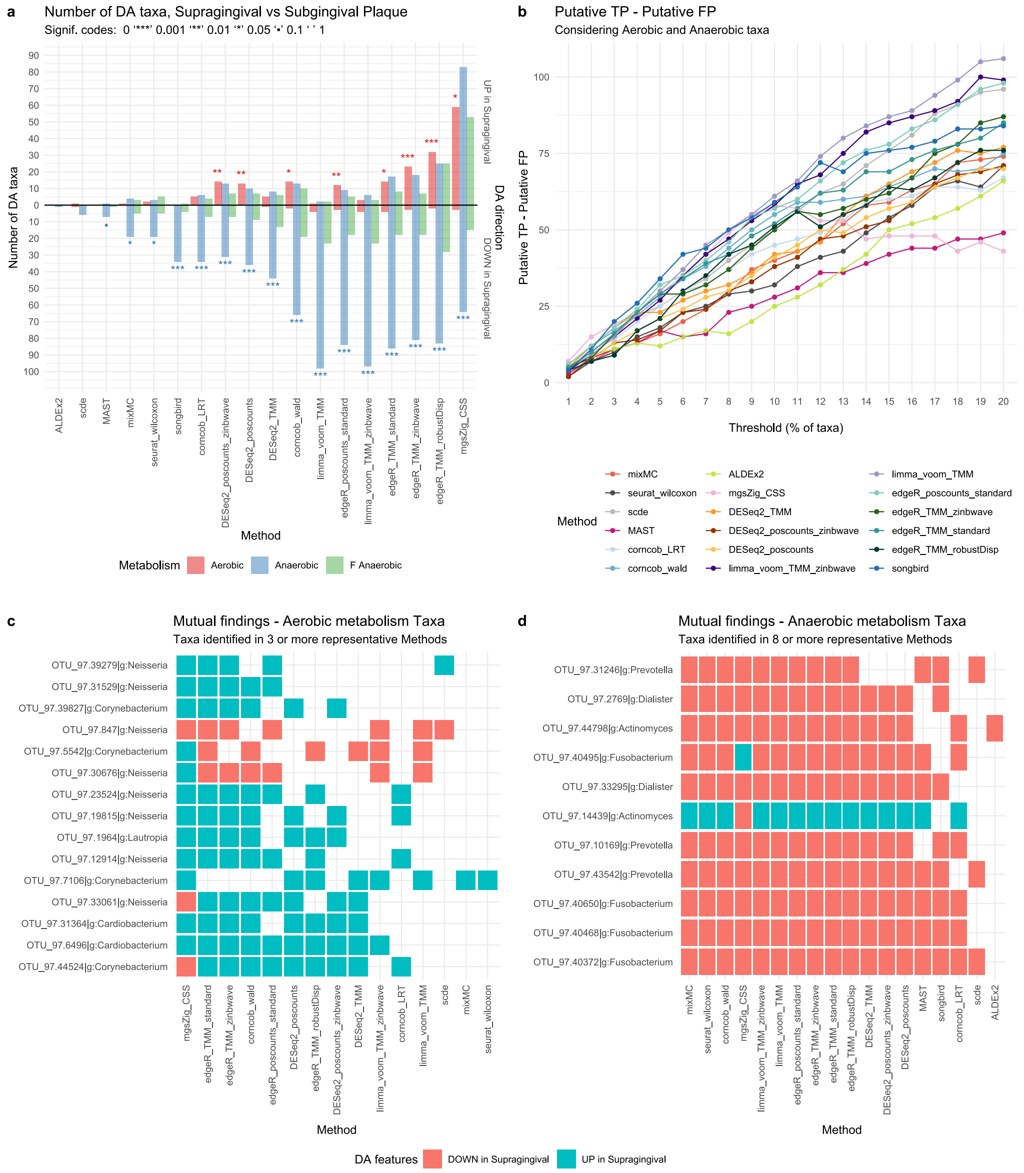


Figure 3.6: 38vs38 Supragingival vs. Subgingival Plaque 16S samples. **a** Barplot of the enrichment tests performed on the DA taxa found by each method using an adjusted p-value of 0.1 as threshold for significance (top 10% ranked taxa for songbird). Each bar represents the number of findings, UP-Abundant in Supragingival or DOWN-Abundant in Supragingival Plaque compared to Subgingival Plaque, regarding aerobic, anaerobic, and facultative anaerobic taxa metabolism. A Fisher exact test was performed to establish the enrichment significance represented with signif. codes. **b** Difference between putative true positives (TP) and putative false positives (FP) (y-axis) for several significance thresholds (x-axis). Each threshold represents the top percent ranked taxa, using the ordered raw p-value lists as reference (loading values for mixMC and differentials for songbird). **c** Aerobic metabolism taxa mutually found by 3 or more methods from the subset of the representative methods. **d** Anaerobic metabolism taxa mutually found by 8 or more methods from the subset of the representative methods.

independently of their power, most methods are able to highly rank true positive taxa. This becomes particularly important for the methods with a low power, suggesting that in these cases a more liberal p-value threshold may be applied. However, metagenomeSeq's performance deteriorates after the 10% threshold, suggesting that this method starts to identify more false positives (Fig. 3.6 b): this is particularly problematic since its adjusted p-value threshold identifies 34% of DA taxa. Among the other methods, MAST and ALDEx2 showed a consistently lower performance, while limma-voom was the best performer at permissive thresholds, and songbird was the best performer at strict thresholds (Fig. 3.6 b).

The majority of aerobic taxa were found DA by just a handful of methods, with only 15 out of 75 unique aerobic taxa identified as DA by 3 or more representative methods (see Methods; Fig. 3.6 c). All of them belonged to the genera *Cardiobacterium*, *Neisseria*, *Lautropia*, *Corynebacterium*, found to be among the most prevalent genera in supragingival plaques in an independent study [35]. On the other hand, 57 out of 161 unique anaerobic taxa were found DA by 5 or more representative methods (see Methods; Fig. 3.6 d; Additional file 1: Supplementary Fig. S9). Among these, *Fusobacterium*, *Prevotella*, *Porphyromonas*, *Treponema* are known to be abundant in the subgingival plaque [36, 37]. Despite the small sample size for WMS data (n=10), enrichment and DA analysis were largely consistent, including several strains of *Neisseria* and several species of *Treponema* found to be DA (Additional file 1: Supplementary Fig. S8 c, d). Overall, similar methods tended to identify a higher number of mutual taxa, confirming our previous findings in the concordance analysis (Additional file 1: Supplementary Fig. S6) and highlighting how different statistical test and normalisation approaches have a big impact on the identified DA.

3.3.6 Parametric simulations

Given the results of our GOF analysis (Fig. 3.2), we only used the NB and ZINB distributions to simulate 7200 and 19,200 scenarios, respectively, mimicking both 16S and WMS data. The simulated data differed in sample size, proportion of DA features, effect size, proportion of zeros, and whether there was an interaction between the amount of zeros and DA (sparsity effect, see the Methods section for details).

In general, we found that the results confirmed our expectations that methods perform well on simulated data that conforms to the assumptions of the method (Additional file 2: Supplementary Fig. S11). The parametric distribution that generated the data had a great influence on the method performances and the methods that rely on NB and ZINB generally performed better compared to the other methods. As an example, MAST, which showed overall good results in real data, did not behave in simulations, partly because of the misspecified model with respect to the data generating distribution.

As expected, all methods' performances increased as the sample size and/or the effect size increased. Confirming our real data results, we finally observed that metagenomeSeq, scde, and edgeR-robust performed poorly. Details on the simulated data analysis can be found in Additional file 2.

3.4 Discussion

We investigated different theoretical and practical issues related to the analysis of metagenomic data. The main objective of the study was to compare several DA detection methods adapted from bulk RNA-Seq, scRNA-Seq, or specifically developed for metagenomics. Unsurprisingly, there is no single method that outperforms all others in all the tested scenarios. As is often the case in high-throughput biology, the results are data-dependent and careful data exploration is needed to make an informed decision on which workflow to apply to a specific dataset. We recommend applying our exploratory analysis framework to gain useful insights about the assumptions of each method and their suitability given the data at hand. To this end, we provide all the R scripts to easily reproduce the analyses of this paper on any given dataset (see the Availability of data and materials section).

Our GOF analysis highlighted the advantages of using count models for the analysis of metagenomics data. The goodness of fit of zero-inflated models seemed dependent on whether the data come from 16S or WMS experiments. The difference between these two approaches translates to different count data structures: while for WMS many features are characterised by a clearly visible bimodal distribution (with a point mass at zero and another mass, quite far from zero, at the second positive mode), 16S data are as sparse as or even more sparse than WMS data, presenting for many features a less clearly bimodal distribution (Additional file 1: Supplementary Fig. S4 a). This difference is probably due to a mix of factors: primarily sequencing depth, but also different taxonomic classification between technologies (entire metagenomic sequences versus clusters of similar amplicon sequences), bioinformatics methods for data preprocessing, etc. However, comparing the distribution of several genera on the same samples assayed with 16S and WMS, we observed that many of the zero counts were consistent across platforms and very different read depths, suggesting that many observed zeros are biological and not technical in nature (Fig. Additional file 1: Supplementary Fig. S4 a). Further analyses are needed to inspect this unsolved issue and related efforts are ongoing in the scRNA-Seq literature, where similar differences are observed between protocols with and without unique molecular identifiers [38, 39].

Metagenomic data are inherently compositional, but whether incorporating compositionality into the statistical model provides benefits greater than the tradeoffs they may introduce is a debated topic in the literature [9, 13, 40–42]. While other data resulting from sequencing are also compositional, some in the microbiome data analysis community believe that compositionality has greater relevance in metagenomics due to the potential presence of dominant microbes. Here, we found that compositional methods did not outperform non-compositional methods designed for count data, indicating that their benefits did not outweigh the drawbacks they may introduce.

This can be explained by two considerations. First, some compositional methods assume that the data arise from a multinomial distribution, with n trials (reads) and a vector p indicating the probability of the reads to be mapped to each taxon. In metagenomic studies, we have a large n (number of sequenced reads) and small p (since there are many taxa, the probability of each read to map to any given taxon is small). In this setting, the Poisson distribution is a good approximation of the multinomial. Similarly, the NB is a good approximation of the DM [31]. Secondly, some normalisations, such as the geometric mean method implemented in DESeq2 or the trimmed mean of M-values of edgeR, have size factors mathematically equivalent or very similar to the CLR proposed by Aitchison [40, 43]. This has been shown to reduce the impact of compositionality on DA results [44]. We did not test the ANCOM package [45] because it was too slow for assessment. However, we included three recent analysis methods that address compositionality, namely, ALDEx2, songbird, and mixMC. This allowed us to perform an adequate assessment of compositional vs. non-compositional approaches. Similarly, multivariate methods, such as songbird and mixMC, did not outperform methods based on univariate tests, suggesting that these simpler approaches are often sufficient to detect the most relevant biological signals.

The lack of ground truth makes the assessment of DA correctness very challenging. However, we can rely on mock datasets, within-method concordance, and enrichment analysis to obtain a principled ranking of method performances (Fig. 3.7). Although each analysis by itself does not imply correctness, taken together these assessments are a good proxy to evaluate methods performances in terms of their ability to limit the amount of false discoveries, give replicable results in datasets contrasting the same groups, and identify as significant the taxa that are expected to be DA.

The parametric simulation framework is useful to inspect how individual characteristics of the data-generating distribution impact the sensitivity and specificity of the methods. As the entire analysis was supported by

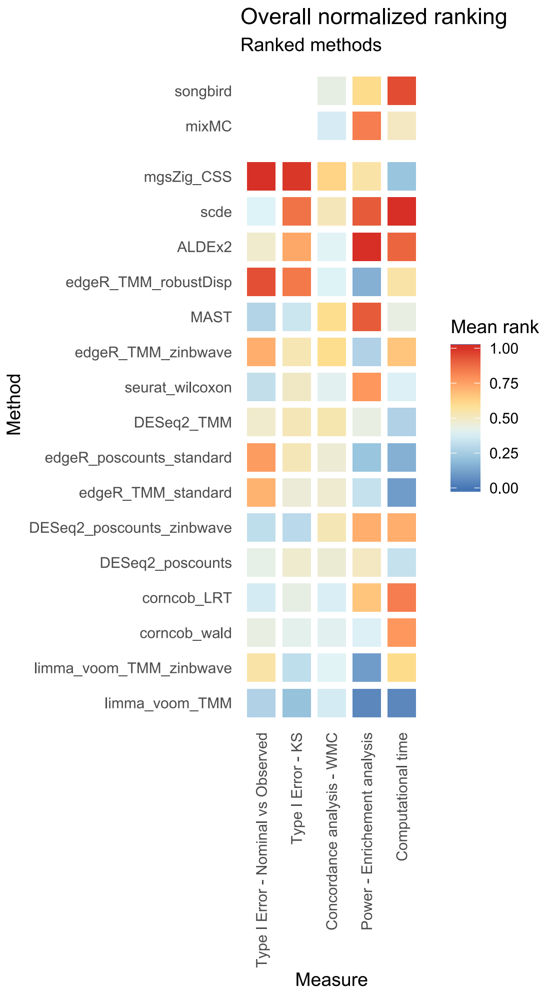


Figure 3.7: Overall method ranking based on 5 evaluation criteria. Average normalised ranks range from 0 to 1, lower values correspond to better performances. The type I error columns are based on the analysis of the 1000 mock comparisons from HMP 16S and WMS Stool datasets; the concordance analysis column is based on the average WMC values across the 100 random subset comparisons for each of the 6 datasets used. The power enrichment analysis and computational time columns are based on the Supragingival vs. Subgingival Plaque 16S dataset evaluations. Each method's ordering is computed using the first 4 columns. Since the type I error analysis was not available for songbird and mixMC, these methods were not included in the final ranking.

real data, we decided to focus only on a very simple but easily reproducible implementation of the NB and ZINB distributions for the simulations. The choice was justified by our GOF analysis on real datasets. Unsurprisingly, the sample size and the effect size were the characteristics that had the most impact on method performances. This translates into an evident suggestion for experimental design: large sample sizes are needed to detect low effect sizes. Our simulation framework can in principle be used for power calculations in the context of DA analysis.

In the 16S dataset used for the enrichment analysis, with a total of 76 samples and almost 900 unique taxa, the most time-consuming methods were scde and songbird with more than 5 minutes needed to identify DA taxa. ALDEx2 and corncob-based methods took about 40s, zinbwave-weighted methods took approximately 20s while mixMC, MAST and seurat_wilcoxon around 10s. DESeq2 and edgeR were under the 10s with limma-voom which was the fastest method taking less than a second (Fig. 3.7). A consistent ranking was found in simulated datasets with interesting changes determined by different sample-sizes (Additional file 2: Supplementary Table S5 and Supplementary Fig. S10).

3.5 Conclusions

As already noted in recent publications [10–12], the perfect method does not exist. However, taken together, our analyses suggested that limma-voom, corncob, and DESeq2 showed the most consistent performance across all datasets, metagenomeSeq had the worst performance, and scde and ALDEx2 suffered from low power (Fig. 3.7). Among compositional data analysis methods, songbird showed a greater ability to identify the correct taxa in the enrichment analysis, while mixMC had a better within-method concordance.

In general, we recommend a careful exploratory data analysis and we present a framework that can help scientists make an informed choice in a dataset-specific manner. We did not find evidence that bespoke differential abundance methods outperform methods developed for the differential expression analysis of RNA-Seq data. However, our analyses also suggested that further research is required to overcome the limitations of currently available methods: in this respect, new directions in DA method development, *e.g.*, leveraging the phylogenetic tree [46, 47], log-contrast models [48], or compositional balances [49] are promising, but efforts to make these methods scalable are needed.

3.6 Methods

3.6.1 Datasets

The HMP16Sdata [25] (v1.2.0) and curatedMetagnomicData [26] (v1.12.3) Bioconductor packages were used to download high-quality, uniformly processed, and manually annotated human microbiome profiles for thousands of people, using 16S and WMS technologies, respectively. HMP16SData comprises the collection of 16S data from the Human Microbiome Project, while curatedMetagnomicData contains data from several projects. Gene-level counts for a collection of public scRNA-Seq datasets were downloaded from the scRNA-Seq (v1.99.8) Bioconductor package.

While the latter datasets are used only for a comparison between technologies, the former are widely used for all the analyses. A complete index with dataset usage is reported in Additional file 1: Supplementary Table S1.

Phyloseq objects were obtained from the HMP16SData and curatedMetagenomicData packages using the function *as_ phyloseq()* and setting the *bugs.as.phyloseq=TRUE* argument, respectively. The *otu_ table* and *sample_ data* slots of the phyloseq objects that contain, respectively, the taxa count table and the metadata associated to each sample were used for all downstream analyses. For the WMS datasets, absolute raw count data were estimated from the metaPhlAn2-produced relative count data by multiplying the columns of the *ExpressionSet* data by the number of reads for each sample, as found in the *pData* column "*number_reads*" (*counts=TRUE* argument).

HMP16SData was split by body subsite in order to obtain 18 separated datasets. Stool and Tongue Dorsum datasets were selected for example purposes thanks to their high sample size. The same was done on curatedMetagenomicData HMP dataset, obtaining 9 datasets. Moreover, for the evaluation of type I error control, 41 stool samples with equal RSID, in both 16S and WMS, were used to compare DA methods. For each research project, curatedMetagenomicData was split by body site and treatment or disease condition, in order to create homogeneous sample datasets. A total of 82 WMS datasets were created.

A total of 100 datasets were evaluated; however, for the CAT analysis, datasets not split by condition or body subsite were evaluated (*e.g.*, Tongue Dorsum vs. Stool in HMP, 2012 for both 16S and WMS).

To consider the complexity and the variety of several experimental scenarios, an attempt to select a wide variety of datasets for the analysis was done. The datasets were chosen based on several criteria: sample size, homogeneity of the samples, or availability of the same subjects (identified by RSID) assayed by both technologies.

3.6.2 Statistical models

The following distributions were fitted to each dataset, either by directly modelling the read counts or by first applying a logarithmic transformation:

- Negative Binomial (NB) model, as implemented in the edgeR (v3.24.3) Bioconductor package (on read counts);
- Zero-Inflated Negative Binomial (ZINB), as implemented in the zinbwave (v1.4.2) Bioconductor package (on read counts);
- Truncated Gaussian hurdle model, as implemented in the MAST (v1.8.2) Bioconductor package (on log count);
- Zero-Inflated Gaussian (ZIG), as implemented in the metagenomeSeq (v1.24.1) Bioconductor package (on log count).

- Dirichlet-Multinomial (DM), as implemented in the MGLM (v0.2.0) CRAN R package.

Negative Binomial (NB)

The edgeR Bioconductor package was used to implement the NB model. In particular, normalisation factors were calculated with the Trimmed Mean of M-values (TMM) normalisation [50] using the *calcNormFactors* function; common, trended, and tagwise dispersions were estimated by *estimateDisp*, and a NB generalised log-linear model was fit to the read counts of each feature, using the *glmFit* function.

Zero-Inflated Negative Binomial (ZINB)

The zinbwave Bioconductor package was used to implement the ZINB model. We fitted a ZINB distribution using the *zinbFit* function. As explained in the original paper, the method can account for various known and unknown technical and biological effects [23]. However, to avoid giving unfair advantages to this method, we did not include any latent factor in the model (*K=0*). We estimated a common dispersion for all features (*common_dispersion=TRUE*) and we set the likelihood penalisation parameter epsilon to 1e10 (within the recommended set of values [24]).

Truncated Gaussian Hurdle model

We used the implementation of the MAST Bioconductor package. After a $\log_2$ transformation of the reascaled counts with a pseudocount of 1, a zero-truncated Gaussian distribution was modelled through generalised regression on positive counts, while a logistic regression modelled feature expression/abundance rate. As suggested in the MAST paper [7], cell detection rate (CDR) which is computed as the proportion of positive count features for each sample, was added as a covariate in the discrete and continuous model matrices as a normalisation factor.

Zero-Inflated Gaussian (ZIG)

The metagenomeSeq Bioconductor package was used to implement a ZIG model for log2 transformed counts with a pseudocount of 1, rescaled by the median of all normalisation factors or by 1e03 which gives the interpretation of "count per thousand" to the offsets. The *CumNormStat* and *CumNorm* functions were used to perform Cumulative Sum Scaling (CSS) normalisation, which accounts for specific data characteristics. Normalisation factors were included in the regression through the *fitZig* function.

Note that both MAST and metagenomeSeq were applied to the normalised, log-transformed data. We evaluated both models, using their default scale factor $\log_2\left(\frac{counts \cdot 10^6}{libSize} + 1\right)$ for MAST and $\log_2\left(\frac{normFacts}{1000} + 1\right)$ for metagenomeSeq, as well as by rescaling the data to the median library size [13], $\log_2\left(\frac{counts \cdot median(libSize)}{libSize} + 1\right)$ and $\log_2\left(\frac{normFacts}{median(normFacts)}\right)$, respectively.

Dirichlet-Multinomial (DM)

The MGLM package was used to fit a DM regression model for counts. The *MGLMreg* function with *dist="DM"* allowed the implementation of the above model and the estimation of the parameter values.

3.6.3 Goodness of fit

To evaluate the goodness of fit of the models, we computed the Mean-Differences (MD) between the estimated and observed values for several datasets.

For each model, we evaluated two distinct aspects: its ability to correctly estimate the mean counts (plotted in logarithmic scale with a pseudo-count of 1) and its ability to correctly estimate the probability of observing a zero, computed as the difference between the probability of observing a zero count according to the model and the observed zero frequencies (Zero Probability-Difference, ZPD). We summarised the results by computing the Root Mean Squared Errors (RMSE) of the two estimators. The lower the RMSE, the better the fit of the model.

This analysis was repeated for 100 datasets available in HMP16SData and curatedMetagenomicData (Additional file 1: Supplementary Table S1 and Supplementary Fig. S2).

Assuming homogeneity between samples inside the same body subsite or study condition, we specified a model consisting of only an intercept or including a normalisation covariate.

3.6.4 Differential abundance detection methods

DESeq2

The DESeq2 (v1.22.2) Bioconductor package fits a NB model for count data. DESeq2 default data normalisation is the so-called Relative Log Expression (RLE) based on scaling each sample by the median ratio of the sample counts over the geometric mean counts across samples. As 16S and WMS data sparsity may lead to a geometric mean of zero, it is replaced by n^{th} root of the product of the non-zero counts (which is the geometric mean of the positive count values) as proposed in the phyloseq package [51] and implemented in the DESeq2 *estimateSizeFactors* function with option *type="poscounts"*. We also tested DESeq2 with TMM normalisation (see below). As proposed in [24], observational weights were supplied in the weights slot of the *DESeqDataSet* class object to account for zero inflation. Observational weights were computed by the *computeObservationalWeights* function of the zinbwave package. To test for DA, we used a Likelihood Ratio Test (LRT) to compare the reduced model (intercept only) to the full model with intercept and group variable. The p-values were adjusted for multiple testing via the Benjamini-Hochberg (BH) procedure. Some p-values were set to *NA* via the *cooksCutoff* argument that prevents rare or outlier features from being tested.

edgeR

The edgeR Bioconductor package fits a NB distribution, similarly to DESeq2. The two approaches differ mainly in the normalisation, dispersion parameter estimation, and default statistical test. We examined different procedures by varying the normalisation and the dispersion parameter estimation: edgeR_TMM_standard involves TMM normalisation and tagwise dispersion estimation through the *calcNormFactors* and *estimateDisp* functions, respectively (with default values). Analogously to DESeq2, "poscounts" normalisation was used in addition to TMM in edgeR_poscounts_standard to investigate the normalisation impact. We also evaluated the impact of employing a robust dispersion estimation, accompanied with a quasi-likelihood F test through the *estimateGLMRobustDisp* and *glmQLFit* functions respectively (edgeR_TMM_robustDisp). As with DESeq2, zinbwave observational weights were included in the *weights* slot of the *DGEList* object in edgeR_TMM_zinbwave to account for zero inflation, through a weighted F test. BH correction was used to adjust p-values for multiple testing.

Limma-voom

The limma Bioconductor package (v3.38.3) includes a *voom* function that (i) transforms previously normalised counts to logCPM, (ii) estimates a mean-variance relationship, and (iii) uses this to compute appropriate observational-level weights [21]. To adapt the limma-voom framework to zero-inflation, zinbwave weights have been multiplied by *voom* weights as done previously [24]. The residual degrees of freedom of the linear model were adjusted before the empirical Bayes variance shrinkage and were propagated to the moderated statistical tests. BH correction method was used to correct p-values.

ALDEx2

ALDEx2 is a Bioconductor package (v1.14.1) that uses a DM model to infer abundance from counts [14]. The *aldex* method infers biological and sampling variation to calculate the expected false discovery rate, given the variation, based on several tests. Technical variation within each sample is estimated using Monte-Carlo draws from the Dirichlet distribution. This distribution maintains the proportional nature of the data while scale-invariance and sub-compositionally coherence of data is ensured by CLR. This removes the need for a between-sample normalisation step. In order to obtain symmetric CLRs, the *iqlr* argument is applied, which takes, as the denominator of the log-ratio, the geometric mean of those features with variance calculated from the CLR between the first and the third quantile. Statistical testing is done through Wilcoxon rank sum test, even if Welch's t, Kruskal-Wallis, generalised linear models, and correlation tests were available. BH correction method was used to correct the p-values for multiple testing.

metagenomeSeq

metagenomeSeq is a Bioconductor package designed to address the effects of both normalisation and under-sampling of microbial communities on disease association detection and testing feature correlations. The underlying statistical distribution for $\log_2(count + 1)$ is assumed to be a ZIG mixture model. The mixture parameter is modelled through a logistic regression depending on library sizes, while the Gaussian part of the model is a generalised linear model with a sample-specific intercept which represent the sample baseline, a sample-specific offset computed by CSS normalisation and another parameter which represents the experimental group of the sample. We opted for the implementation suggested in the original publication [13], where CSS scaling factors are divided by the median of all the scaling factors instead of dividing them by 1000 (as done in the Bioconductor package). An Expectation-Maximisation algorithm is performed by the *fitZig* function to estimate all the parameters. An empirical Bayes approach is used for variance estimation and a moderated t test is performed to identify differentially abundant features between conditions. BH correction method was used to account for multiple testing.

Corncob

corncob is an R package (v0.1.0 [52]) for the differential abundance and differential variability analysis of microbiome data [17]. Specifically, corncob is designed to account for the challenges of modelling sequencing data from microbial abundance studies. It is based on a hierarchical model in which the latent relative abundance of each taxon is modelled as a beta distribution, and the observed absolute presence of a taxon is modelled as a binomial process with the previously specified beta as the probability of success. This hierarchical structure gives flexibility to the method, which can account for changes in the average count values as well as their dispersion. A generalised linear model framework, with a logit link function, is used to allow the study of covariates in the feature count distributions. The model fit is performed by maximum likelihood using the trust region optimisation algorithm [17]. Likelihood-ratio or Wald tests can be used to test the null hypothesis of no DA.

Songbird

songbird is a python package [53] that ranks microbes that are changing the most relative to each other [16]. The method is based on a compositional approach in which the underlying count distribution is assumed to be multinomial. The coefficients from multinomial regression can be ranked to determine which taxa are changing the most between samples. The compositionality is addressed using the differential abundance of each taxon as reference to each other when they are ranked numerically. Since songbird has been developed as an extension tool for Qiime2, we converted all our data tables to the .biom format to serve as input for this method. The authors' suggested analysis pipeline requires several manual adjustments to the tuning parameters on the basis of the comparison of the results after several runs, making it difficult to implement this method within a benchmarking framework. For this reason, we used the default values for all the tuning parameters.

mixMC

mixMC is a multivariate framework implemented in mixOmics, a Bioconductor package (v6.6.1), for omic data analysis [18]. It handles compositional and sparse data, repeated-measures experiments, and multiclass problems. After the addition of a pseudo-count value of 1, the TSS normalisation is applied to the count table and the CLR transformation is performed to account for compositionality. The method is based on a Partial Least Squares Discriminant Analysis (PLS-DA), a multivariate regression model which maximises the covariance between linear combinations of the feature counts and the outcome (in our case, a dummy variable indicating the body site/group of each sample). Covariance maximisation is achieved in a sequential manner via the use of latent component scores [18]. Each component is a linear combination of the feature counts and characterises a source of covariation between the feature and the groups. The sparse version of PLS-DA, sPLS-DA uses Lasso penalisations to select the most discriminative features in the PLS-DA model. The penalisation is applied component-wise and the resulting selected features reflect the particular source of covariance in the data highlighted by each PLS component. We specified the number of features to select per component at 100 or more, and we optimised it using leave-one-out cross-validation. Since we always compared two groups in this manuscript, only the first component is necessary for the analysis. The multivariate regression coefficients, one for each feature, were ranked in order to obtain the most discriminant features for the first component.

MAST

MAST is a Bioconductor package for managing and analysing quantitative PCR and sequencing-based single-cell gene expression data, as well as data from other types of single-cell assays. The package also provides functionality for significance testing of differential expression using a Hurdle model. Zero rate represents the discrete part, modelled as a binomial distribution while $\log_2\left(\frac{counts_{i,j} \cdot median(libSize)}{libSize_j} + 1\right)$ where i and j represent the i^{th} feature and the j^{th} sample, respectively, is used for the continuous part, modelled as a Gaussian distribution. The kind of data considered, different from scRNA-Seq, does not allow the usage of the adaptive thresholding procedure suggested in the original publication [7]. Indeed, because of the amount of feature loss, if adaptive thresholding is applied, the comparison of MAST with other methods would be unfair. However, a normalisation variable is included in the model. This variable captures information about each feature sparsity related to all the others; hence, it helps to yield more interpretable results and decreases background correlation between features. The function *zlm* fits the Hurdle model for each feature: the regression coefficients of the discrete component are regularised using a Bayesian approach as implemented in the *bayesglm* function; regularisation of the continuous model variance parameter helps to increase the robustness of feature-level differential expression analysis when a feature is only present in a few samples. Because the discrete and continuous parts are defined conditionally independent for each feature, tests with asymptotic χ^2 null distributions, such as the likelihood-ratio or Wald tests, can be summed and remain asymptotically χ^2, with the degrees of freedom of the component tests added. BH correction method was used to correct p-values.

Seurat with Wilcoxon rank sum test

Seurat (v2.3.4) R package is a data analysis toolkit for the analysis of scRNA-Seq [22]. Briefly, counts were scaled, centred, and log-normalised. Wilcoxon rank sum test for detecting differentially abundant features was performed via the *FindMarkers* function. Rare features, which are present in a fraction lower than 0.1 of all samples, and weak signal features, which have a log fold change between conditions lower than 0.25, are not tested. BH correction method was used to correct p-values.

SCDE

The scde Bioconductor package (v1.99.1) with flexmix package (v2.3-13) implements a Bayesian model for scRNA-Seq data [8]. Read counts observed for each gene are modelled using a mixture of a NB distribution (for the amplified/detected transcripts) and low-level Poisson distribution (for the unobserved or background-level signal of genes that failed to amplify or were not detected for other reasons). The *scde.error.models* function was used to fit the error models on which all subsequent calculations rely. The fitting process is based on a subset of robust genes detected in multiple cross-cell comparisons. Error models for each group of cells were fitted independently (using two different sets of "robust" genes). Translating in a metagenomic context, cells correspond to samples and genes to taxa or amplicon sequence variants. Some adjustments were needed to calibrate some function default values such as the minimum number of features to use when determining the expected abundance magnitude during model fitting. This option, defined by the *min.size.entries* argument, set by default at 2000, was too big for many 16S or WMS experiment scenarios: as we usually observe around 1000 total features per dataset (after filtering out rare ones), we decided to replace 2000 with the 20% of the total number of features, obtaining a dataset-specific value. Particularly, poor samples may result in abnormal fits and were removed as suggested in the scde manual. To test for differential expression between the two groups of samples a Bayesian approach was used: incorporating evidence provided by the measurements of individual samples, the posterior probability of a feature being present at any given average level in each subpopulation was estimated. To moderate the impact of high-magnitude outlier events, bootstrap resampling was used and posterior probability of abundance fold-change between groups was computed.

3.6.5 Type I error control

For this analysis, we used the collection of HMP Stool samples in HMP16SData and curatedMetagenomicData. The Multi-Dimensional Scaling (MDS) plot of the beta diversity did not show patterns associated with known variables (Additional file 1: Supplementary Fig. S3); hence, we assumed no differential abundance. All samples with the same RSID in 16S and WMS were selected in order to easily compare the two technologies. Forty-one biological samples were included.

Starting from the 41 samples, we randomly split the samples into two groups: 21 assigned to group 1 and 20 to group 2. We repeated the procedure 1000 times. We applied the DA methods to each randomly split dataset. Every method returned a p-value for each feature. DESeq2, seurat_wilcoxon, and corncob methods returned some *NA* p-values. This is due to feature exclusion criteria, based on distributional assumptions, performed by these methods (see above), or convergence issues.

We compared the distribution of the observed p-values to the theoretical uniform distribution, as no truly DA features are present. This was summarised in the qq-plot where the bisector represents a perfect correspondence between observed and theoretical quantiles of p-values. For each theoretical quantile, the corresponding observed quantile was obtained averaging the observed p-values' quantiles from all 1000 datasets. Departure from uniformity was evaluated with a Kolmogorov-Smirnov statistic. p-values were also used to compare the number of false discoveries with 3 common thresholds: 0.01, 0.05, and 0.1.

3.6.6 Concordance

We used the Concordance At the Top (CAT) to evaluate concordance for each differential abundance method. Starting from two lists of ranked features (by p-values, fold-changes, or other measures), the CAT statistic was computed in the following way. For a given integer i, concordance is defined as the cardinality of the intersection of the top i elements of each list, divided by i, *i.e.*, $\frac{\#(L_{1:i} \cap M_{1:i})}{i}$, where L and M represent the two lists. This concordance was computed for values of i from 1 to R.

Depending on the study, only a minority of features may be expected to be differentially abundant between two experimental conditions. Hence, the expected number of differentially abundant features is a good choice as the maximum rank R. In fact, CAT displays high variability for low ranks as few features are involved, while concordance tends to 1 as R approaches the total number of features, becoming uninformative. We set $R = 100$, considering this number biologically relevant and high enough to permit an accurate concordance evaluation. In our filtered data, the total number of features was close to 1000, and 100 corresponds to 10% of total taxa. We used CAT for two different analyses:

- Between Methods Concordance (BMC), in which a method was compared to other methods in the same dataset;
- Within Method Concordance (WMC), in which a method is compared to itself in random splits of the datasets.

To summarise this information for all pairwise method comparisons, we computed the area under the curve, hence giving a better score to two methods that are consistently concordant for all values of i from 1 to 100.

We selected several datasets, with different α- and β-diversity, for our concordance analysis. Additional file 1: Table S3 describes the six datasets used. For each dataset, the same sample selection step, described next, was used.

The concordance evaluation algorithm can be easily summarised by the following steps:

1. Each dataset was randomly divided in half to obtain two subsets (Subset1 and Subset2) with two balanced groups;
2. DA analysis between the groups was performed with all evaluated methods independently on each subset;

3. For each method, the list of features ordered by p-values (or differentials, or loadings) obtained from Subset1 was compared to the analogous list obtained from Subset2 and used to evaluate WMC;
4. For each method, the list of features ordered by p-values (or differentials, or loadings) obtained from Subset1 was compared to the analogous list obtained from Subset1 by all the other methods and used to evaluate BMC for Subset1. The same was done in Subset2;
5. Steps 1–4 were repeated 100 times; and
6. WMC and BMC were averaged across the 100 values (and between Subset1 and Subset2 for BMC) to obtain the final values.

Sample selection step

For each dataset, a subset was chosen in order to have a balanced number of samples for each condition. In lower diversity studies (*e.g.*, Subgingival vs. Supragingival Plaque) different biological samples from the same subject may be strongly correlated. Hence, we selected only one sample per individual, no matter the condition. To further increase the homogeneity of the datasets, we selected only samples from the same sequencing center.

3.6.7 Enrichment analysis

The same low-diversity dataset used in the concordance analysis (*i.e.*, 16S Subgingival vs. Supragingival Plaque) was used for the enrichment analysis. The dataset is balanced as it is composed of 38 samples for each body subsite, for a total of 76 samples. DA analysis was performed using Subgingival Plaque as the reference level. Taxa with an adjusted p-value less than 0.1 were chosen as DA, for all the methods except songbird and mixMC that return a list of differentials and loadings, respectively. For songbird, a threshold corresponding to the 10% of the total number of taxa was chosen to select the most associated taxa for the considered comparison. mixMC implements a variable selection procedure that automatically selects the most discriminant taxa. We annotated each taxon with the information on genus-level metabolism (available at `https://github.com/waldronlab/nychanesmicrobiome`), classifying each taxon in aerobic, anaerobic, facultative anaerobic, or unassigned. Enrichment analysis was performed via a Fisher exact test, using the function *fisher.test* (*table*, *alternative="greater"*) where *table* is a contingency table. Six contingency tables were built for each method to inspect enrichment of the following:

- Over-abundant (UP) aerobic taxa in Supragingival Plaque;
- Under-abundant (DOWN) aerobic taxa in Subgingival Plaque;
- Over-abundant (UP) anaerobic taxa in Supragingival Plaque;
- Under-abundant (DOWN) anaerobic taxa in Subgingival Plaque;
- Over-abundant (UP) facultative anaerobic taxa in Supragingival Plaque; and
- Under-abundant (DOWN) facultative anaerobic taxa in Subgingival Plaque.

All the information retrieved from the enrichment analysis was summarised in a bar plot, where for each method, the number of differentially abundant taxa together with their direction were represented as a positive (negative) bar for over- (under-) abundant taxa in Supragingival Plaque samples, colored by genus level metabolism.

To calculate log odds-ratio for each contingency table, the Haldane-Anscombe correction is applied since it allows the odds-ratio calculation in presence of zero cells. Briefly, it consists in adding a pseudo-count value of 0.5 to each cell of the contingency table to calculate the odds-ratio and a pseudo-count value of 1 to calculate the variance.

To compare all the evaluated methods without considering their power, the followings steps were followed:

1. Raw p-values, songbird's differentials, and mixMC's loadings were properly ordered;
2. Several thresholds from 1 to 20% of the top ranked taxa in the previously ordered lists were used to select the DA taxa for each method;

3. Putative true positives (TP) were calculated as the sum of aerobic taxa over-abundant in Supragingival Plaque and anaerobic taxa under-abundant in Supragingival Plaque;
4. Putative false positives (FP) were calculated as the sum of aerobic taxa under-abundant in Supragingival Plaque and anaerobic taxa over-abundant in Supragingival Plaque; and
5. The differences between Putative TP and Putative FP were plotted.

To rank all the methods, the same difference was computed, this time using the list of DA taxa based on the adjusted p-values less than 0.1 and the 10% threshold for songbird.

To inspect the concordance of DA taxa between methods, mutual findings were collected and added between the methods. As similar methods tend to identify the same taxa, only one method for each normalisation or weighting procedure was considered as representative. This subset contains edgeR with TMM normalisation, DESeq2 with poscounts normalisation, limma-voom with TMM normalisation, MAST, scde, seurat-wilcoxon, corncob (Wald test), mgsZig, ALDEx2, mixMC, and songbird. The taxa found by most methods in this subset were extracted, but for the graphical representation, all methods were reintroduced.

The same analysis was performed in the WMS dataset. However, the sample size was limited to only 5 for the subgingival body subsite, while 88 (with unique RSID) for the supragingival site. For this reason, a 5 vs. 5 sample analysis was performed, randomly selecting five samples from the supragingival dataset. Songbird was not included in the analysis because of an error during the parameter estimation that we were not able to solve. Given the low sample-size, corncob methods with bootstrap were added to the analysis.

3.6.8 Parametric simulations

Several real datasets were used as templates for the simulations:

- 41 Stool samples available for both 16S and WMS from HMP.

- 208 16S samples and 90 WMS samples of Tongue Dorsum body subsite from HMP.
- 67 Stool and 56 Oral cavity WMS data of Fijian adult women from BritoIl_2016.

Each dataset was filtered to obtain only a sample per individual. 16S and WMS samples were pruned to keep sequencing runs with library sizes of more than 103 and 106, respectively. Moreover, only features present in more than 1 sample with more than 10 reads were kept. After the data filtering step, the simulation framework was established, by specifying the parametric distribution and other data characteristics, described in Additional file 2: Supplementary Table S4.

For each combination of parameters, we simulated 50 datasets, yielding a total of 28,800 simulations. Variables to be included in the simulation framework were chosen based on the role they may play in the analysis of a real experiment.

NB and ZINB are simple parametric distributions, easy to fit on real data through a reliable Bioconductor package, and above all, seemed to fit 16S and WMS data better than other statistical models (see Fig. 3.2). The *zinbSim* function from the zinbwave Bioconductor package easily allows the user to generate both NB and ZINB counts after the *zinbFit* function estimates model parameters from real data. The user can set several options in *zinbFit*, we used *epsilon=1e14*, *common_ dispersion=TRUE*, and *K=0*.

Generating two experimental groups requires the specification of enough samples for each condition and a more or less substantial biological difference between them.

Sample size is a crucial parameter: many pilot studies start with 10 or even fewer samples per condition, while clinical trials and case-control studies may need more samples in order to achieve the needed power. We included 10, 20, and 40 samples per condition in our simulation framework.

We considered two different scenarios for the number of features simulated as DA: 10%, representing a case where the majority of the features are not

DA, a common assumption made by analysis methods; and 50%, a more extreme comparison. Similarly, we simulated a fold change difference for the DA features of 2 or 5. This is obviously a simplification, since in reality, a continuum gradient of fold effects is present. Nevertheless, it allowed us to characterise the role of the effect size in the performance of the methods. For the DA features, the fold change between conditions was applied to the mean parameter of the ZINB or NB distributions, with or without "compensation" as introduced by [10]. Without compensation, the absolute abundance of a small group of features responds to a physiological change. This simple procedure modifies the mean relative abundances of all features, a microbiologist would only want to detect the small group that initially reacted to the physiological change. For this reason, significant results for other features will be considered as false discoveries. Compensation prevents the changes in DA features to influence the other, non-DA, features. The procedure comprises the following steps:

1. The relative mean for each feature is computed using estimated mean parameter of NB.
2. 10 or 50% of features are randomly sampled.
3. If there is no compensation, half of their relative means are multiplied by *foldEffect* while the remainings are divided by *foldEffect* generating up- and down-regulated features, respectively. If there is compensation, $\frac{1}{1+foldEffect}$ of the selected feature relative means are multiplied by *foldEffect* while the remaining ones are multiplied by $\frac{a}{b}(1 - foldEffect) + 1$, where a is the sum of the relative means of the features that will be up-regulated while b is the sum of the features that will be down-regulated.
4. The resulting relative means are normalised to sum to 1.

Sparsity is a key characteristic of metagenomic data. The case in which a bacterial species presence rate varies between conditions was emulated in the simulation framework via the so called *sparsityEffect* variable. Acting on the mixture parameter of the ZINB model it is possible to exacerbate down-regulation and up-regulation of a feature, adding zeros for the former and reducing zeroes for the latter. This scenario provided by 0 (no sparsity change at all), 0.05 and 0.15 of sparsity change should help methods to identify more differentially abundant features. As the mixing parameter can only take values between 0 and 1, when the additive sparsity effect yielded a value outside this range, it was forced to the closer limit.

The previously described DA methods were tested in each of the simulated datasets (50 for each set of simulation framework parameters) and the adjusted p-values were used to compute the False Positive Rate (FPR=1-Specificity) and the True Positive Rate (TPR=Sensitivity). Partial areas under the receiver operating characteristic (pAUROC) curve with an FPR from 0 to 0.1 values were computed and then averaged in order to obtain a single value for each set of variables.

3.6.9 Computational complexity

To measure the computational times for all the 18 methods, we used the Subgingival vs. Supragingival Plaque HMP 16S dataset where a total of 76 samples and approximately 900 taxa were available. The evaluation was performed on a laptop computer with O.S. Windows 10 64bit, Intel® i7-8th Gen CPU with 16GB of RAM. Moreover, the Stool 16S and WMS parametric simulation datasets (9200 total datasets) were used in order to measure each method's computational complexity (except for mixMC and songbird). Time evaluation was performed on a single core for each dataset where all methods are tested sequentially and then properly averaged with the values of all the simulations. The methods' performance evaluations in power analysis on the 28,800 total parametric simulations were performed in the same way, equally dividing the simulated datasets across 30 cores. The working machine was a Linux x86_64 architecture server with 2 Intel® Xeon® Gold 6140 CPU with 2.30 GHz for a total of 72 CPUs and 128 GB of RAM.

3.6.10 Availability of data and materials

The real datasets used in this article are available in the HMP16SData Bioconductor package [25], available at `https://bioconductor.org/packages/HMP16SData`, and in the curatedMetagenomicData Bioconductor package [26], available at `https://bioconductor.org/packages/curatedMetagenomicData`. The scripts to reproduce all analyses and figures of this article are available at `https://github.com/mcalgaro93/sc2meta` [54] under a MIT license (archived source code at time of publication: `https://zenodo.org/record/3942108#.XwyN1ygzZPY`).

References

1. 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 (2007).
2. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* **12,** 902–903 (2015).
3. Zhu, S., Qing, T., Zheng, Y., Jin, L. & Shi, L. Advances in single-cell RNA sequencing and its applications in cancer research. *Oncotarget* **8,** 53763–53779 (2017).
4. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology* **34,** 1145–1160 (2016).
5. Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nature Reviews Immunology* **18,** 35–45 (2018).
6. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* **16,** 133–145 (2015).
7. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16,** 278 (2015).
8. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nature Methods* **11,** 740–742 (2014).
9. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology* **8** (2017).
10. Hawinkel, S., Mattiello, F., Bijnens, L. & Thas, O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Briefings in Bioinformatics* **20,** 210–221 (2019).

11. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5,** 27 (2017).
12. Thorsen, J. *et al.* Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome* **4,** 62 (2016).
13. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods* **10,** 1200–1202 (2013).
14. Fernandes, A. D. *et al.* 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 (2014).
15. Russel, J. *et al. DAtest: a framework for choosing differential abundance or expression method* 2018.
16. Morton, J. T. *et al.* Establishing microbial composition measurement standards with reference frames. *Nature Communications* **10,** 2719 (2019).
17. Martin, B. D., Witten, D. & Willis, A. D. Modeling microbial abundances and dysbiosis with beta-binomial regression. *The Annals of Applied Statistics* **14,** 94–115 (2020).
18. Lê Cao, K.-A. *et al.* MixMC: A Multivariate Statistical Framework to Gain Insight into Microbial Communities. *PLoS ONE* **11** (2016).
19. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
20. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 550 (2014).

21. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15,** R29 (2014).
22. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36,** 411–420 (2018).
23. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications* **9,** 284 (2018).
24. Van den Berge, K. *et al.* Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biology* **19,** 24 (2018).
25. Schiffer, L. *et al.* HMP16SData: Efficient Access to the Human Microbiome Project Through Bioconductor. *American Journal of Epidemiology* **188,** 1023–1026 (2019).
26. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nature Methods* **14,** 1023–1024 (2017).
27. Murdoch, D. J., Tsai, Y.-L. & Adcock, J. P -Values are Random Variables. *The American Statistician* **62,** 242–245 (2008).
28. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology* **10,** 766 (2014).
29. Castro-Nallar, E. *et al.* Composition, taxonomy and functional diversity of the oropharynx microbiome in individuals with schizophrenia and controls. *PeerJ* **3,** e1140 (2015).
30. Consortium, T. H. M. P. Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).
31. Irizarry, R. A. *et al.* Multiple-laboratory comparison of microarray platforms. *Nature Methods* **2,** 345–350 (2005).
32. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11,** 94 (2010).

33. Beghini, F. *et al.* Tobacco exposure associated with oral microbiota oxygen utilization in the New York City Health and Nutrition Examination Study. *Annals of Epidemiology* **34,** 18–25.e3 (2019).
34. Thurnheer, T., Bostanci, N. & Belibasakis, G. N. Microbial dynamics during conversion from supragingival to subgingival biofilms in an in vitro model. *Molecular Oral Microbiology* **31,** 125–135 (2016).
35. Xiao, C., Ran, S., Huang, Z. & Liang, J. Bacterial Diversity and Community Structure of Supragingival Plaques in Adults with Dental Health or Caries Revealed by 16S Pyrosequencing. *Frontiers in Microbiology* **7** (2016).
36. Socransky, S. S., Haffajee, A. D., Cugini, M. A., Smith, C. & Kent, R. L. Microbial complexes in subgingival plaque. *Journal of Clinical Periodontology* **25,** 134–144 (1998).
37. Paster, B. J. *et al.* Bacterial Diversity in Human Subgingival Plaque. *Journal of Bacteriology* **183,** 3770–3783 (2001).
38. Townes, F. W., William Townes, F., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. *Feature Selection and Dimension Reduction for Single Cell RNA-Seq based on a Multinomial Model* 2019.
39. Svensson, V. *Droplet scRNA-seq is not zero-inflated* 2019.
40. Quinn, T. P., Erb, I., Richardson, M. F. & Crowley, T. M. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics* **34,** 2870–2878 (2018).
41. Quinn, T. P., Crowley, T. M. & Richardson, M. F. Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. *BMC Bioinformatics* **19,** 274 (2018).
42. Calle, M. L. Statistical Analysis of Metagenomics Data. *Genomics & Informatics* **17,** e6 (2019).
43. Aitchison, J. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society: Series B (Methodological)* **44,** 139–160 (1982).

44. Kumar, M. S. *et al.* Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genomics* **19,** 799 (2018).
45. Mandal, S. *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease* **26** (2015).
46. Mao, J., Chen, Y. & Ma, L. Bayesian Graphical Compositional Regression for Microbiome Data. *Journal of the American Statistical Association* **115,** 610–624 (2019).
47. Bogomolov, M., Peterson, C. B., Benjamini, Y. & Sabatti, C. *Testing hypotheses on a tree: new error rates and controlling strategies* 2017.
48. Lu, J., Shi, P. & Li, H. Generalized linear models with linear constraints for microbiome compositional data. *Biometrics* **75,** 235–244 (2019).
49. Rivera-Pinto, J. *et al.* Balances: a New Perspective for Microbiome Analysis. *mSystems* **3** (2018).
50. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11,** R25 (2010).
51. McMurdie, P. J. & Holmes, S. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **8** (2013).
52. Martin, B. D. *bryandmartin/corncob* `https://github.com/bryandmartin/corncob` (2020).
53. Morton, J. T. *biocore/songbird* `https://github.com/biocore/songbird` (2020).
54. Calgaro, M. *mcalgaro93/sc2meta* `https://github.com/mcalgaro93/sc2meta` (2020).

Chapter 4

benchdamic: benchmarking of differential abundance methods for microbiome data

The work described in this chapter is taken from: M. Calgaro, C. Romualdi, D. Risso, N. Vitulo; benchdamic: benchmarking of differential abundance methods for microbiome data. Bioinformatics 39, 1 (2023). `https://doi.org/10.1093/bioinformatics/btac778`

Contributions

MC and NV conceived the project. CR and DR co-developed the evaluation strategies. **MC** drafted the manuscript. DR, CR, and NV reviewed and edited the manuscript. **MC** developed the package and curated its manual. All Authors read and approved the final manuscript.

Challenges and future perspectives

This work represents the most tangible result of my doctoral programme. While the previous chapter assessed many of the available tools for Differential Abundance Analysis (DAA), in this chapter I explain the structure of an R software package that can be used to perform such evaluations on a generic dataset. The package itself is easily described in a few pages, but the real content is represented by its comprehensive manual and its utility. Thousands of lines of code, dozens of functions, and their parameters are intended to provide flexibility and robustness. The main challenges in creating this package, named benchdamic, were: organising the code, optimising it, and maintaining the manual. These challenges were also the main requirements to join the Bioconductor project. Bioconductor is an open-source software platform for the analysis and visualisation of biological data. Since joining the platform in October 2021, the package has reached a wide audience with more than 400 downloads.

A lot has changed in the last year when this application note article was published: parallel processing has been enabled for the most time-consuming tasks, code quality has been further improved and more DAA tools have been introduced. From a practical point of view, the introduction of a new DAA tool is associated with many problems. On the one hand, it requires input and output management to ensure comparability with the other tools. On the other hand, it leads to increased complexity of the package, both computationally and methodologically speaking. While the latter problem could be solved by increasing computation times or, where possible, by code optimisation, input and output management requires some special precau-

tions. For this reason, I am continuously updating the package and in the meantime have set up a section of the manual where users can learn how to add custom tools themselves.
The future of benchdamic is focused on improving and adding new DAA tools. Indeed, new normalisations, DAA tools, or similar approaches will be released.

Article location

`https://academic.oup.com/bioinformatics/article/39/1/btac778/6881076`

Supplementary material

The supplementary material for this article consists of a static version of the package manual at the time of the publication. It is available at the supplementary-data section of the journal page `https://academic.oup.com/bioinformatics/article/39/1/btac778/6881076#supplementary-data`. Alternatively, the up-to-date package vignette is available at the Bioconductor repository `https://www.bioconductor.org/packages/release/bioc/html/benchdamic.html`

4.1 Abstract

Recently, an increasing amount of methodological approaches have been proposed to tackle the complexity of metagenomics and microbiome data. In this scenario, reproducibility and replicability have become two critical issues, and the development of computational frameworks for the comparative evaluations of such methods is of utmost importance. Here, we present benchdamic, a Bioconductor package to benchmark methods for the identification of differentially abundant taxa. benchdamic is available as an open-source R package available through the Bioconcutor project at `https://bioconductor.org/packages/benchdamic/`.

4.2 Introduction

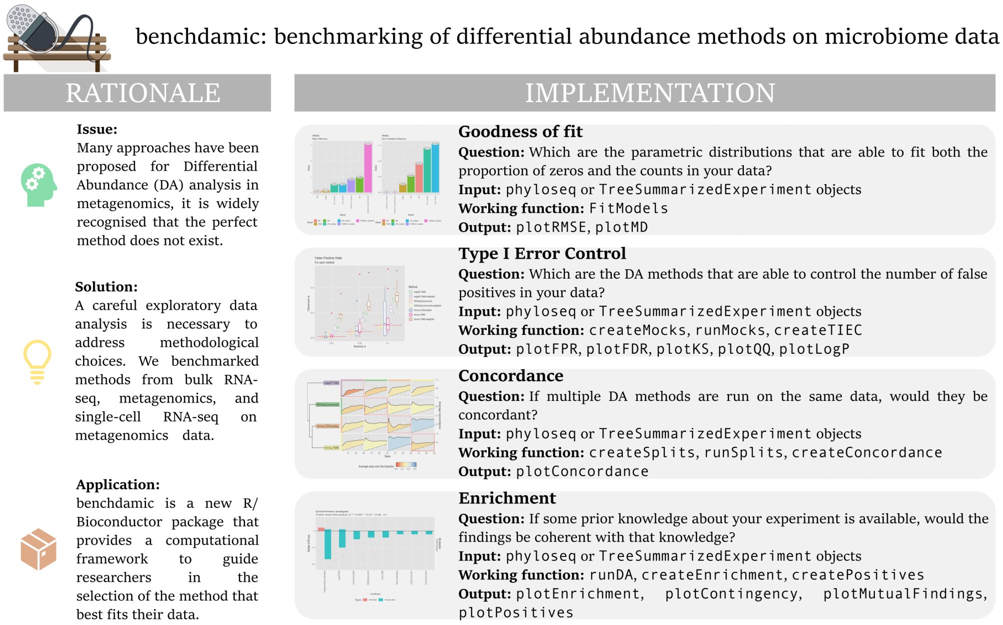
Differential abundance (DA) analysis identifies significant differences in the microbial community composition between groups of samples, providing new insights on the composition of microbial communities and on their associations with the environment. Although many approaches have been proposed for DA analysis, it is widely recognised that the best method (*i.e.*, a method with performances uniformly better than all the others) does not exist and that a careful exploratory data analysis is necessary to address methodological choices [1–5].

Building on our previous work [1], we present the benchdamic R/Bioconductor package, which provides a computational framework to guide researchers in the selection of the method that best fits their data.

The structure of benchdamic can be summarized into 4 main parts (Fig. 4.1). Each section is developed to answer specific questions when comparing samples from different experimental groups, namely: i) the ability for a given statistical distribution to successfully fit the input data, with particular focus on sparsity and their count nature; ii) the ability of the DA methods to control the type I error; iii) the concordance among methods; and iv) the accuracy of the findings based on a priori biological knowledge. Altogether, benchdamic is a flexible and customisable framework that can be used for the benchmarking of new and existing DA methods.

4.3 Implementation

benchdamic builds on existing R/Bioconductor infrastructure packages: the primary input of benchdamic's main functions is a phyloseq or a TreeSummarizedExperiment object [6, 7]. Ready-to-use normalisation and DA methods included in benchdamic are based on the edgeR [8], DESeq2 [9], limma-voom [10–12], metagenomeSeq [13], ALDEx2 [14, 15], corncob [16], MAST [17], Seurat [18], dearseq [19], NOISeq [20], ANCOMBC [21, 22], and zinbwave [23, 24] packages. Combinations of parameters are possible as well as



benchdamic: benchmarking of differential abundance methods on microbiome data

RATIONALE

Issue:
Many approaches have been proposed for Differential Abundance (DA) analysis in metagenomics, it is widely recognised that the perfect method does not exist.

Solution:
A careful exploratory data analysis is necessary to address methodological choices. We benchmarked methods from bulk RNA-seq, metagenomics, and single-cell RNA-seq on metagenomics data.

Application:
benchdamic is a new R/Bioconductor package that provides a computational framework to guide researchers in the selection of the method that best fits their data.

IMPLEMENTATION

Goodness of fit
Question: Which are the parametric distributions that are able to fit both the proportion of zeros and the counts in your data?
Input: `phyloseq` or `TreeSummarizedExperiment` objects
Working function: `FitModels`
Output: `plotRMSE`, `plotMD`

Type I Error Control
Question: Which are the DA methods that are able to control the number of false positives in your data?
Input: `phyloseq` or `TreeSummarizedExperiment` objects
Working function: `createMocks`, `runMocks`, `createTIEC`
Output: `plotFPR`, `plotFDR`, `plotKS`, `plotQQ`, `plotLogP`

Concordance
Question: If multiple DA methods are run on the same data, would they be concordant?
Input: `phyloseq` or `TreeSummarizedExperiment` objects
Working function: `createSplits`, `runSplits`, `createConcordance`
Output: `plotConcordance`

Enrichment
Question: If some prior knowledge about your experiment is available, would the findings be coherent with that knowledge?
Input: `phyloseq` or `TreeSummarizedExperiment` objects
Working function: `runDA`, `createEnrichment`, `createPositives`
Output: `plotEnrichment`, `plotContingency`, `plotMutualFindings`, `plotPositives`

Figure 4.1: Graphical abstract. Each box on the right represents a step of the analysis where information about the research question, type of input data, working functions, and outputs are reported.

the inclusion of custom methods (Additional file: Section 3).

In the following sections, we briefly outline the main functionality of the package. See [1] for technical details on how these metrics are computed.

4.3.1 Goodness of fit

DA statistical models are based on different statistical distributions. Five different distributions are available in benchdamic for testing the goodness of fit on user-provided data: Negative Binomial, Zero-Inflated Negative Binomial, Zero-Inflated Gaussian, Truncated Gaussian, Dirichlet-Multinomial (Additional file: Section 2). Goodness of fit is measured by the ability of each method to correctly estimate the average counts and the probability of observing a zero.

4.3.2 Type I error control

To investigate the Type I error rate control of each DA method (*i.e.*, the probability of the statistical test to call a feature DA when it is not) mock datasets with no true DA are generated starting from the user-provided data (Additional file: Section 4).

Briefly, the dataset is split into two random subsets and DA analysis, based on a chosen list of methods, is performed. The process is repeated N times (N $\geq$ 1000 suggested). The performances of each method are then summarized and graphically represented considering the false positive rate, false discovery rate, and departure from uniformity for the p-values distribution.

4.3.3 Concordance

benchdamic can be used to measure the Between Methods Concordance (BMC), in which a DA method is compared to other methods in the same dataset, and the Within Method Concordance (WMC), in which a method is compared to itself in two random subsets of the same dataset (Additional file: Section 5). Firstly, the dataset is randomly divided in half to obtain two subsets (Subset1 and Subset2) with samples from two or more biological groups, then DA analysis is performed between two groups, independently on each subset. The process is repeated N times (N $\geq$ 100 suggested) and average WMC and BMC metrics are computed and summarized using a heatmap representation.

4.3.4 Enrichment analysis

Enrichment analysis can provide an alternative way of ranking methods in terms of their ability to identify, as significantly different, taxa that are known to be differentially abundant between two groups. DA analysis needs to be performed on a dataset where some a priori knowledge is available (Additional file: Section 6). Given the direction of the DA features (over- or under-abundant) and the expected group in which the features should be differentially abundant according to the prior knowledge, several contingency tables are created for each DA method. A Fisher exact test is then performed to test the enrichment and the DA features identified by more than one method are highlighted. Additionally, the users will be able to rank the methods based on the difference between the total number of True Positives and False Positives for several thresholds (based on p-values, adjusted p-values, or other statistics). The same approach can also be used to perform power analysis using simulated data (Additional file: Section 6.8)

4.4 Conclusions

The benchdamic R/Bioconductor package aims to be a support tool for the identification of DA microbial taxa and the benchmarking of new methods. We envision two main uses of our package: (i) for practitioners interested in performing DA analysis on a new dataset, benchdamic can be used to identify the best DA methods among those already in the literature; (ii) for method developers interested in benchmarking their new approach, benchdamic can be used as an impartial tool to evaluate the relative merits of the new method compared to what is already available. benchdamic is available as an open-source package through the Bioconductor project. The package includes a vignette with a detailed tutorial.

The future of benchdamic is oriented to the addition of new aspects of analysis, *e.g.*, new normalisation methods and new DA approaches.

References

1. Calgaro, M., Romualdi, C., Waldron, L., Risso, D. & Vitulo, N. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biology* **21,** 191 (2020).
2. Thorsen, J. *et al.* Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome* **4,** 62 (2016).
3. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5,** 27 (2017).
4. Hawinkel, S., Mattiello, F., Bijnens, L. & Thas, O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Briefings in Bioinformatics* **20,** 210–221 (2019).
5. Nearing, J. T. *et al.* Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications* **13,** 342 (2022).
6. McMurdie, P. J. & Holmes, S. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **8** (2013).
7. Huang, R. *et al.* TreeSummarizedExperiment: a S4 class for data with hierarchical structure. *F1000Research* **9,** 1246 (2021).
8. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
9. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 550 (2014).
10. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **43,** e47–e47 (2015).

11. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* **15,** R29 (2014).
12. Phipson, B., Lee, S., Majewski, I. J., Alexander, W. S. & Smyth, G. K. Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression. *The Annals of Applied Statistics* **10** (2016).
13. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods* **10,** 1200–1202 (2013).
14. Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G. & Gloor, G. B. ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. *PLoS ONE* **8,** e67019 (2013).
15. Fernandes, A. D. *et al.* 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 (2014).
16. Martin, B. D., Witten, D. & Willis, A. D. Modeling microbial abundances and dysbiosis with beta-binomial regression. *The Annals of Applied Statistics* **14,** 94–115 (2020).
17. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology* **16,** 278 (2015).
18. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36,** 411–420 (2018).
19. Gauthier, M., Agniel, D., Thiébaut, R. & Hejblum, B. P. dearseq: a variance component score test for RNA-seq differential analysis that effectively controls the false discovery rate. *NAR Genomics and Bioinformatics* **2,** lqaa093 (2020).

20. Tarazona, S. *et al.* Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research,* gkv711 (2015).

21. Lin, H. & Peddada, S. D. Analysis of compositions of microbiomes with bias correction. *Nature Communications* **11,** 3514 (2020).

22. Mandal, S. *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease* **26** (2015).

23. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications* **9,** 284 (2018).

24. Van den Berge, K. *et al.* Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biology* **19,** 24 (2018).

Part II

Case Studies

Chapter 5

Metabarcoding analysis of gut microbiota of healthy individuals reveals impact of probiotic and maltodextrin consumption

The work described in this chapter is taken from: M. Calgaro, M. Pandolfo, E. Salvetti, A. Marotta, I. Larini, M. Pane, A. Amoruso, A. Del Casale, N. Vitulo, M. Fiorio and G.E. Felis; Metabarcoding analysis of gut microbiota of healthy individuals reveals impact of probiotic and maltodextrin consumption. Beneficial Microbes 12:2, 121-136 (2021). https://doi.org/10.3920/BM2020.0137

Contributions

MP1 performed the bioinformatics analysis, **MC** performed the statistical analyses. AM and MF helped in the interpretation of psychological variables. MP2 and AA supplied the probiotics and the placebo. ADC provided support for the trial. **MC**, MP, ES, and IL drafted the manuscript. NV and GEF reviewed and edited the manuscript. All Authors read and approved the final manuscript.

Challenges and future perspectives

The work of Marotta and colleagues [1] showed that probiotics intake exerted a positive effect on sleep quality and a general improvement across time in different aspects of the profile of mood state, like sadness, anger, and fatigue in 33 healthy individuals. This work extends the findings of that publication by conducting a metabarcoding analysing of the stool samples collected from the same cohort during the experiment.

From an analytical perspective, the complexity of this experiment lies in its longitudinal nature. Initially, the analyses were meaningless due to the heterogeneity of the samples. The baseline microbiota of each subject could respond differently to probiotics or placebo assumption. Only through careful exploratory data analysis was I able to establish the presence of different groups of subjects in the β-diversities of the samples at baseline. This allowed the other authors and I to hypothesise the presence of 3 or 4 groups of subjects who had a similar microbiota. Finally, 3 groups of subjects were selected through a data-driven approach that favoured the best clustering performance.

In the context of sample heterogeneity, another challenge in this study was that the strong sample-specific effect over time masked the probiotics or placebo effects. To adjust the data for the sample-specific effects, I sought to examine the first principal coordinates of the ordinated β-diversities using mixed effects regression models. This approach, in combination with the 3 previously identified subject groups, allowed me to identify differential probiotics and placebo effects over time between groups.

The major limitation of this study was the small sample size, which became even more apparent when I started working on the subgroups of patients. This limited the scope of the research, which was more observational, and the results became more specific to this dataset. Nutritional data were also collected, but without proper refinement (*e.g.*, by adopting some macronutrient groups) they were useless.

As a future perspective, a higher sample size is needed to address covariates, including diet, physical activity habits, and seasonal aspects related to psychological states (*e.g.*, exams and other stressful periods of the year). In addition, the collection of metabolomics and metagenomics data, beyond metabarcoding, could facilitate a more comprehensive investigation of the causal relationships between probiotics intake, psychological variables, and thus the gut-brain axis.

Article location

`https://www.wageningenacademic.com/doi/10.3920/BM2020.0137`

Supplementary material

The supplementary material for this article consists of 1 additional file available at `https://www.wageningenacademic.com/doi/suppl/10.3920/BM2020.0137?role=tab`

5.1 Abstract

In a previously published double-blind, placebo-controlled study, we showed that probiotics intake exerted a positive effect on sleep quality and a general improvement across time in different aspects of the profile of mood state, like sadness, anger, and fatigue in 33 healthy individuals. The present work investigates the impact of the probiotic product, constituted of *Limosilactobacillus fermentum* LF16, *Lacticaseibacillus rhamnosus* LR06, *Lactiplan-*

tibacillus plantarum LP01 (all former members of *Lactobacillus* genus), and *Bifidobacterium longum* 04, on the gut microbiota composition of the same cohort through a metabarcoding analysis. Both the placebo and probiotic treatments had a significant impact on the microbiota composition. Statistical analysis showed that the microbiota of the individuals could be clustered into three groups, or bacteriotypes, at the baseline, and, inherently, bacterial compositions were linked to different responses to probiotic and placebo intakes. Interestingly, *L. rhamnosus* and *L. fermentum* were retrieved in the probiotic-treated cohort, while a bifidogenic effect of maltodextrin, used as placebo, was observed. The present study shed light on the importance of defining bacteriotypes to assess the impact of interventions on the gut microbiota and allowed to reveal microbial components which could be related to positive effects (*i.e.*, sleep quality improvement) to be verified in further studies.

5.2 Introduction

A great interest has emerged in the last years on the impact of the gut-brain axis in psychiatric disorders, pointing to stressed and unhealthy conditions of the microbial communities inside the human gut as possible causes for psychological diseases and conditions, such as depression, anxiety and schizophrenia [1, 2].

It has been observed that the gut microbiota communicates with the brain, exerting effects over several neurobiological mechanisms and related systems; among these the hypothalamic-pituitary-adrenal axis, the immune system, the tryptophan metabolism and the production of various neuroactive compounds [2].

For those reasons, the gut microbiota has become a new target to obtain antidepressant effects; remarkably, the diversity of studies performed and the functional redundancy of the microbiome make it difficult to understand if specific microbial components are more related than others to psychiatric symptoms [3].

Since microbiota composition can be modified in a variety of ways, such as through the use of probiotics, prebiotics and dietary changes [3, 4], several clinical and translational studies have been published over the years, showing that the prolonged prebiotic and probiotic consumption can positively affect aspects of mood, anxiety, and cognition in both healthy individuals as well as in patients diagnosed with clinical psychiatric disorders [2, 4, 5]. However, in some clinical trials, lack of evidence of an effect on depression and related symptoms have also been reported either in depressed [6] as well as in healthy individuals (in particular in older adults [4]) even though probiotic strains used were also previously successfully applied.

Probiotic supplements used in clinical trials for the treatment of depression, either alone [7–9], in combination with prebiotics (*i.e.*, galactooligosaccharides) [10], or as adjunctive therapy with antidepressants (*i.e.*, sertraline) [11] mainly include *Lactobacillus* (*L. acidophilus*, *L. helveticus*, *L. brevis*, *L. casei*, and *L. salivarius*), *Lactococcus* (*L. lactis*) and *Bifidobacterium* spp. (*B. bifidum*, *B. lactis*, and *B. longum*). Generally speaking, such treatments led to (1) a significant reduction of depression scores on Hospital Anxiety and Depression Scale and improvement of the cognitive reactivity scores in mild/moderate depression patients [8, 9], (2) a decrease of the anxiety symptoms in individuals with anxiety disorders [11] and (3) an improvement of the depression scores on Beck Depression Inventory in patients with a diagnosis of Major Depressive Disorder (MDD) [7, 10].

In healthy subjects, administration of probiotics (*L. casei* Shirota, *B. bifidum* W23, *B. lactis* W52, *L. acidophilus* W37, *L. brevis* W63, *L. casei* W56, *L. salivarius* W24, and *L. lactis* W19 and W58) improved the mood of subjects having lowest baseline mood levels and in general reduced the cognitive reactivity to sadness [12, 13].

Moreover, in other two studies led by Messaoudi *et al.* [14] and Mohammadi *et al.* [15], a significant reduction in overall anxiety and depression scores was shown after the treatment with *L. helveticus* R0052 and *B.*

longum R0175 as well as with a polybiotic combination of various *Lactobacillus* strains (*L. acidophilus*, *L. delbrueckii subsp. bulgaricus*, *L. casei*, and *L. rhamnosus*), *Bifidobacterium* (*B. breve*, *B. longum*) and *Streptococcus thermophilus* strains. Besides mood, anxiety and depression scores, it has also been shown that the short-term administration of *L. gasseri* improved stress-associated symptoms in terms of sleep disturbance [16].

Within the framework of the impact of probiotics on mood, we have previously reported [5] on a double-blind, placebo-controlled study on 33 healthy volunteers who received daily either a probiotic mixture containing *Limosilactobacillus fermentum* LF16, *Lacticaseibacillus rhamnosus* LR06, *Lactiplantibacillus plantarum* LP01 (former members of *Lactobacillus* genus, [17]), and *B. longum* 04 in maltodextrin, or a maltodextrin-only placebo, for 6 weeks, followed by a 3-weeks washout (Fig. 5.1). Data obtained showed that the probiotics exerted a general improvement and persistence over time in different aspects of the mood state, including sadness, anger, and fatigue, accompanied by improvement in the sleep quality, which indicates that probiotics may increase the production of neuroactive precursors involved in emotional modulation, brain functions and circadian rhythms. These findings corroborated the positive effect of probiotics on mental well-being, possibly determining changes in cognitive strategies to deal with problems by reducing sensitivity to negative situations.

The aim of the present study was to apply metabarcoding analysis of the faecal microbiota to determine (1) the microbial arrangement at baseline in the same cohort of healthy adults, which were randomised based on other characteristics, and (2) determine the effects of the probiotic and placebo consumption during and after the administration.

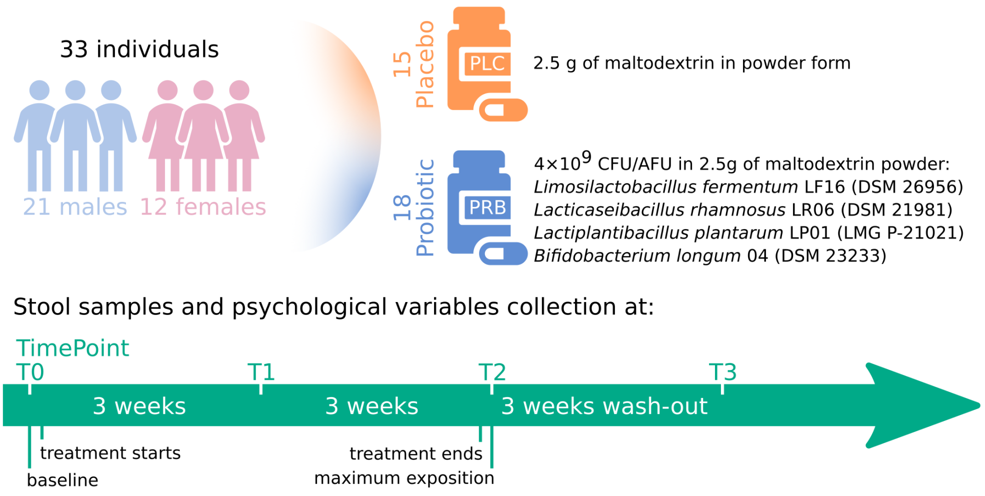

33 individuals
21 males
12 females
15 Placebo
PLC
2.5 g of maltodextrin in powder form
18 Probiotic
PRB
4×10^9 CFU/AFU in 2.5g of maltodextrin powder:
Limosilactobacillus fermentum LF16 (DSM 26956)
Lacticaseibacillus rhamnosus LR06 (DSM 21981)
Lactiplantibacillus plantarum LP01 (LMG P-21021)
Bifidobacterium longum 04 (DSM 23233)
Stool samples and psychological variables collection at:
TimePoint
T0
T1
T2
T3
3 weeks
3 weeks
3 weeks wash-out
treatment starts
baseline
treatment ends
maximum exposition

Figure 5.1: Experimental design of the study.

5.3 Materials and Methods

5.3.1 Sample collection

The samples analysed derive from 33 healthy subjects enrolled in the study. Stool samples and psychological variables were collected at four time points (see Additional file: Supplementary Table S1): before the intake of probiotic/placebo (T0), at 3 (T1) and 6 (T2) weeks after the first intake and at the end of the third week of washout (T3).

The experimental group received 42 sachets of the product (one for each day), each containing 4×10^9 cfu/active fluorescent units (AFU) of four probiotic species: *L. fermentum* LF16 (DSM 26956), *L. rhamnosus* LR06 (DSM 21981), *L. plantarum* LP01 (LMG P-21021), and *B. longum* 04 (DSM 23233) in 2.5 g of freeze-dried powder mixture containing maltodextrin (around 85% of the total weight) (Probiotical S.p.A., Novara, Italy). The control group received 42 sachets of placebo, each containing 2.5 g of maltodextrin in powder form. The placebo powder was indistinguishable from the probiotics powder in colour, taste, and smell. Participants were instructed to dissolve the powder in water or milk and drink it in the morning with breakfast. The probiotic sachets were analysed by Biolab Research S.r.l. (Novara, Italy), via flow cytometry (ISO 19344:2015 IDF 232:2015,

$\geq 4 \times 10^9$ AFU) and plate count method (Biolab Research Method 014-06, $\geq 4 \times 10^9$ cfu) to confirm target cell count. Product stability was monitored to ensure minimum cell counts were maintained. The study was approved by the Ethical committee of Verona Hospital (Azienda Ospedaliera Universitaria Integrata, AOUI Verona, 766CESC) and it is registered in `https://ClinicalTrials.gov` with the number ID: NCT03539263.

5.3.2 Library preparation and sequencing

The 132 collected samples from 33 healthy subjects were stored at -20 °C until analysis. DNA extraction and sequencing were performed at BMR Genomics S.r.l. (Padua, Italy). DNA was isolated with the Mobio Powerfecal kit (Mo Bio Laboratories, Inc., Carlsbad, CA, USA) adapted for QIAcube HT extractor (Qiagen, Hilden, Germany). V3-V4 regions of 16S rRNA gene were amplified with previously described primers [18], modified with forward and reverse overhangs necessary for dual index library preparation generating amplicons of $\sim$ 460 bp. The paired-end sequencing of the 16S rRNA gene amplicons was performed using the MiSeq Illumina platform (dual-indexing approach, 2 × 300 bp) (Illumina, San Diego, CA, USA). A mock community was included as control. The resulting output was a set of 264 raw files in FASTQ format. All the reads have been submitted to SRA archive and are available under the bioproject PRJNA644097

5.3.3 Bioinformatics data analysis

The whole analysis was performed on R (v3.6.1, R Core Team, 2019). Primarily, the FASTQ sequences were analysed using DADA2 (v1.13) [19], a tool that implements an error correction model and allows to identify exact sample sequences that differ as little as a single nucleotide. The final output of DADA2 was an Amplicon Sequence Variants (ASV) table which recorded the number of times each ASV was observed in each sample. DADA2 was run as described in `https://benjjneb.github.io/dada2/bigdata.html` using default parameters. In order to improve the overall quality of the sequences, the reads were filtered and trimmed using the *filterAndTrim* function implemented in DADA2. Consequently, to remove low quality bases at the end of reads, the *truncLen* option was set to (280, 220) for the forward and reverse FASTQ files respectively. Similarly, to remove adapter sequences at the 5' end, the *trimLeft* option was set to (17, 21), for forward and reverse reads respectively. The *removeBimeraDenovo* function was used to remove chimeras, via consensus method, and then the *collapseNoMismatch* function collapsed together all the reads that are identical up to shifts or length variation. Finally, the taxonomic assignment was performed using the naïve Bayesian classifier method implemented in DADA2 (*assignTaxonomy* and *addSpecies* functions) using as reference the EzBioCloud 16S database for QIIME pipeline (version 2018.05, `https://ezbiocloud.net/resources/16s_download`), correctly formatted to work with the taxonomic classifier implemented within DADA2 (`https://benjjneb.github.io/dada2/assign.html`). A phylogenetic tree of the ASVs was obtained using the function *AlignSeq* implemented in DECIPHER (v2.12) [20], an R package to create multiple sequence alignments. FastTree (v2.1.10) [21] was used to create the final tree.

5.3.4 Data quality assessment and filtering

Rarefaction curves on raw data were evaluated to assess the species richness among samples as a function of the sequencing depth. Data were pre-processed filtering taxa (ASVs) with low prevalence (where prevalence is the fraction of total samples in which an ASV is observed), setting a threshold of 0.5% for the cumulative relative abundance across all the samples; furthermore, taxa present in less than 2 samples were discarded. Synergistetes phylum members taxa (cumulative relative abundance=0.34%) and Lentisphaerae phylum members taxa (cumulative relative abundance=0.03%) were discarded by this filter. The pre-processing output data were then transformed to their relative abundances, and the 10 most present genera were plotted to phylum level. Mann-Whitney tests were performed on ASVs detected in these genera and the Benjamini-Hochberg correction was applied to adjust the p-values because of multiple testing.

In order to investigate the presence of probiotic related taxa, a further taxonomy classification was performed. The softwares Kraken2 [22] and Bracken [23] were used to check both raw FASTQ data and DADA2 inferred list of ASV, using two different pre-built Kraken2/Bracken databases (minikraken2_v2_8GB_201904 and k2_standard_16gb_20200919) and a custom database containing bacteria, archaea, virus, fungi and plants sequences, built using RefSeq [24] sequences.

5.3.5 Biodiversity measurements

Shannon-Wiener index was used to calculate α-diversity, which was plotted stratifying the samples according to time points, gender and treatment type; the Kruskal-Wallis tests were performed to verify statistical differences in the α-diversity among the samples. To measure β-diversity, data were normalised by three different methods (Cumulative Sum Scaling [CSS], Total Sum Scaling [TSS], Rarefaction) through the *phyloseq_transform_-css*, *phyloseq_standardize_otu_abundance*, and *rarefy_even_depth* functions respectively. The first two functions are part of the vmikk/metagMisc package (`https://github.com/vmikk/metagMisc`) while the latter belongs to the phyloseq package (v1.30.0) [25]. Each type of normalised data was inspected using four different distance metrics (Unweighted UniFrac, Weighted UniFrac, Bray-Curtis, Jaccard) and ordinated using the Principal Coordinates Analysis (PCoA) and Detrended Correspondence Analysis (DCA) ordination methods, through the ordinate function of the Vegan package (v2.5-5) [26].

A rigorous procedure was applied to evaluate the best combination of normalisation, distance metric, and ordination method. Normalisation based on rarefaction was not considered as it performs very similarly to TSS due to the similar library sizes between samples.

At first, a hierarchical clustering was applied to the β-diversity bidimensional plot at the baseline grouping the samples in 3 and 4 groups. To test which of the two clustering methods performed better, the homogeneity of the cluster dispersions were tested using PERMANOVA F-test on *betadisper* function's output. A significant p-value indicated that the cluster dispersions were not homogeneous and that data needed to be taken with care. Secondly, the silhouette value was calculated, that is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranged from -1 to +1, where a high value indicated that the object was well matched to its own cluster and poorly matched to neighbouring clusters. If most objects had a high value, then the clustering configuration was appropriate. On the other hand, if many samples had a low or negative value, then the clustering configuration might have too many or too few clusters. Finally, the cluster memberships found at the baseline were extended to all the other time points; cluster dispersions and silhouette indexes were computed again to verify the performances of the clustering on the whole dataset.

5.3.6 Mixed effects regression models statistical analysis

Amongst all the tested combinations, the TSS-normalised data, ordinated using the PCoA method and the unweighted-UniFrac distance metric, showed the most consistent results in cluster dispersions homogeneity and silhouettes, hence it was chosen for deeper exploration. To investigate the biological meaning of each PCoA coordinate, mixed effects regression model analysis was performed on each, using the *lme* function of the nlme package (v3.1-140) [27].

Firstly, the model formulation involved the *Sample* variable as a random component for each individual, and several categorical variables as fixed effects, such as *TimePoint*, *Gender*, *Treatment* and their interactions. Since all the variables were categorical, the regression framework set a baseline formed by *TimePoint=T0*, *Treatment=Placebo* and *Gender=Female* samples. Variable significance was guaranteed through an iterative process. Starting from the complete model, nonsignificant variables were dropped one by one. Every time a variable was dropped a Likelihood Ratio Test (LRT) was performed in order to compare the likelihood of the model with the likelihood of the nested one (p-value<0.1). This procedure allowed us to reach the most informative as well as parsimonious formulation of the model. Moreover, two versions of each model were compared: the first, where no correlation structure was specified, and the second, where the type of correlation was specified as an AR(1) process through the option *correlation=corAR1(form= 1|Sample)* of *lme* function.

Secondly, mixed effects regression models were used to study correlation between sample variables and the PCoA components, with the new information about cluster memberships. The *TimePoint*, *Gender*, *Treatment* and *Cluster* variables were tested in the model, together with the interactions between *TimePoint* and *Gender*, *TimePoint* and *Treatment*, *TimePoint* and *Cluster*, *Treatment* and *Cluster*, and *TimePoint*, *Treatment* and *Cluster*. The already described model selection procedure was performed to choose the best model.

5.3.7 Biomarkers investigation

To retrieve information about the most discriminant features (Amplicon Sequence Variantss, ASVs) of the clusters identified with the hierarchical clustering procedure, a discriminant analysis was computed using PLS-DA and sPLS-DA methods. Following the default mixOmics (v6.8) [28, 29] pipeline (`https://mixomics.org/case-studies/splsda-srbct/`), a pseudo-count value of 1 was added to the counts table, which was then normalised with TSS and CLR transformed. At first, the pipeline was performed on the clusters at baseline T0 to identify the most discriminant ASVs of each group. The discriminant analysis was then applied to each significant interaction resulted from the mixed effects models, to investigate the effect of treatments. For each interaction a summary image was plotted using the *HotLoadings* function of the homonym package (`https://github.com/mcalgaro93/HotLoadings`), displaying the discriminant ASVs loadings and the related heatmap.

5.3.8 Psychological variables analysis

To find significant associations between psychological variables, treatments and clusters, Wilcoxon Rank Sum tests were performed between time points T0 and T1, T0 and T2, and T0 and T3 for placebo and probiotics groups. The p-values were also corrected for multiple testing using the Benjamini-Hochberg correction method.

5.4 Results

5.4.1 16S metabarcoding sequencing depth and taxonomy classification

A total of 5,382,700 paired-end sequences (an average of 40,778 reads per sample) with a read length of 300 bp were obtained from the samples of the 33 subjects summarised in Fig. 5.1. After read quality assessment, denoising and chimera filtering, 1,728 different ASVs were obtained. ASVs artefacts were removed with several filters and a total of 730 unique ASVs were obtained (Additional file: Supplementary Fig. S1). The taxonomy classification allowed to identify 10 phyla, 20 classes (730 ASVs), 27 orders (728 ASVs), 46 families (727 ASVs), 170 genera (720 ASVs) and 263 species (273 ASVs). The comparison of rarefaction curves (Additional file: Supplementary Fig. S2) as a function of sampling depth showed that all curves are close to saturation, therefore the richness of the samples has been fully observed or sequenced. The only exception was for subject number 8 at time point T2 that had a library size of 538, while the second lower had a value of 8,848; for this reason, the former was discarded from the analysis.

The most abundant phylum was Firmicutes, with a relative frequency of 62.3% followed by Bacteroidetes 17.9%, Proteobacteria 9.1%, Verrucomicrobia 4.9%, and Actinobacteria 4.7%. The remaining 1% accounted for Euryarchaeota, Tenericutes, Saccharibacteria, Fusobacteria and Cyanobacteria. At genus level, the most abundant populations were *Agathobacter*, *Blautia*, *Dialister*, *Faecalibacterium*, *Ruminococcus*, *Subdoligranulum* (belonging to Firmicutes phylum), *Bacteroides* (Bacteroidetes phylum), *Escherichia*(Proteobacteria phylum), *Akkermansia* (Verrucomicrobia phylum), and *Bifidobacterium* (Actinobacteria phylum) (Additional file: Supplementary Fig. S3).

5.4.2 α-diversity analysis confirmed that the subjects of the cohort were comparable

Samples were stratified according to *TimePoint*, *Treatment* (placebo or probiotics) and *Gender* using Shannon-Wiener index, as shown in Additional file: Supplementary Fig. S4. No significant differences among samples were observed, neither in the experimental nor in the control group (Kruskal-Wallis tests had p-value>0.05). This finding was in line with expectations, as the subjects enrolled in the study were comparable when related to their internal diversity; neither alterations nor major shifts were expected on gut microbiota species richness or evenness regarding probiotic consumers.

5.4.3 β-diversity analysis revealed three clusters and a strong sample-specific effect

The flow chart in Additional file: Supplementary Fig. S5, summarises the following steps of the analysis. All the β-diversity plots are shown in Additional file: Supplementary Fig. S6, while the homogeneity of cluster dispersions and silhouettes are presented in the Additional file: Supplementary Results S1. The choice of the number of clusters was performed using only the samples at T0, which represents a snapshot of the microbiome composition before any type of treatment and allows to stratify the samples according to different bacteriotypes.

The identification of the best combination of number of clusters, normalisations, distances, and type of ordinations was then chosen. Specifically, the metrics that performed better in terms of homogeneity of cluster dispersions and silhouette values, when the cluster membership was extended also to all the other time points, were selected. Three clusters grouping with PCoA ordination method, based on unweighted UniFrac distances and TSS normalisation (Fig. 5.2 a-c), produced the most consistent results (see Materials and Methods - Biodiversity measurements and Additional file: Supplementary Results S1, for details). Indeed, using these combinations of metrics, the stability of the clusters was maximised over time. In other words, the clusters identified at T0 tended to be the most consistent when the information of cluster membership is extended also to the other time points. The underlying idea was that the microbial signatures of the bacteriotype we identified at T0 should be stable over time, even though individual hosts may switch between enterotypes over long time periods [30].

The composition of each cluster is reported in Table 5.1. As expected, samples of the same subject tended to form close subclusters, regardless the considered time point or treatment (Additional file: Supplementary Fig. S7). This suggests that the differences among subjects are stronger than the effects determined by the treatment.

Cluster	**Placebo**	**Probiotic**
1	9	12
2	3	3
3	3	3

Table 5.1: Cluster membership for individuals at the baseline (*TimePoint=T0*).

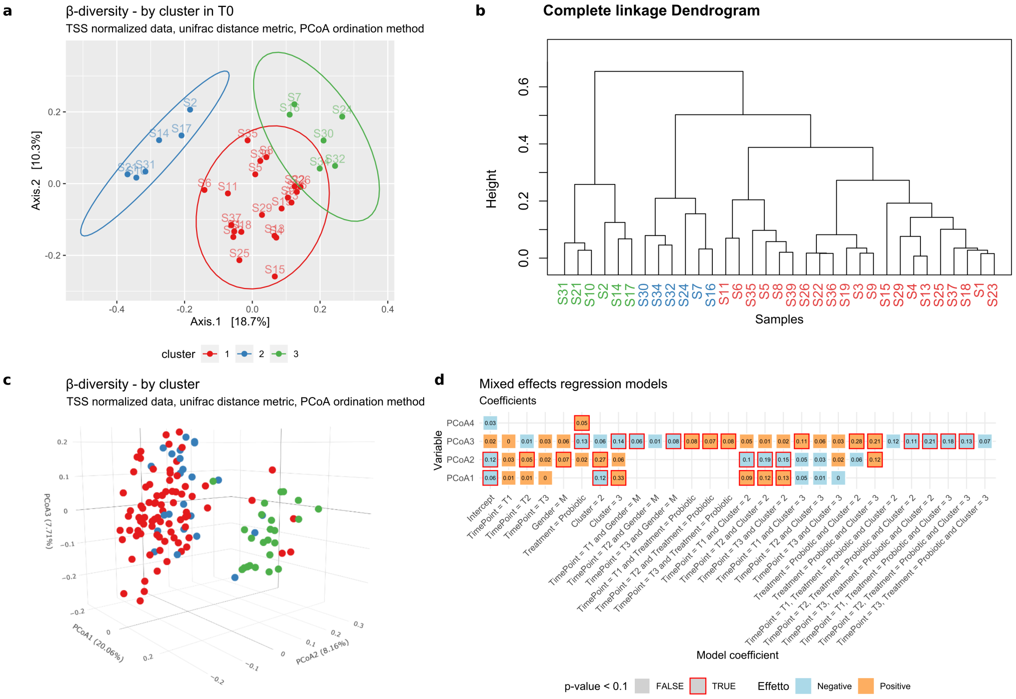


Figure 5.2: **a** Bidimensional representation of β-diversity for the samples in *TimePoint=T0* (PCoA ordination method on UniFrac distance matrix of TSS normalised count data). Coloured by cluster membership obtained cutting the dendrogram in **b** in order to obtain 3 groups of individuals (see Materials and Methods for details of normalisation, distance, and ordination's choice). **b** Hierarchical clustering dendrogram built with the complete linkage method on the euclidean distance matrix. Distance matrix based on PCoA1 and PCoA2 coordinates of the β-diversity for the samples in *TimePoint=T0* (PCoA ordination method on UniFrac distance matrix of TSS normalised count data). **c** Tridimensional representation of β-diversity coloured by cluster membership for the samples in all time points (PCoA ordination method on UniFrac distance matrix of TSS normalised count data). **d** Linear mixed effects regression model coefficients. Blue (orange) coloured tiles represent a negative (positive) effect of the variable referred to the model baseline (*TimePoint=T0*, *Gender=F*, *Treatment=Placebo*, *Cluster=1*). Statistically significant (p-value<0.1) effects are represented by red squared tiles.

5.4.4 Mixed effects regression models found associations between β-diversity and sample variables

To inspect the variability held by the first four coordinates of the PCoA, four mixed effects regression models were at first estimated without considering cluster membership (see Materials and Methods). This regression framework allowed us to find significant correlations between PCoA coordinates and metadata such as *Gender*, *Treatment* and *TimePoint* and to remove sample-specific effects. In this context, we implicitly considered the *Treatment=Placebo*, *Gender=Female*, and *TimePoint=T0* as the baseline level. As shown in Additional file: Supplementary Fig. S8, a correlation between the *Treatment* variable and the fourth component of the PCoA was found, while the third component showed a statistically significant difference in *Gender* at time point T3 compared to the baseline. This first analysis did not allow us to identify any statistically significant effects for the interactions between time points and treatments.

To further investigate if adding bacteryotype information would help in identifying significant effects for the interactions between time points and treatments, new mixed effects regression models were estimated adding the cluster membership variable as a fixed effect to the framework. Several significant interactions between time points and treatments were found for the third coordinate (Fig. 5.2 d): (1) *TimePoint=T1*, *Treatment=Probiotic* and *Cluster=3*; (2) *TimePoint=T2*, *Treatment=Probiotic* and *Cluster=2*; (3) *TimePoint=T2*, *Treatment=Probiotic* and *Cluster=3*; (4) *TimePoint=T3*, *Treatment=Probiotic* and *Cluster=2*. Clusters 2 and 3 were commonly affected by the variable *TimePoint=T2* compared to the baseline: Cluster 2 responded later in the treatment (time points T2 and T3) while Cluster 3 responded at the beginning (time points T1 and T2). Although each interaction should be interpreted very carefully, these results highlighted a difference between the considered variable categories and the baseline (T0, Placebo, Cluster 1). Biologically speaking, the identified interactions could be an indicator of a distinct effect of the treatment considering different groups/bacteriotypes.

5.4.5 sPLS-DA analysis showed that clusters were characterised by a specific bacteriotype

A sPLS-DA analysis was performed to identify the most discriminant ASVs at the baseline T0. This multivariate approach identified two main components which were able to discriminate the clusters. The first component highlighted 5 taxa associated with Cluster 3 (Fig. 5.3 a): all the members of this group were characterised by the presence of SV33, assigned to *Methanobrevibacter smithii*, while 66% of them also displayed *Sporobacter*, *Eubacterium*, and *Oscillibacter* spp. (SV168, SV256, SV37).

The second component highlighted the top 30 taxa associated with Clusters 1, 2 and 3 which created two different patterns as shown in the heatmap (Fig. 5.3 b). Cluster 1 individuals showed the general presence of *Faecalibacterium* spp. (SV3, SV14), while Cluster 3 were also characterised by *Faecalibacterium* spp. and *Alistipesputredinis* (SV4 and SV39); in Cluster 2, the second component revealed the presence of SV94-*Eubacterium* and SV77-Lachnospiraceae in almost all the members; 50% of them were also characterised by *Blautia* spp. (SV562).

5.4.6 *Lacticaseibacillus rhamnosus* is the only probiotic SV that increases significantly in the probiotic cohort in Cluster 1 and 2

The sPLS-DA analysis revealed that SV232, associated with *L. rhamnosus*, was present in the probiotic cohort at time point T1 and T2 in Cluster 1 and Cluster 2, respectively (Additional file: Supplementary Results S2 a, g). As for bifidobacteria, SV34 associated with *B. longum* was found to increase in Cluster 1 at T2 (where it was abundant also in the placebo individuals) and T3 (Additional file: Supplementary Results S2 e, k). Interestingly, other Sequence Variant (SV)s associated with *Bifidobacterium* spp. displayed a

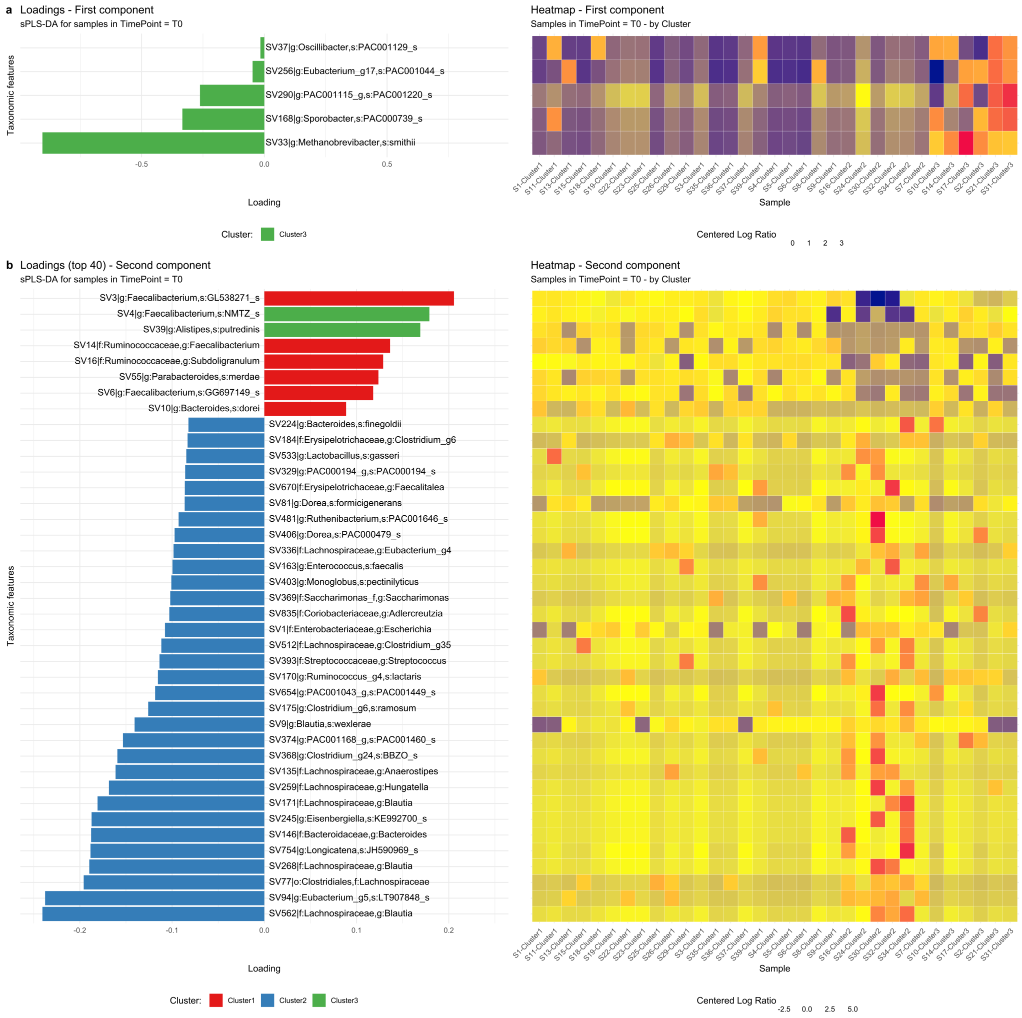


Figure 5.3: **a** sPLS-DA analysis at the baseline (*TimePoint=T0*). Loading values represent the 5 discriminant taxa of the first component, associated with Cluster 3. Bigger the loading absolute value, stronger the association. Heatmap shows the CLR values of the discriminant taxa in all the samples. **b** sPLS-DA analysis at the baseline (*TimePoint=T0*). Loading values represent the first 30 (out of 135) most discriminant taxa of the second component, associated with Cluster 2 and Cluster 1. Bigger the loading absolute value, stronger the association. Heatmap shows the CLR values of the discriminant taxa in all the samples.

different behaviour: SV228 was found to increase in the placebo cohort in Cluster 1 at T1 (Additional file: Supplementary Results S2 b), while the relative abundance of SV121 and SV15 decreased in Cluster 1 and 2 at T3 (Additional file: Supplementary Results S2 l, m). These observations showed that *L. rhamnosus* is the only probiotic SV that increases during the probiotic administration until T1 and T2 in Cluster 1 and Cluster 2, respectively.

5.4.7 Bacteriotypes changed distinctly in the probiotic and placebo cohorts

The probiotic intake in Cluster 1 was associated with an increase of *Coproiciproducens leptum* (*Clostridium letpum*), *Romboutsia timonensis* and *Mogibacterium* spp. (SV264, SV25 and SV664) from T1 to T3, respectively; the same cohort displayed a decrease of SVs related to *Butyricimonas* (SV785), *Lachnospira* (SV144) and *Faecalibacterium* spp. (SV4) at the same time points (Additional file: Supplementary Results S2 a, e, k). The placebo individuals featured a decrease of *Butyricimonas*, *Alistipes*, and *Ruthenibacterium lactatiformans* (SV774, SV465 and SV90) and a higher abundance of *Anaerotignum* (SV370), *S. thermophilus* (ST32) and *Turicibacter* spp. (SV56) from T1 to T3 (Additional file: Supplementary Results S2 b, f, l).
Individuals who took probiotics in Cluster 2 showed a significant decrease of Ruminococcaceae (SV348) in T2 (Additional file: Supplementary Results S2 g) while the placebo group were characterised by an increment of *Alistipesonderdonkii* (SV73) and Lachnospiraceae spp. (SV144) in T2 and T3, respectively, and a drop of *Blautia* spp. (SV562) and *Clostridium* spp. (SV512) in the same time points (Additional file: Supplementary Results S2 h, n).

In Cluster 3, *Phascolarctobacterium faecium* (SV97) and *Subdoligranulum* spp. (SV16) distinguished the probiotic cohort at T1 and T2 which, conversely, showed negative CLR values for *Dialister invisus* (SV5) and *Eubacterium* spp. (SV256) at the same time points (Additional file: Supplementary Results S2 d, j); this latter species (SV94) increased together with *Roseburia hominis* (SV161) in the placebo subjects, which also showed a decrease of *Intestinibacter bartlettii* (SV63) and *Bacteroides* (SV47) at T1 and T2, respectively (Additional file: Supplementary Results S2 c, i).

5.4.8 Maltodextrin exerted an effect on the bacteriotype of each cluster

Since maltodextrins are included both in the placebo and in the probiotic products, their impact on each cluster's bacteriotype (included Cluster 1) was investigated (Additional file: Supplementary Results S3). Focusing on SVs related to probiotics, SV34 - *B. longum* generally increased in members of Cluster 1 at T2 and T3 and in Cluster 2 at T1 (Additional file: Supplementary Results S3 b, d, g); as for SVs related to other *Bifidobacterium* spp. a general reduction of SV121 and SV15 was observed in both probiotic and placebo groups in Cluster 1 at T2 and in Cluster 2 at T3 (Additional file: Supplementary Results S3 g, h).

Considering other taxa, Cluster 1 was characterised by a general increase in relative abundance of *S. thermophilus* (SV32), *R. timonensis* (SV25), *Turicibacter* spp. (SV56), and a decrease of *Butyricimonas* spp. (SV774) and *Lachnospira* spp. (SV144) from T1 to T3 (Additional file: Supplementary Results S3 a, d, g).

Cluster 2 individuals were characterised by higher levels of *Faecalibacterium* (SV14), and *Roseburiainulinivorans* (SV54) at T1 and T3; while SVs related to *Escherichia* (SV1), *Agathobaculum* (SV87), *Blautia* (SV17), and *Eubacterium* (SV94) decreased from T1 to T3 (Additional file: Supplementary Results S3 b, e, h).

Finally, in Cluster 3 positive CLR values were associated to *Anaerotignum*, *Pseudoflavonifractor* and *Sporobacter* (SV282, SV428, SV145) while negative values were related to *D. invisus* (SV5), *Bacteroides* (SV140) and *Blautia obeum* (SV29) (Additional file: Supplementary Results S3 c, f, i).

5.4.9 Sequence variants related to *Limosilactobacillus fermentum* were detected in only one individual treated with probiotics

ASVs associated with *L. fermentum* and *L. plantarum*, included in the probiotic product, were investigated and checked through the 16S-based ID tool of EzBioCloud.net, `https://ezbiocloud.net/identify`; database version 2020.10.12). SV1273 associated with *L. fermentum* was detected only in one probiotic cohort sample at T1, while conflicting results were obtained using different databases related to *L. plantarum*, confirming that the V3-V4 region for this species is not informative (Fig. 5.4).

5.4.10 Sleep quality and mood changes were detected in probiotics treated individuals of Cluster 1

As shown in Fig. 5.5 a, a significant reduction (p-value=0.03) was detected between time points T0-T1 and confirmed for T0-T2 and T0-T3 for the Pittsburgh Sleep Quality Index (PSQI). The PSQI global score is inversely correlated to the sleep quality (the lower the score, the better the sleep quality). The identified reduction indicates a sleep quality improvement for the probiotics treated individuals of Cluster 1.

Other significant changes were detected for the depression, anger, and fatigue subscales of the Profile Of Mood State (POMS) psychological variables. Specifically, between T0-T1 and T0-T3 for anger (p-value=0.08, 0.02; Fig. 5.5 b) and depression (p-value=0.08, 0.06; Fig. 5.5 c) indexes, and between T0-T1 (p-value=0.04), T0-T2 (p-value=0.02), and T0-T3 (p-value=0.03) for the fatigue subscale (Fig. 5.5 d). It is noteworthy a clear descending trend for all mentioned psychological variables also in Cluster 3, even though these differences were not significant, probably due to the low sample size of the cluster. A similar pattern was not visible in Cluster 2.

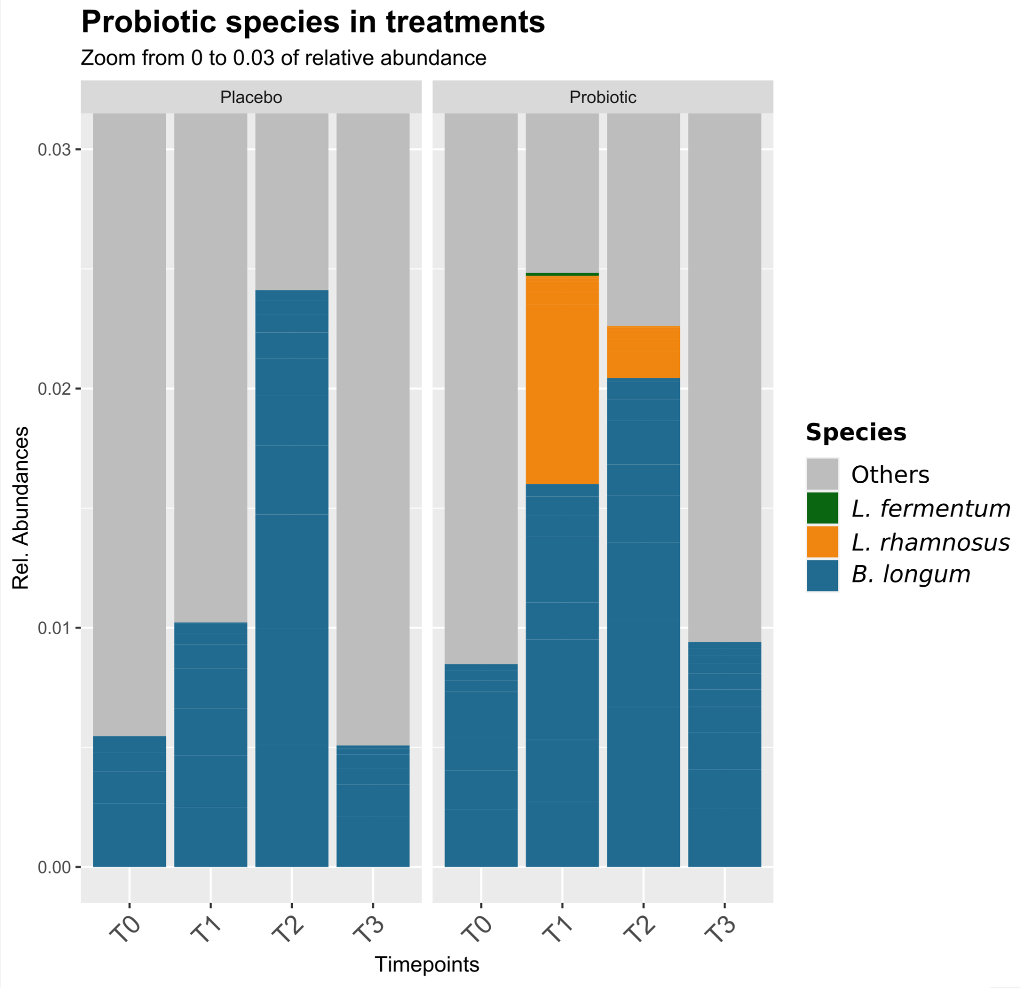


Figure 5.4: The species related to the probiotic compound were isolated and plotted in this barplot. The relative abundances percentages were zoomed to visualise the portion from 0 to 0.3 and stratified by time points and treatment type. The *Bifidobacterium longum* species is present in each time point for both the treatment types, showing a shared increasing trend. *Limosilactobacillus fermentum* was detected only for the second time point (T1) in the probiotic cohort. *Lacticaseibacillus rhamnosus* taxa were detected for both the second and the third time points, relative to the probiotic cohort.

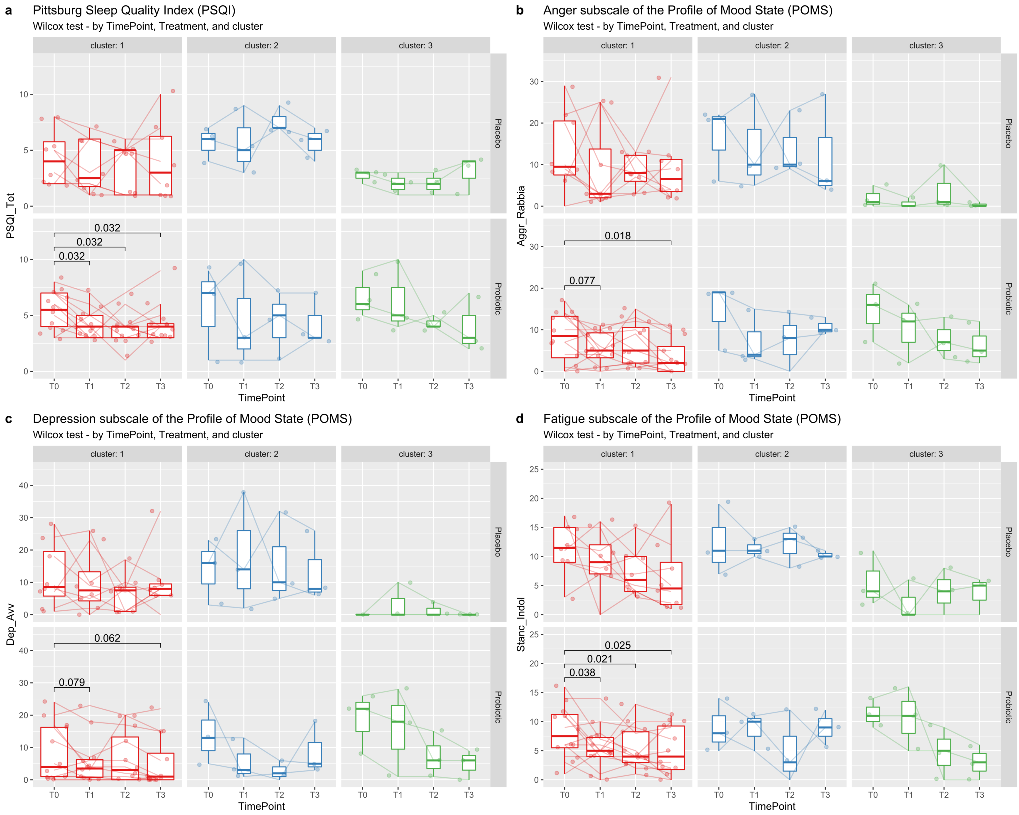


Figure 5.5: Wilcoxon Rank Sum tests between time points T0-T1, T0-T2, T0-T3. P-values are corrected for multiple testing using the Benjamini-Hochberg correction method and only adjusted p-values lower than 0.1 are reported. **a** PSQI, stratified by cluster and treatment. **b** Anger subscale for the POMS psychological variable, stratified by cluster and treatment. **c** Depression subscale for the POMS psychological variable, stratified by cluster and treatment. **d** Fatigue subscale for the POMS psychological variable, stratified by cluster and treatment.

5.5 Discussion

5.5.1 Biodiversity measures stratified individuals in three clusters related to their microbiota

In the present work the possible effect caused by the intake of *B. longum*, *L. fermentum*, *L. rhamnosus* and *L. plantarum* strains for 6 weeks (followed by a 3 week-washout) on the gut microbiota composition of a cohort of 33 healthy subjects was investigated.

A robust bioinformatic pipeline was implemented to analyse and characterise the metabarcoding data; a series of exploratory analyses were performed targeting particular effects with a possible biological correspondence, which could be related to the cognitive and emotional improvements we assessed in our previous study [5]. Biodiversity measures did not detect a significant diversity within the samples (α-diversity) but a sample-specific effect was found between samples (β-diversity). This finding led to perform a statistical analysis using a mixed effects model, through which several minor significant effects were found. The 33 individuals were clustered into three groups/bacteriotypes at the baseline; each one of them responded distinctly to the treatments. Cluster 1 and 2 were most impacted by the probiotic treatment, while Cluster 3 responded more to the placebo treatment. Furthermore, Cluster 1 responded throughout the whole treatment, while Cluster 2 and Cluster 3 had, respectively, a late and an early response. Those behaviours reinforce the concept that individuals with diverse bacteriotypes might respond differently to the same treatment. In addition, stratification of individuals according to their bacterial composition may be useful to better understand and predict the responses to specific treatments, such as probiotic interventions [31, 32].

5.5.2 Maltodextrin has a bifidogenic effect

Focusing on SVs related to the probiotic species, it was observed that some *Bifidobacterium*-related SVs (SV15 and SV121) decreased both in the placebo and probiotic groups, but others, such as SV34 - *B. longum* and SV228 - *Bifidobacterium* significantly increased in the placebo cohort both in Cluster 1 and Cluster 2. Their presence in the control subjects, who were administered maltodextrin, is also in line with data reported in a previous work [33] where authors observed that the majority of the culturable bifidobacterial strains (including 10 strains of *B. longum*) were capable of growing in maltodextrin rich media. Tandon and colleagues [34] observed the same behaviour of the bifidobacterial population in a randomized, double-blind, placebo-controlled, dose-response relationship study led to investigate the efficacy of fructo-oligosaccharides on human gut microbiota, where maltodextrin was used as control.

From this perspective, this study does not include a true placebo cohort which may have prevented to capture time dependent oscillations in abundances of relevant taxa; further, since maltodextrins are broadly used as placebo treatment, their bifidogenic effect needs to be deeply evaluated in future clinical trials involving bifidobacteria. Considering our data, we suggest that participants of this study were subjected to two different treatments rather than one: a "synbiotic" administration (probiotics and maltodextrin) and a "prebiotic" assumption (maltodextrin).

5.5.3 Cluster 1 individuals displayed different gut composition following the prebiotic and synbiotic treatments

The gut composition of individuals who received the synbiotics in Cluster 1 are selectively characterised by the presence of *L. rhamnosus* and *C. leptum* at the beginning of the treatments and then of *Mogibacterium* (described in 2000 to include strains isolated from the human periodontal environment

[35]) at the end. *C. leptum* belongs to *Clostridium Cluster IV* which was reduced in patients with depression and anhedonia and it was negatively associated with scores in Quick Inventory of Depressive Symptoms-Self-Rated (QIDS-SR) and the Generalized Anxiety Disorder (GAD)-7 [36]. The effect of probiotics administration on the abundance of *C. leptum* was also observed by Sato and colleagues [37] where patients with type-2 diabetes had higher counts of *C. leptum* after 16 weeks of probiotics (L. casei Shirota) assumption. The synbiotic treatment specifically reduced the levels of *Faecalibacterium*, which was among the signature taxa of this cluster. This taxon is usually lower in MDD patients but there is still a lack of congruence across investigations [3, 38].

The prebiotic treatment is associated with higher levels of *Anaerotignum* (described in 2017 following the isolation of strains from a methanogenic reactor [39]), *S. thermophilus*, *Turicibacter* and *D. invisus.* Among these species, it is interesting to report that a reduced abundance of *Turicibacter* spp. was observed in socially defeated mice and it was strongly correlated to pro-inflammatory cytokine changes within the prefrontal cortex [40]. However, this has to be further investigated, as *Turicibacter* levels were also found to be higher in depressed subjects [41]. As for *D. invisus*, it is usually lower in MDD patients and in other autoimmune diseases, including Crohn's disease, ulcerative colitis and rheumatoid arthritis [42].

Prebiotic administration was also related to reduced levels of *Butyricimonas*, *Lachnospira*, *Alistipes* and *R. lactatiformans* (isolated in 2016 from human faeces [43]). The decrease of *Butyricimonas* spp. may have a positive effect on the individuals: *Butyricimonas* members were found to be at higher levels in the gut microbiota of patients with clinically significant depression compared to control patients [44]. As for *Lachnospira*, no consensus data have been obtained so far, as Cheung and colleagues reported that this taxon could be related to MDD as well as to healthy subjects [3, 38].

A significant association with depression was shown for *Alistipes* both in human cases as well as in mice subjected to stress over an extended time period. Since high levels of this taxon in the gut microbiota were also linked to chronic fatigue syndrome and Irritable Bowel Syndrome (IBS), it has been suggested that *Alistipes* may promote depression through inflammatory pathways. In addition, *Alistipes* species are indole-positive and may thus influence tryptophan availability (the precursor of serotonin), disrupting the balance in the intestinal serotonergic system [45]. They are also high metabolisers of proteins and amino acids and, as such, they could trigger the production of toxic compounds such as ammonia, putrescine, and phenol [3]. However, these data are not concordant with what observed from Zheng and colleagues who reported that *Alistipes* were overrepresented in healthy control subjects compared to patients diagnosed with MDD [46].

5.5.4 Prebiotic treatment in Cluster 2 has an opposite effect of the synbiotic in Cluster 1

In Cluster 2, the prebiotic supplementation induced a general oscillation of the abundance levels of *Eubacterium*, *Blautia* and Lachnospiraceae spp. which characterised the bacteriotype of this group. Kim and colleagues [47] suggested that the reduction in the relative abundances of *Eubacterium* is related to the increase of the brain-derived neurotrophic factor in the serum, improving brain functions. On the contrary respect to Cluster 1, the comparison between the prebiotic and synbiotic treatments showed that *Alistipes* spp. and *Lachnospira* spp. increased in Cluster 2 individuals.

The synbiotic intervention specifically led to a lower abundance of Ruminococcaceae (heterotypic synonym of family Oscillispiraceae): at family level, it was observed that these taxa were lower in depressed subjects compared to the control group [44] and were correlated with behavioural changes induced by stress in mice [48]. Conversely, prebiotics reduced the levels of *Blautia*, *Clostridium*, *Escherichia*, and *Agathobaculum* spp. although no data have been reported yet on the association of *Agathobaculum* (a strictly anaerobic and butyrate-producing strain isolated from the faeces of a healthy 23-year-old Korean female [49]), with stress-related disorders and there is a lack of consensus related to the presence of *Escherichia* in MDD patients, this effect could be considered beneficial for this cluster, as both *Blautia* spp. and *Clostridium* are usually found at higher levels in patients with MDD [3, 44]. Conversely, *Faecalibacterium* and *R. inulinivorans* increased at the end of prebiotic treatment: this species has been shown to have beneficial effects in specific conditions (*i.e.*, atherosclerosis, [50]) but no particular correlation has been found with mental or stress-related disorders [3].

5.5.5 Treatments in Cluster 3 changed the relative abundance of *Phascholarctobacterium faecium*, *Subdoligranulum*, and *Eubacterium*

The assumption of the synbiotic in Cluster 3 individuals increased the relative abundance of *Subdoligranulum* and *P. faecium* and *Eubacterium* spp. (which characterised the bacteriotype at the baseline). It is interesting to note that *P. faecium* and, in general, family Acidaminococcaceae are more correlated to patients with active MDD [44] and with both IBS and depression [51] rather that with healthy subjects. On the contrary, *Subdoligranulum* are depleted in subjects with IBS and depression, so its presence in the synbiotic cohort can be interpreted as a positive effect of this treatment [52]. This taxon is capable of producing Short Chain Fatty Acids (SCFAs) (in particular butyrate) that protect the intestinal mucosa and regulate the immune system. More specifically, SCFAs play an important role in the differentiation of T cells and as histone deacetylase inhibitors, which were found to have immunosuppressive and anti-inflammatory functions and have been explored as potential novel antidepressants [3]. *R. hominis*, *Anaerotignum* spp. (similarly to Cluster 1), *Pseudoflavonifractor*, *Eubacterium* (conversely to the synbiotic treatment) and *Sporobacter* (among the signature taxa of this cluster) were significantly abundant following the prebiotic treatment while *I. bartletti*, *Bacteroides* and *B. obeum* decreased. *Sporobacter* and

Pseudoflavonifractor are among the common taxa found in the human gut microbiota; as for *R. hominis*, although no data are available regarding the direct positive or negative connection of this species and stress-related disorders, the reduction in the abundance of butyrate-producing Lachnospiraceae members, (including *R. hominis*) which are beneficial for the integrity and function of intestinal barrier, was involved in the formation of stress-induced visceral hypersensitivity for which *R. hominis* was proposed as a candidate potential probiotic [53]. As for *I. bartlettii*, it is interesting to report that it was found more frequently in the faecal samples of children with neurodevelopmental disorders compared to the control subjects [54]. Finally, *Bacteroides* spp. exhibited divergent directionality and were found to be associated both with MDD as well as with healthy status, so no conclusions can be made on the effect of their reduction in this cohort [38].

5.6 Conclusions

Although no consensus observations on the biological significance of particular components of the gut microbiota on mood disorders have been obtained yet, the present study shows that both the "synbiotic" and the "prebiotic" intake over a period of 6 weeks significantly changed the composition of the gut microbiota.

A debate is still ongoing whether the probiotic supplementation alters successfully the microbiota composition [55]; in this perspective, Pinto-Sanchez and colleagues [9] observed that probiotic administration in patients with IBS led to changes in urine metabolic profiles, brain activity and to antidepressant effects, but no detectable effects on the gut microbiota composition were noticed. Nevertheless, it has been demonstrated that probiotic treatments impact on both the gut microbiota gene expression (with potential anti-inflammatory effects) and the gut barrier function (as also shown by the probiotic strains used in the present study – M. Pane, personal communication); which can lead to an effect on the cognitive function [8].

Overall, this study offers evidence that probiotics supplementation has variable impacts depending on the gut microbiota bacteriotypes (*i.e.*, Cluster 1, 2 and 3). In some cases, the shifts were towards microbial populations generally related to a healthy mood status (*i.e.*, higher abundance of *C. leptum* and *Subdoligranulum* in Cluster 1 and 3, respectively) and suggests some mechanisms (*i.e.*, SCFA production related to *Subdoligranulum*) which might rationalise the positive effects of the supplementation on the depressive mood state and sleep quality we observed in our previous work [5]. Particularly, variations of microbiota compositions were found to be statistically related to sleep quality improvement and to a descending rate of depression, anger and fatigue in probiotic-treated individuals of Cluster 1.

Overall, these findings should be interpreted with caution: first of all, further studies are necessary on a larger and more homogeneous cohort of individuals, taking fully into account the effects of gender, diet, body mass index, presence of inflammation, bacteriotypes, and other factors that may be important covariates affecting the faecal microbiota.

As for diet, it is well established that it is one of the major modulators of the microbiota, therefore its monitoring is of utmost interest to better understand microbial dynamics and link them to other metabolic and physiological parameters [56]. However, the monitoring of young healthy individuals for 6 weeks (9 including washout) proved to be a very challenging task, with too partial data that could not be used for associations.

Indeed, in this study, we tried to move from an effectiveness perspective (the whole cohort) to an efficacy-focused one (the clusters/bacteriotypes), revealing some complexities on the microbial background related to the effects described by Marotta *et al.* [5] on almost the same cohort. Shedding light on these variables, especially on a healthy individual's cohort, is expected to allow a better development of psychobiotic treatment strategies. This will contribute to the definition of probiotics as an adjunct therapy or for the prevention of mood-related disorders.

References

1. Bastiaanssen, T. F. S. *et al.* Gutted! Unraveling the Role of the Microbiome in Major Depressive Disorder. *Harvard review of psychiatry* **28,** 26–39 (2020).
2. Kelly, J. R., Keane, V. O., Cryan, J. F., Clarke, G. & Dinan, T. G. Mood and Microbes: Gut to Brain Communication in Depression. *Gastroenterology clinics of North America* **48,** 389–405 (2019).
3. Cheung, S. G. *et al.* Systematic review of gut microbiota and major depression. *Frontiers in Psychiatry* **10** (2019).
4. Butler, M. I., Sandhu, K., Cryan, J. F. & Dinan, T. G. From isoniazid to psychobiotics: The gut microbiome as a new antidepressant target. *British journal of hospital medicine* **80,** 139–145 (2019).
5. Marotta, A. *et al.* Effects of probiotics on cognitive reactivity, mood, and sleep quality. *Frontiers in Psychiatry* **10** (2019).
6. Romijn, A. R., Rucklidge, J. J., Kuijer, R. G. & Frampton, C. A double-blind, randomized, placebo-controlled trial of Lactobacillus helveticus and Bifidobacterium longum for the symptoms of depression. *Australian and New Zealand Journal of Psychiatry* **51,** 810–821 (2017).
7. Akkasheh, G. *et al.* Clinical and metabolic response to probiotic administration in patients with major depressive disorder: A randomized, double-blind, placebo-controlled trial. *Nutrition* **32,** 315–320 (2016).
8. Chahwan, B. *et al.* Gut feelings: A randomised, triple-blind, placebo-controlled trial of probiotics for depressive symptoms. *Journal of affective disorders* **253,** 317–326 (2019).
9. Pinto-Sanchez, M. I. *et al.* Probiotic Bifidobacterium longum NCC3001 Reduces Depression Scores and Alters Brain Activity: A Pilot Study in Patients With Irritable Bowel Syndrome. *Gastroenterology* **153,** 448–459.e8 (2017).

10. Kazemi, A., Noorbala, A. A., Azam, K., Eskandari, M. H. & Djafarian, K. Effect of probiotic and prebiotic vs placebo on psychological outcomes in patients with major depressive disorder: A randomized clinical trial. *Clinical Nutrition* **38,** 522–528 (2019).

11. Eskandarzadeh, S. *et al.* Efficacy of a multispecies probiotic as adjunctive therapy in generalized anxiety disorder: a double blind, randomized, placebo-controlled trial. *Nutritional Neuroscience* **24,** 102–108 (2021).

12. Benton, D., Williams, C. & Brown, A. Impact of consuming a milk drink containing a probiotic on mood and cognition. *European journal of clinical nutrition* **61,** 355–361 (2007).

13. Steenbergen, L., Sellaro, R., van Hemert, S., Bosch, J. A. & Colzato, L. S. A randomized controlled trial to test the effect of multispecies probiotics on cognitive reactivity to sad mood. *Brain, behavior, and immunity* **48,** 258–264 (2015).

14. Messaoudi, M. *et al.* Assessment of psychotropic-like properties of a probiotic formulation (Lactobacillus helveticus R0052 and Bifidobacterium longum R0175) in rats and human subjects. *British Journal of Nutrition* **105,** 755–764 (2011).

15. Mohammadi, A. A. *et al.* The effects of probiotics on mental health and hypothalamic–pituitary–adrenal axis: A randomized, double-blind, placebo-controlled trial in petrochemical workers. *Nutritional Neuroscience* **19,** 387–395 (2016).

16. Nishida, K., Sawada, D., Kuwano, Y., Tanaka, H. & Rokutan, K. Health benefits of lactobacillus gasseri cp2305 tablets in young adults exposed to chronic stress: A randomized, double-blind, placebo-controlled study. *Nutrients* **11** (2019).

17. Zheng, J. *et al.* A taxonomic note on the genus Lactobacillus: Description of 23 novel genera, emended description of the genus Lactobacillus beijerinck 1901, and union of Lactobacillaceae and Leu-

conostocaceae. *International Journal of Systematic and Evolutionary Microbiology* **70,** 2782–2858 (2020).

18. Takahashi, S., Tomita, J., Nishioka, K., Hisada, T. & Nishijima, M. Development of a prokaryotic universal primer for simultaneous analysis of Bacteria and Archaea using next-generation sequencing. *PLoS ONE* **9** (2014).
19. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* **13,** 581–583 (2016).
20. Wright, E. S. Using DECIPHER v2.0 to analyze big biological sequence data in R. *R Journal* **8,** 352–359 (2016).
21. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5** (2010).
22. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome biology* **20** (2019).
23. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: Estimating species abundance in metagenomics data. *PeerJ Computer Science* **2017** (2017).
24. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic acids research* **44,** D733–D745 (D1 2016).
25. McMurdie, P. J. & Holmes, S. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **8** (2013).
26. Oksanen, J. *et al. vegan: Community Ecology Package* `https://CRAN.R-project.org/package=vegan`.
27. Pinheiro, J., Bates, D. & R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models* `https://CRAN.R-project.org/package=nlme`.
28. Lê Cao, K.-A. *et al.* MixMC: A Multivariate Statistical Framework to Gain Insight into Microbial Communities. *PLoS ONE* **11** (2016).

29. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology* **13** (2017).
30. Moeller, A. H. *et al.* Chimpanzees and humans harbour compositionally similar gut enterotypes. *Nature Communications* **3** (2012).
31. Cheng, M. & Ning, K. Stereotypes About Enterotype: the Old and New Ideas. *Genomics, Proteomics and Bioinformatics* **17,** 4–12 (2019).
32. Christensen, L., Roager, H. M., Astrup, A. & Hjorth, M. F. Microbial enterotypes in personalized nutrition and obesity management. *American Journal of Clinical Nutrition* **108,** 645–651 (2018).
33. Watson, D. *et al.* Selective carbohydrate utilization by lactobacilli and bifidobacteria. *Journal of applied microbiology* **114,** 1132–1146 (2013).
34. Tandon, D. *et al.* A prospective randomized, double-blind, placebo-controlled, dose-response relationship study to investigate efficacy of fructo-oligosaccharides (FOS) on human gut microflora. *Scientific Reports* **9** (2019).
35. Nakazawa, F. *et al.* Description of Mogibacterium pumilum gen. nov., sp. nov. and Mogibacterium vescum gen. nov., sp. nov., and reclassification of Eubacterium timidum (Holdeman et al. 1980) as Mogibacterium timidum gen. nov., comb. nov. *International Journal of Systematic and Evolutionary Microbiology* **50,** 679–688 (2000).
36. Mason, B. L. *et al.* Reduced anti-inflammatory gut microbiota are associated with depression and anhedonia. *Journal of affective disorders* **266,** 394–401 (2020).
37. Sato, J. *et al.* Probiotic reduces bacterial translocation in type 2 diabetes mellitus: A randomised controlled study. *Scientific Reports* **7** (2017).
38. Amirkhanzadeh Barandouzi, Z., Starkweather, A. R., Henderson, W. A., Gyamfi, A. & Cong, X. S. Altered composition of gut mi-

crobiota in depression: A systematic review. *Frontiers in Psychiatry* **11,** 1–10 (2020).

39. Ueki, A., Goto, K., Ohtaki, Y., Kaku, N. & Ueki, K. Description of anaerotignum aminivorans gen. Nov., sp. nov., a strictly anaerobic, amino-acid-decomposing bacterium isolated from a methanogenic reactor, and reclassification of clostridium propionicum, clostridium neopropionicum and clostridium lactatifermentans as species of the genus anaerotignum. *International Journal of Systematic and Evolutionary Microbiology* **67,** 4146–4153 (2017).
40. Szyszkowicz, J. K., Wong, A., Anisman, H., Merali, Z. & Audet, M. .-. Implications of the gut microbiota in vulnerability to the social avoidance effects of chronic social defeat in male mice. *Brain, behavior, and immunity* **66,** 45–55 (2017).
41. Kelly, J. R. *et al.* Transferring the blues: Depression-associated gut microbiota induces neurobehavioural changes in the rat. *Journal of psychiatric research* **82,** 109–118 (2016).
42. Lee, J. .-. *et al.* Comparative analysis of fecal microbiota composition between rheumatoid arthritis and osteoarthritis patients. *Genes* **10** (2019).
43. Shkoporov, A. N. *et al.* Ruthenibacterium lactatiformans gen. nov., sp. nov., an anaerobic, lactate-producing member of the family Ruminococcaceae isolated from human faeces. *International Journal of Systematic and Evolutionary Microbiology* **66,** 3041–3049 (2016).
44. Jiang, H. *et al.* Altered fecal microbiota composition in patients with major depressive disorder. *Brain, behavior, and immunity* **48,** 186–194 (2015).
45. Naseribafrouei, A. *et al.* Correlation between the human fecal microbiota and depression. *Neurogastroenterology and Motility* **26,** 1155–1162 (2014).

46. Zheng, P. *et al.* Gut microbiome remodeling induces depressive-like behaviors through a pathway mediated by the host's metabolism. *Molecular psychiatry* **21,** 786–796 (2016).
47. Kim, C. *et al.* Probiotic supplementation improves cognitive function and mood with changes in gut microbiota in community- dwelling older adults: A randomized, double-blind, placebo-controlled, multicenter trial. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences* **76,** 32–40 (2021).
48. Bangsgaard Bendtsen, K. M. *et al.* Gut Microbiota Composition Is Correlated to Grid Floor Induced Stress and Behavior in the BALB/c Mouse. *PLoS ONE* **7** (2012).
49. Ahn, S. *et al.* Agathobaculum butyriciproducens gen. nov. sp. nov., a strict anaerobic, butyrate-producing gut bacterium isolated from human faeces and reclassification of Eubacterium desmolans as Agathobaculum desmolans comb. nov. *International Journal of Systematic and Evolutionary Microbiology* **66,** 3656–3661 (2016).
50. Liu, S., Zhao, W., Liu, X. & Cheng, L. Metagenomic analysis of the gut microbiome in atherosclerosis patients identify cross-cohort microbial signatures and potential therapeutic target. *FASEB Journal* **34,** 14166–14181 (2020).
51. Jeffery, I. B. *et al.* An irritable bowel syndrome subtype defined by species-specific alterations in faecal microbiota. *Gut* **61,** 997–1006 (2012).
52. Liu, T. *et al.* Microbial and metabolomic profiles in correlation with depression and anxiety co-morbidities in diarrhoea-predominant IBS patients. *BMC Microbiology* **20** (2020).
53. Zhang, J. *et al.* Beneficial effect of butyrate-producing Lachnospiraceae on stress-induced visceral hypersensitivity in rats. *Journal of Gastroenterology and Hepatology (Australia)* **34,** 1368–1376 (2019).

54. Bojović, K. *et al.* Gut Microbiota Dysbiosis Associated With Altered Production of Short Chain Fatty Acids in Children With Neurodevelopmental Disorders. *Frontiers in Cellular and Infection Microbiology* **10** (2020).
55. Kristensen, N. B. *et al.* Alterations in fecal microbiota composition by probiotic supplementation in healthy adults: A systematic review of randomized controlled trials. *Genome Medicine* **8** (2016).
56. Bowyer, R. C. E. *et al.* Use of dietary indices to control for diet in human gut microbiota studies. *Microbiome* **6,** 77 (2018).

Chapter 6

Salivary microbiota composition may discriminate between patients with eosinophilic oesophagitis (EoE) and non-EoE subjects

The work described in this chapter is taken from: S. Facchin, M. Calgaro, M. Pandolfo, F. Caldart, M. Ghisa, E. Greco, E. Sattin, G. Valle, E. S. Dellon, N. Vitulo, E. V. Savarino; Salivary microbiota composition may discriminate between patients with eosinophilic oesophagitis (EoE) and non-EoE subjects. Aliment Pharmacol Ther. 2022; 56: 450-462. `https://doi.org/10.1111/apt.17091`

Contributions

SF, **MC**, NV and EVS conceptualised the work. SF, MP, MG curated the metadata. SF, **MC**, MP, ESD, NV, EVS drafted the original manuscript. ES and GV processed the biological samples. **MC**, MP, and NV performed the bioinformatics and statistical analyses. FC acquired the fundings and administered the project with MG and EG, under the management of EVS. NV, GV, and ESD supervised the project at various stages with EVS as the principal investigator.

Challenges and future perspectives

In this work, saliva samples and oesophageal biopses were collected prospectively from 49 adult patients, either diagnosed with Eosinophilic oEsophagitis (EoE) or with symptoms of oesophageal disfunction, undergoing upper endoscopy. The first aim was to characterise the salivary, oesophageal, and gastric microbiome in EoE patients through 16S rRNA analysis. Instead, the second objective was to correlate the findings with this specific disease and its activity from saliva samples. Indeed, the high accessibility of saliva samples during routine outpatient visits, combined with high-throughput technologies, could produce huge amounts of data. Statistical models could be trained on these data and used to help clinicians in diagnosis and prognosis, moving towards a more personalised medicine approach.

In the context of this research the main challenge was the low sample size which could reduce the scope of the results by reflecting specific microbial characteristics of this cohort. Despite the low sample size and all the derived limits, I was able to train a classifier based on a sparse Partial Least Squares Discriminant Analysis (sPLS-DA), and validate it on a second group of samples, to discriminate between cases and controls with decent accuracy. sPLS-DA is an extension of the popular Partial Least Squares Discriminant Analysis (PLS-DA) method, specifically designed to handle high-dimensional data with a limited number of samples. One key advantage of sPLS-DA over traditional PLS-DA is its ability to achieve sparsity, *i.e.*, the property of having only a subset of features that are truly informative for discrimination, while the remaining features are effectively disregarded. This is achieved through LASSO penalisation, by incorporating a penalty into the model optimisation process. This penalty encourages a smaller subset of features to have non-zero coefficients, effectively selecting the most informative variables for discrimination while shrinking the rest towards zero. The sparsity level parameter can be determined using cross-validation and is chosen based on a performance metric, such as classification accuracy.

The future perspectives of this project rely on the inclusion of more samples, both by directly sampling new saliva specimens from patients during routine outpatient visits and by integrating available data from public repositories or other similar research. In both cases, but especially in the latter, the batch effect correction holds the key to successful integration.

Article location

`https://onlinelibrary.wiley.com/doi/full/10.1111/apt.17091`

Supplementary material

The supplementary material for this article consists of 3 additional files available at `https://onlinelibrary.wiley.com/doi/full/10.1111/apt.17091#support-information-section`

6.1 Summary

Background

Data on the role of the microbiome in adult patients with Eosinophilic oEsophagitis (EoE) are limited.

Aims

To prospectively collect and characterise the salivary, oesophageal and gastric microbiome in patients with EoE, further correlating the findings with disease activity.

Methods

Adult patients with symptoms of oesophageal dysfunction undergoing upper endoscopy were consecutively enrolled. Patients were classified as EoE patients, in case of more than 15 eosinophils per high-power field, or non-EoE controls, in case of lack of eosinophilic infiltration. Before and during endoscopy, saliva, oesophageal and gastric fundus biopsies were collected. Microbiota assessment was performed by 16S rRNA analysis. A sparse Partial Least Squares Discriminant Analysis (sPLS-DA) was implemented to identify biomarkers.

Results

Saliva samples were collected from 29 EoE patients and 20 non-EoE controls; biopsies from 25 EoE and 5 non-EoE controls. In saliva samples, 23 Amplicon Sequence Variants (ASVs) were positively associated with EoE and 27 ASVs with controls, making it possible to discriminate between EoE and non-EoE patients with a Classification Error (CE) of 24%. In a validation cohort, the accuracy, sensitivity, specificity, positive predictive value and negative predictive value of this model were 78.6%, 80%, 75%, 80% and 60%, respectively. Moreover, the analysis of oesophageal microbiota samples observed a clear microbial pattern able to discriminate between active and inactive EoE (CE=8%).

Conclusions

Our preliminary data suggest that salivary metabarcoding analysis in combination with machine learning approaches could become a valid, cheap, non-invasive test to segregate between EoE and non-EoE patients.

6.2 Introduction

Eosinophilic oEsophagitis (EoE) is an allergen/immune-mediated disease characterised by symptoms of oesophageal dysfunction and eosinophilic infiltration of the oesophageal mucosa in the absence of secondary causes of eosinophilia [1]. The prevalence (0.5-1 case per 1000) and the incidence (5-10 cases per 100,000 per year) markedly increased in the last decade and is now considered to be one of the most important causes of dysphagia in children and young adults. The diagnosis is based on suggestive clinical features (*e.g.* dysphagia and/or bolus impaction), the presence of eosinophilic inflammation ($\geq$15 eosinophils per High-Power Field [eos/HPF] in at least one of multiple oesophageal biopsies) and exclusion of other causes of eosinophilia [2, 3]. EoE affects more males than females (3:1), and the mean age at diagnosis is between 30 and 50 years in adults and 5 and 10 years among children [4].

The pathogenesis is still uncertain. Among genetic factors, thymic stromal lymphopoietin (TSLP), Calpain 14 (CAPN14), chemokine C-C motif Ligand 26 (STAT6) appear to be involved in the development of EoE [5, 6]. Moreover, environmental factors, including aero- and alimentary allergens, and early life conditions (*e.g.*, caesarean section, use of antibiotics, preterm birth) seem to have a predominant role in causing EoE and suggest that alterations in the microbiota may play a role in EoE pathogenesis [7–9]. In this context, the role of oesophageal microbiome has been evaluated in the evolution of this disease. In fact, a change in the composition or in the load of gastrointestinal microbiota has been involved in molecular pathogenic pathways and in promoting diseases [10–12].

To date, little is known about the possible role of the gut microbiome in EoE, with most of the studies focusing on oesophageal and salivary microbiome [13–17]. These preliminary studies showed that active EoE is associated with an increase in *Haemophilus*, *Neisseria*, and *Corynebacterium* in the oesophageal microbiome and, in contrast, inactive EoE patients and healthy controls have a predominance of Gram-positive (especially *Strep-*

tococcus) bacteria [13–15, 18]. Comparing the salivary microbiome to the oesophageal one in paediatric EoE patients, a study demonstrated that both have an abundance of *Streptococcus*, *Neisseria*, and *Prevotella* [14]. Moreover, there are no data on the composition of the gastric microbiome in EoE subjects, whereas in healthy subjects it seems to be composed by Actinobacteria (*Rothia*, *Actinomyces*, and *Micrococcus*), Bacteroidetes (*Prevotella*), Firmicutes (*Streptococcus* and *Bacillus*), and Proteobacteria (*H. pylori*, *Haemophilus*, *Actinobacillus*, and *Neisseria*) [19–22].

Given the limited knowledge about the characteristics of salivary, oesophageal, and gastric microbiome in EoE and its correlation with the progression of the disease, we aimed to prospectively collect and characterise the salivary, gastric, and oesophageal microbiome in active and inactive EoE patients, and to correlate these findings with disease activity.

6.3 Methods

6.3.1 Study design and case definitions

Adult patients with symptoms of oesophageal dysfunction undergoing OesophagoGastroDuodenoscopy (OGD) with biopsies at Gastroenterology Unit, Academic Hospital of Padua (Italy), between October 2018 and November 2020 were consecutively and prospectively enrolled. The diagnosis of EoE was established according to international guidelines in case of symptoms of oesophageal dysfunction, the presence of an eosinophilic inflammation ($\geq$15 eos/HPF in at least one of the multiple oesophageal biopsies), and the exclusion of other causes of eosinophilia [2, 3]. Active EoE and inactive disease were defined per the 2018 consensus guidelines as a peak eosinophil count of $\geq$ or $<$15 eos/HPF in all oesophageal biopsies performed, respectively [23–25]. To compare the gastro-oesophageal microbiome, adult control patients with gastro-oesophageal symptoms but lacking of eosinophilic inflammation were included. Moreover, additional control patients were enrolled to obtain a higher number of saliva samples for in-depth analysis. Some of them underwent endoscopy and biopsies during the same endoscopic sessions for oesophageal symptoms and had a normal upper gastrointestinal endoscopy, while others were EoE patients who underwent follow-up visits to monitor the maintenance of remission and agreed to participate.

The study was approved by the Regional Ethical Committee for Clinical Trials (n=3312/AO/14 and n=4204/AO/17). Written informed consent was obtained from all eligible participants before participation.

6.3.2 Clinical, endoscopic, and histological data

Clinical data including demographics, coexisting allergic conditions (*e.g.*, allergic rhinitis, asthma, food allergies, environmental allergies, pharmacological allergies), current and recent (within 4 weeks) exposure to medication like Proton Pump Inhibitors (PPI) and topical corticosteroids, were recorded at the time of the endoscopy. All OGDs were performed by an EoE-trained investigator (EVS) and any oesophageal mucosal changes such as oedema (0-2), rings (0-3), exudates (0-2), furrows (0-2), and strictures (0-1) were recorded for the evaluation of EREFS scores (range 0-10; higher scores indicate more severe endoscopic findings) [26].

6.3.3 Biopsies sample collection and preprocessing

We obtained from each patient at least six oesophageal biopsies (*i.e.*, from the upper, middle, and lower sites) for histology for EoE diagnosis and monitoring (in the case of follow-up endoscopies). For the microbiota analysis, we obtained one biopsy from the upper, middle, lower oesophagus, and one from the gastric fundus conserved in a lysis/stabilisation solution until analysis. An expert gastrointestinal pathologist analysed the oesophageal biopsies to determine the Eosinophilic oEsophagitis Histologic Scoring System (EoEHSS) score, based on features of: intensity of eosinophilic inflammation, basal zone hyperplasia, dilated intercellular spaces, eosinophilic microabscess, eosinophil surface layering, surface epithelial alterations, dyskeratotic epithelial cells, and lamina propria thickness when present [27]. Duodenal and gastric biopsies were also collected for the histopathologic evaluation of gastritis, *H. pylori* infection, and eosinophilic infiltration, in particular, to exclude cases of concomitant eosinophilic gastritis or enteritis [27].

6.3.4 Saliva sample collection and preprocessing

Saliva samples were collected just before the OGD. Per standard protocol, participants were fasting for at least 6 hours before the upper endoscopy. After providing informed consent, between 1 and 2 ml of saliva were collected in Omnigene-oral kit (DNAgenotek). Among the additional EoE cases who did not undergo endoscopic assessment, saliva was collected before outpatient clinics, but they were asked to respect the same conditions of the patients who underwent the upper endoscopy (*i.e.* fasting for at least 6 hours before collection). The samples were stored at -20°C until further analysis.

6.3.5 Illumina 16S library construction

Next Generation Sequencing (NGS) protocol was performed by BMR genomics (Padua) using standard techniques. Briefly: V3-V4 regions of 16S rRNA gene were amplified using the primers Pro341F: 5'-CCT ACG GGN BGC ASC AG-3' and Pro805R: Rev 5'-GAC TAC NVG GGT ATC TAA TCC-3' [28]. Primers were modified with forward overhang: 5'-TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG [locus-specific sequence]-3' and with reverse overhang: 5'-GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA G [locus-specific sequence]-3' necessary for dual-index library preparation, following Illumina protocol [29]. Samples (saliva and biopsies) were normalised, pooled, and run on Illumina MiSeq with a 2×300 bp approach.

6.3.6 Bioinformatics data analysis

Analysis was performed using R (v4.0.4) (R Core Team, 2019). Primarily, the sequences in FASTQ format were analysed using DADA2 (v1.18), a tool that implements an error correction model and allows the identification of exact sample sequences that differ as little as a single nucleotide [30]. The final output of DADA2 was an Amplicon Sequence Variants (ASVs) table which recorded the number of times each ASV was observed in each sample. DADA2 was run as described in `https://benjjneb.github.io/dada2/bigdata.html` using default parameters. To improve the overall quality of the sequences, the reads were filtered and trimmed using the *filterAndTrim* function implemented in DADA2. Consequently, to remove low-quality bases at the end of reads, the *truncLen* option was set to (290; 250) for the forward and reverse FASTQ files, respectively. Similarly, to remove adapter sequences at the 5' end, the *trimLeft* option was set to (17; 21), for forward and reverse reads, respectively. The *removeBimeraDenovo* function was used to remove chimaeras, via consensus method and then *collapseNoMismatch* function collapsed together all the reads that were identical up to shifts or length variation. Finally, the taxonomic assignment was performed using the naïve Bayesian classifier method implemented in DADA2 (*assignTaxonomy* and *addSpecies* functions) using as reference the Silva 16S database (Version 138), correctly formatted to work with the taxonomic classifier implemented within DADA2 (`https://benjjneb.github.io/dada2/assign.html`) [31]. A phylogenetic tree of the ASVs was obtained using the function *AlignSeq* implemented in DECIPHER (v2.16.1) an R package to create multiple sequence alignments [32]. FastTree (v2.1.11) was used to create the phylogenetic tree [33]. The phyloseq package was used to perform all the downstream analysis in the R environment [34].

6.3.7 Data quality assessment and filtering

Rarefaction curves on raw data were evaluated to assess the species richness among samples as a function of the sequencing depth. Data were pre-processed filtering taxa (ASVs) with a low average relative abundance, setting a threshold of 0.005%; furthermore, taxa present in less than two samples were discarded. Phylum members of Chloroflexi (cumulative relative abundance=0.0001%), Armatimonadota (0.0001%), Acidobacteriota (0.0002%), Abditibacteriota (0.0003%), Verrucomicrobiota (0.0007%), and Desulfobacterota (0.002%) taxa were discarded by this filter. The counts of all the ASVs were collapsed together by genus and by phylum, and the 10 most present genera were plotted to phylum level. Mann-Whitney tests were performed to test relative abundance differences across active disease activity, inactive disease activity, and control samples at phylum level and for each of the 10 most abundant genera.

6.3.8 Biodiversity measurements

Shannon-Wiener index was used to calculate α-diversity, which was plotted by stratifying the samples according to body site and disease activity. Mann-Whitney tests were performed to verify statistical differences in the α-diversity across active disease activity, inactive disease activity, and control samples. To measure β-diversity, data were normalised using the Total Sum Scaling (TSS) normalisation through the *phyloseq_ standardize_ otu_ abundance* function of the vmikk/metagMisc package (`https://github.com/vmikk/metagMisc`). Bray-Curtis distance metrics was used to measure diversity between sample counts and the Principal Coordinates Analysis (PCoA) ordination method was used to ordinate the samples in a reduced dimensional space using the *ordinate* function of the Vegan package (v2.5-7) [35]. To test the multivariate homogeneity of group dispersions, *betadisper* function of the latter package was used. Finally, the PERMutational ANalysis Of VAriance (PERMANOVA) was performed, using the *adonis* and the *adonis_pairwise* functions, in order to investigate disease activity and condition contributions on the β-diversity variability.

6.3.9 Biomarkers identification

A discriminant analysis was computed using sparse Partial Least Squares Discriminant Analysis (sPLS-DA) methods to identify possible biomarkers associated with the condition (EoE vs non-EoE), disease activity (active based on $\geq$15 eos/HPF vs inactive based on $<$15 eos/HPF) in the three oesophageal biopsies, and the EREFS score. In particular, following the default mixOmics (v6.14) pipeline (`https://mixomics.org/case-studies/splsda-srbct/`), a pseudo-count value of 1 was added to the raw counts, which were then normalised with TSS and Centred Log-Ratio (CLR) transformed [36, 37]. This compositional approach is based on the CLR value which is computed through the ratio of an ASV abundance, and the geometric mean of all the other ASV abundances in the sample. A positive (or negative) value of the CLR indicates that the abundance of the considered ASV is CLR-fold bigger (or smaller) than the geometric mean of the abundances of all the ASVs. Consequently, a zero value does not indicate the absence, instead, it indicates that the difference between the ASV's abundance and the geometric mean of the abundances is null.

The sPLS-DA classification performance was measured with a machine learning approach through the function *tune.splsda*. The tuning was performed with a leave-one-out Cross Validation (CV) process, and a prediction distance (maximal distance) was chosen to predict class membership across all CV runs. The ability of the model to correctly classify samples was summarised by the Classification Error (CE) which is computable by subtracting the classification accuracy to 1: *Classification error=1-Accuracy*. Accuracy is computable as $\frac{TP+TN}{TP+FP+TN+FN}$, where *TP*, *TN*, *FP*, and *FN* are the true positives, true negatives, false positives, and false negatives, respectively. For each comparison a summary image was plotted using the *HotLoadings* function (`https://github.com/mcalgaro93/HotLoadings`), displaying the

discriminant ASVs loadings and the related heatmap.
Finally, to establish the adequacy of the model, it was tested in a validation set of 14 saliva samples. Accuracy, specificity, sensitivity, positive predictive value, and negative predictive value were computed.

6.3.10 Statistical analysis

When continuous parameters were compared, a non-parametric test (Mann-Whitney test or Kruskal-Wallis) was used, while the proportions were compared using Fisher's exact test. For the relative abundance analysis, to assess the main microbial differences between EoE and non-EoE patients, Mann-Whitney tests were performed, independently, on the relative abundances of the 10 most abundant genera at phylum level, stratifying the samples by body site. To better characterise and identify potential biomarkers for EoE condition, differences in the microbial compositions between EoE and non-EoE subjects, for each body site we conducted a multivariate analysis based on sPLS-DA data. The sPLS-DA is a variation of the Partial Least Squares Discriminant Analysis (PLS-DA) and enables the selection of the most predictive or discriminative features in the data to classify the samples. The sPLS-DA performs variable selection and classification in a one-step procedure. This compositional approach is based on the CLR values that indicate the abundance of a taxa relative to the average (geometric mean) abundance of all the other taxa in the sample. To this respect, when interpreting the results, it is important to remember that we examined ratios between values, that was the change in abundance of a taxon relative to all others in the data set, rather than abundances. Moreover, sPLS-DA analyses in saliva, oesophagus (all segments considered together), and gastric fundus were conducted to investigate whether specific taxa were associated with active or inactive-EoE. Finally, to investigate the differences between the three oesophageal biopsies of each subject and the association with eos/HPF counts, sPLS-DA was performed.

6.4 Results

6.4.1 Demographics and clinical parameters

Of 49 adults enrolled (mean age 35 years, range 18-76 years), 29 were EoE-patients (16 inactive and 13 active), and 20 were non-EoE controls. Saliva samples were collected from all the subjects, whereas biopsies for microbiome assessment were collected from 25 out of 29 EoE patients and only 5 out of 20 non-EoE controls. The latter five non-EoE controls, they had symptoms of oesophageal dysfunction, lack of eosinophilic inflammation at upper endoscopy, and no previous treatment with proton pump inhibitors. Demographic and clinical characteristics of the whole population are detailed in Table 6.1. The groups were comparable for age (EoE patients' InterQuartile Range [IQR] 25-50 years vs non-EoE patients' IQR 27-48 years, p-value=0.63), while they differed in terms of sex (p-value=0.01). At the time of OGD for microbiome samples, 26 out of 29 (90%) EoE patients were taking PPIs and the proportion was comparable in both inactive (88%) and active-EoE (92%) groups (p-value=1) but not between all the EoE patients and the controls (55%, p-value=0.01).

Features	**EoE patients**			
	Inactive EoE (N=16)	**Active EoE** (N=13)	**Controls** (N=20)	**p-value**
Demographics				
Male, n %	14 88%	10 77%	8 40%	0.01[a]
Age, median (IQR)	37 (25–52)	29 (21–43)	39 (27–47)	0.65[b]
Clinical symptoms, n %				
Dysphagia	3 19%	4 31%	0 0%	0.02[a]
Bolus impaction	2 13%	4 31%	2 10%	0.29
Heartburn/regurgitation	4 25%	4 31%	6 30%	1.00
Chest pain	1 6%	1 8%	1 5%	1.00
Abdominal pain	1 6%	4 31%	7 35%	0.09
Nausea/vomiting	1 6%	0 0%	1 5%	1.00
Allergic comorbidities, n %				
Rhino/conjunctivitis	4 25%	6 46%	1 5%	0.02
Asthma	2 13%	3 23%	1 5%	0.35
Food allergies	1 6%	3 23%	1 5%	0.29
Environmental allergies	3 19%	7 54%	2 10%	0.02
Other atopic manifestations (*e.g.*, atopic dermatitis)	1 6%	2 15%	3 15%	0.75
Therapies, n %				
Proton pump inhibitors	14 88%	12 92%	11 55%	0.03
Topical steroids	7 44%	9 69%	0 0%	0.00
Endoscopy lesions, n %				
Edema	2 13%	2 15%	—	1.00
Rings	3 19%	10 77%	—	0.00
Exudates	7 44%	8 62%	—	0.46
Furrows	4 25%	6 46%	—	0.27
Stricture	2 13%	1 8%	—	1.00
Histology, median (IQR)				
eos/HPF[c]	1 (0–3.25)	35 (20–45)	—	0.00[d]
16S Analysis, n %				
Saliva	16 100%	13 100%	20 75%	
Upper oesophagus	15 94%	10 77%	5 25%	
Middle oesophagus	15 94%	10 77%	5 25%	
Lower oesophagus	15 94%	10 77%	5 25%	
Gastric Fundus	15 94%	10 77%	5 25%	

[a] Fisher's exact test.

[b] Kruskal–Wallis rank sum test.

[c] We consider the highest eosinophilic peak in one of the biopsies.

[d] Mann–Whitney test.

Table 6.1: Demographic and clinical characteristics of the whole population. P-values refer to the comparisons between EoE and non-EoE patients.

6.4.2 Microbial composition of the samples according to body sites

The 16S rRNA metabarcoding analysis of saliva samples was performed for a total of 16 inactive-EoE patients, 13 active-EoE patients, and 20 non-EoE controls. Moreover, the 16S rRNA metabarcoding analysis of gastro-oesophageal mucosal samples was performed for 15 inactive-EoE patients, 10 active-EoE patients, and 5 non-EoE controls. They resulted in 761 ASVs with a median of 62,333 bacterial reads (IQR 46532, 71358) per sample retained after data processing, quality control, and filtering (Additional file 1: Supplementary Fig. S1-S4). The most abundant phyla overall were Firmicutes, Bacteroidota, and Proteobacteria with more than 86% of the total counts, followed by Fusobacteriota, Actinobacteriota, Patescibacteria, Campilobacterota, and some other low abundant phyla (Additional file 1: Supplementary Table S1). At the genus level, the 10 most abundant genera were *Streptococcus*, *Prevotella*, *Haemophilus*, *Veillonella*, and *Neisseria* which contributed to the 60% of the total counts, followed by *Fusobacterium*, *Alloprevotella*, *Actinobacillus*, *Porphyromonas*, and *Gemella* which contained almost the 25% of the counts (Additional file 1: Supplementary Table S2).

α-diversity was different between body sites, displaying a significantly higher Shannon index in saliva and gastric fundus, compared to the three oesophageal segments (Fig. 6.1 a). β-diversity (Fig. 6.1 b) showed that the dispersion of the samples was homogeneous between body sites, while it was significantly different between active/inactive EoE patients (p-value=0.026), active vs non-EoE patients (p-value=0.015) and tended to be significant between inactive and non-EoE patients (p-value=0.093). Considering the homogeneity of variances between body sites, PERMANOVA analysis highlighted that body sites were significantly associated with the β-diversity measurements (p-value=0.001). Specifically, the pairwise comparisons between body sites displayed non-significant differences only between the three oesophageal segments. Of the total 761 ASVs, 550 were present in all the body sites, while 15 of them were present exclusively in saliva and 25 were present exclusively in the oesophagus (in more than one tract: Fig. 6.1 c).

6.4.3 Oesophageal, gastric, and salivary microbiome composition between EoE and non-EoE patients

Microbial composition by site in active, inactive EoE, and non-EoE samples is summarised in Fig. 6.2 where the 10 most abundant genera are reported. No significant differences were found comparing their relative abundances (see Methods and Additional file 2 for details). However, a minor trend was observed for Bacteroidota phylum that resulted to be less abundant (unadjusted p-value=0.03) in gastric fundus samples of EoE patients compared to that of non-EoE (30% vs 36.7%; Additional file 2 a). Similarly, the *Neisseria* genus was found to be more abundant (unadjusted p-value=0.04) among the saliva samples of EoE patients compared to that of non-EoE (11.09% vs 7.39%; Additional file 2 c).

A similar microbial-richness was shown by the α-diversity analysis between EoE and non-EoE samples stratified by body sites (Fig. 6.1 d), although the α-diversity values of non-EoE were slightly higher than those of EoE without reaching statistical significance. Similarly, the first two principal coordinates of β-diversity were unable to show a clear separation between EoE and non-EoE patients (Fig. 6.1 b).

To further investigate differences in the microbial composition, we applied a multivariate statistical analysis based on sPLS-DA to identify possible biomarkers associated with EoE and non-EoE patients. The analysis performed on the saliva samples revealed that a group of 50 ASVs were able to discriminate between EoE and non-EoE patients with a classification error of 24%. In particular, 23 of them were positively associated with EoE samples, while the remaining 27 were positively associated with the non-EoE ones. Among the most discriminant ASVs positively associated with EoE samples, we found *Streptococcus cristatus*, *Prevotella oris*, *Veillonella mas-*

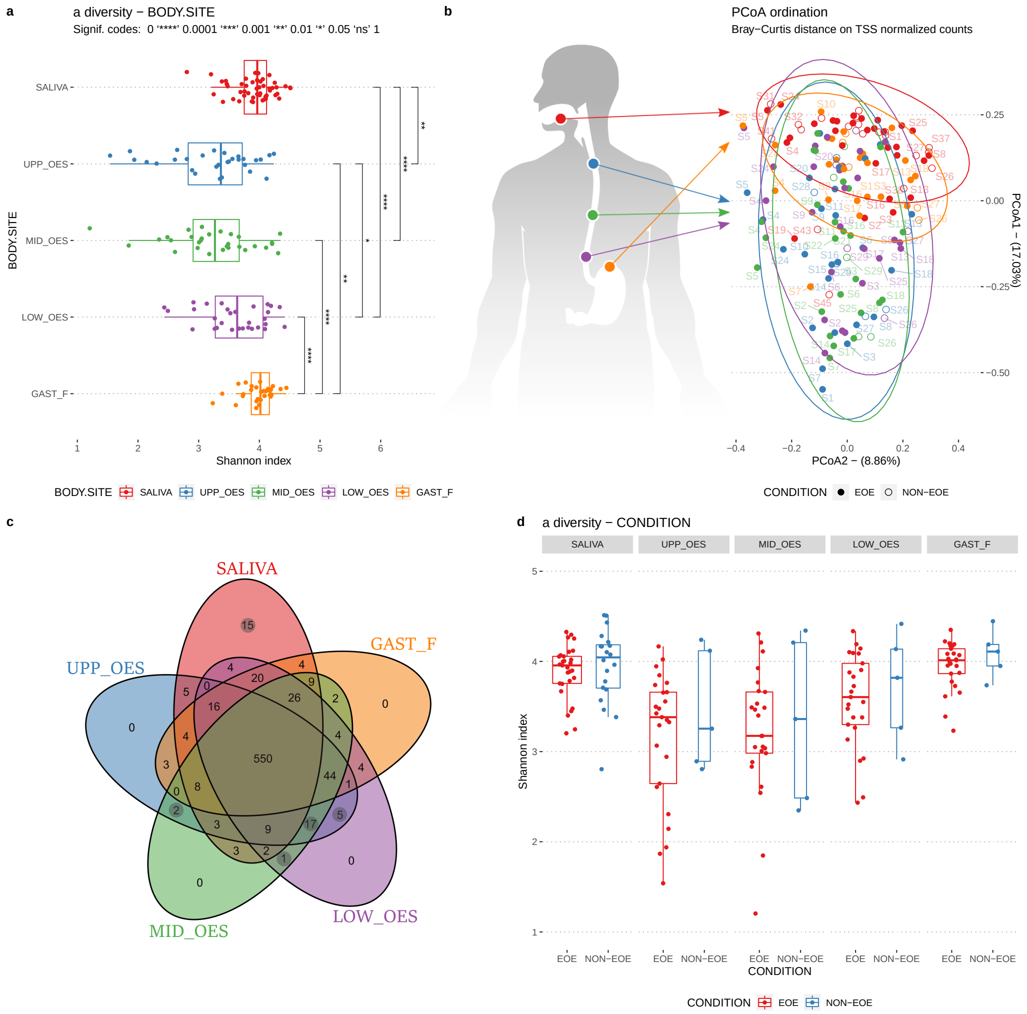


Figure 6.1: **a** Shannon-Wiener α-diversity, over body site. P-values signif. codes are reported for the significant comparisons according to the Mann-Whitney tests. **b** Anatomical body sites and bidimensional representation of β-diversity (PCoA ordination method on Bray-Curtis distance matrix of TSS normalised counts). Colored and circled by body site and shaped by case/control groups condition. **c** Venn diagram for the ASVs in each body site. **d** Shannon-Wiener α-diversity, over condition and faceted by body site. (Mann-Whitney tests between EoE and non-EoE status resulted not statistically significant, p-value>0.05)

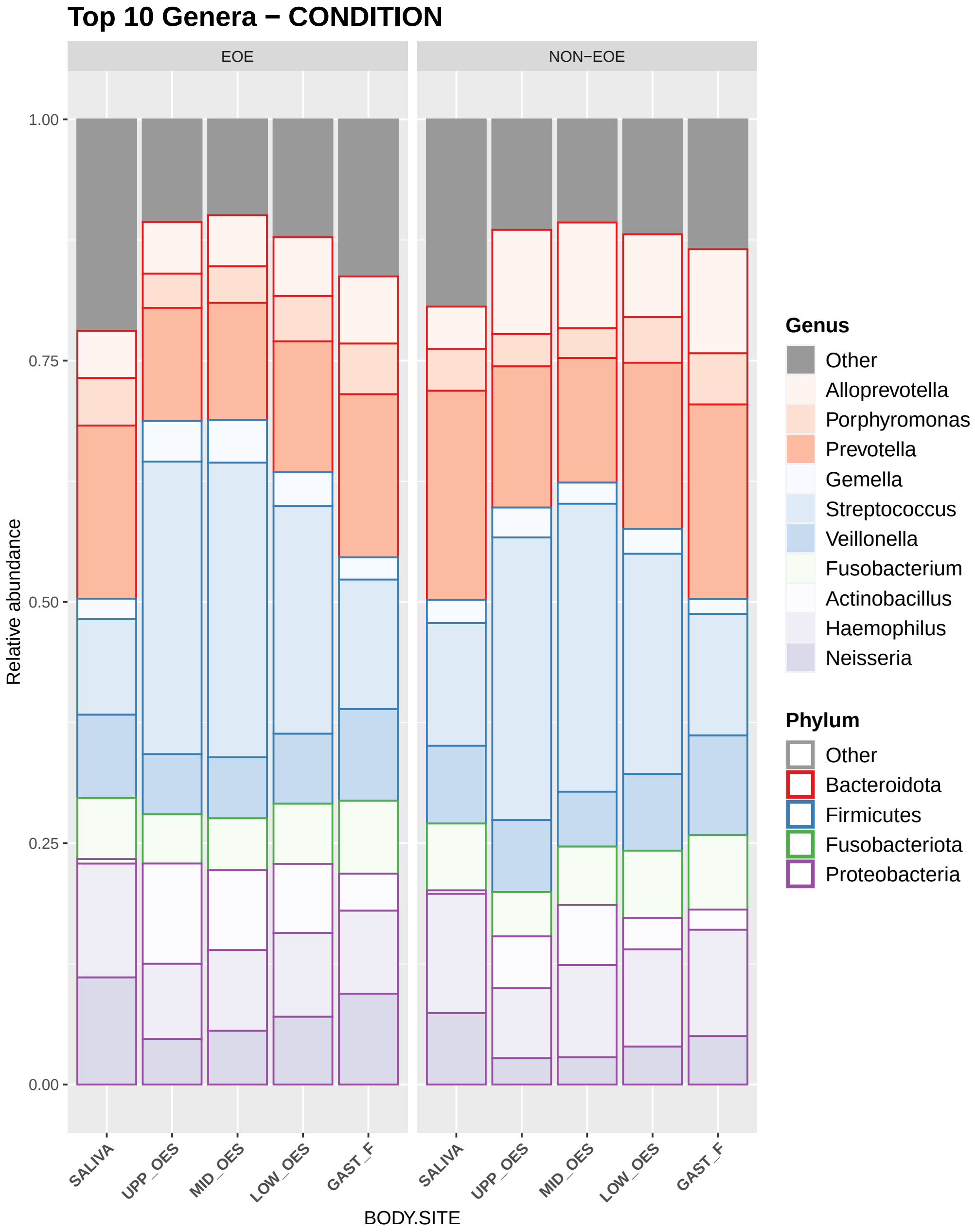


Figure 6.2: The top 10 genera and related phyla are shown; taxa are plotted for their mean relative abundance over body site in case and control groups

siliensis, and *Peptostreptococcus stomatis* spp., together with ASVs of *[Eubacterium] nodatum group*, *Porphyromonas*, *Alloprevotella*, *Selenomonas*, and other *Streptococcus* genera. Conversely, among the ASVs associated with non-EoE patients, we found members belonging to *Prevotella*, *Alloprevotella*, *Porphyromonas*, *Neisseria*, and *Streptococcus* genera, along with *Mogibacterium*, *[Eubacterium] brachy group* genera, and *Haemophilus pittmaniae* spp. (Fig. 6.3).

To establish the adequacy of the model, this was validated on a set of 14 saliva samples (10 from EoE patients and 4 from non-EoE controls) which was comparable to the group where the model was estimated in terms of demographics and clinical characteristics of the patients. As shown in Additional file 1: Supplementary Fig. S8, 8 out of 10 EoE patients and 3 out of 4 non-EoE patients were classified as true positives and true negatives, respectively. The classification accuracy, sensitivity, specificity, positive predictive value, and negative predictive value of the above-specified model were 78.6%, 80%, 75%, 80%, and 60%, respectively.

The sPLS-DA analysis was also performed on the gastric and oesophageal samples and the results are detailed in the Additional file 1: Supplementary Table S3 and Additional file 3.

6.4.4 Oesophageal, gastric, and salivary microbiome composition between active- and inactive-EoE

No significant differences were found in terms of relative abundance analysis. However, a minor trend was observed for *Neisseria* genus that resulted to be less abundant (unadjusted p-value=0.04) in mid oesophagus samples of active-EoE patients compared to that of inactive-EoE (3.02% vs 7.27%; see Additional file 2 b). On the contrary, the *Actinobacillus* genus was found to be slightly more abundant (unadjusted p-value=0.04) among the gastric fundus samples of active-EoE patients compared to that of inactive-EoE (6.48% vs 2.06%; see Additional file 2 d).

The α-diversity analysis performed on active and inactive-EoE, stratified by

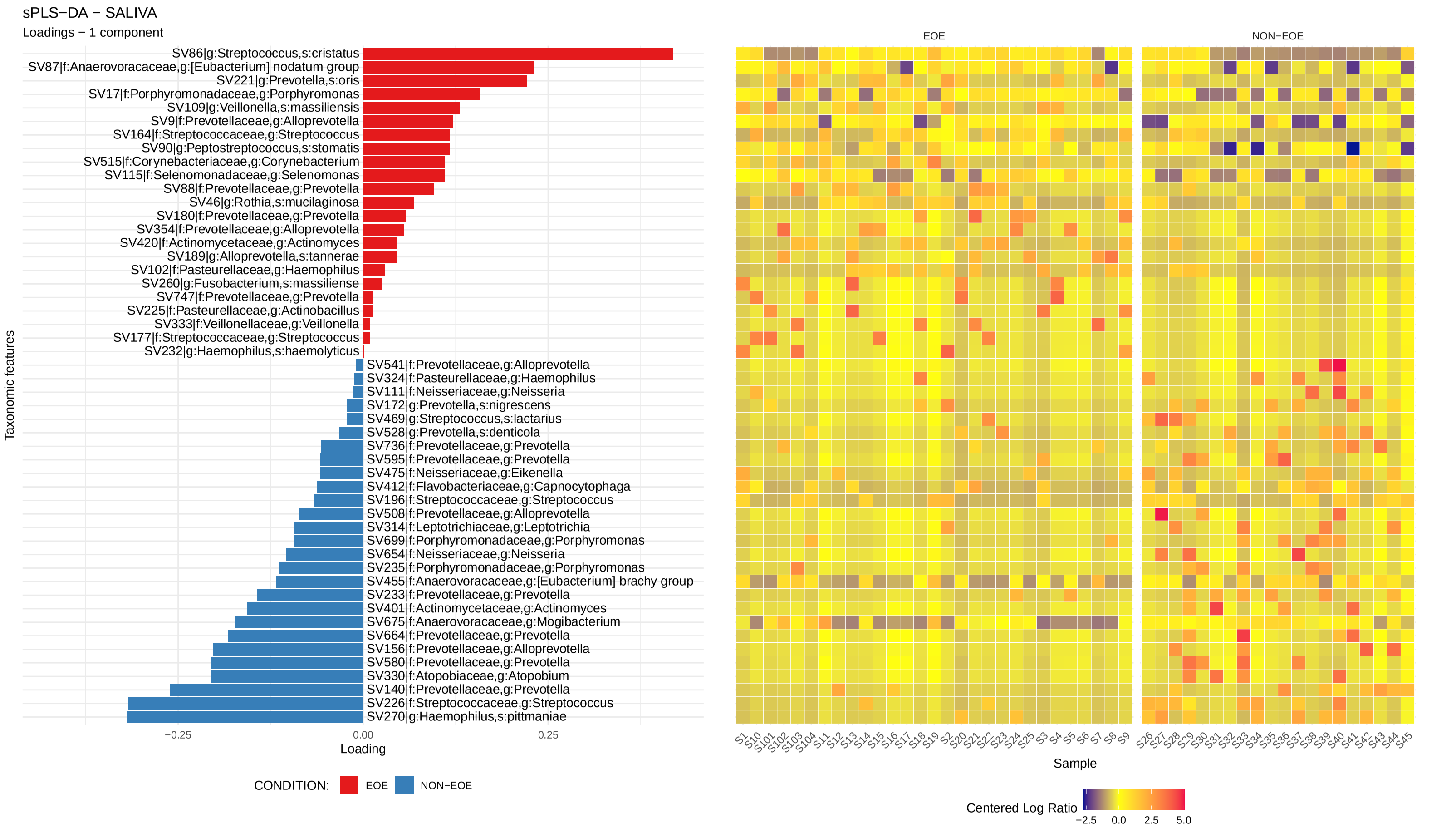


Figure 6.3: sPLS-DA analysis of saliva samples. Loading values (on the left) represent the discriminant taxa of the first component, associated with the condition. Bigger the loading absolute value, stronger the association. Heatmap (on the right) shows the CLR values of the discriminant taxa in all the samples

body sites, showed a similar microbial-richness (Additional file 1: Supplementary Fig. S5). The first two principal coordinates in β-diversity did not show any clear difference between active and inactive-EoE patients (Additional file 1: Supplementary Fig. S6).

Regarding the analysis to identify a potential biomarker for disease activity in EoE, a group of 151 discriminant ASVs was found in saliva samples between active and inactive-EoE patients, with a classification error of 48%. Among the top 50 ASVs (Fig. 6.4), 22 were associated with active-EoE samples, while the remaining 28 were associated with inactive-EoE. We found, as biomarkers of active disease, *Catonella morbi*, *Haemophilus parainfluenzae* species and various ASVs belonging to *Prevotella*, *Alloprevotella*, *Actinobacillus*, *Treponema*, and *Mycoplasma* genera. Instead, other *Prevotella* genera were associated with inactive-EoE samples, together with *gingivalis* and *leadbatteri* species of *Capnocytophaga* genera, *Streptococcus*, and *Actinomyces* genera. Moreover, *Oribacterium asaccharolyticum* and *Streptococcus cristatus* species were characterised by some samples with negative CLR values in active-EoE samples. Further information about the biomarkers found in the other body sites are available in Additional file 3.

6.4.5 Oesophageal microbiome composition according to the different sites and eosinophil counts in EoE patients

With a classification error of 17%, sPLS-DA revealed that 243 ASVs were associated with the dichotomic separation of the histological values (<15 eos/HPF and ≥ 15 eos/HPF). Fig. 6.5 reports the top 50 discriminant ASVs showing a heterogeneous scenario. Members of the *Actinobacillus*, *Bergeyella*, *Porphyromonas*, and *Alloprevotella* genera were associated with biological samples with ≥ 15 eos/HPF, while *Oribacterium asaccharolotycum*, *Streptococcus cristatus*, *Veillonella atypica*, *Prevotella melaninogenica* species, and others were associated with <15 eos/HPF. Interestingly, for

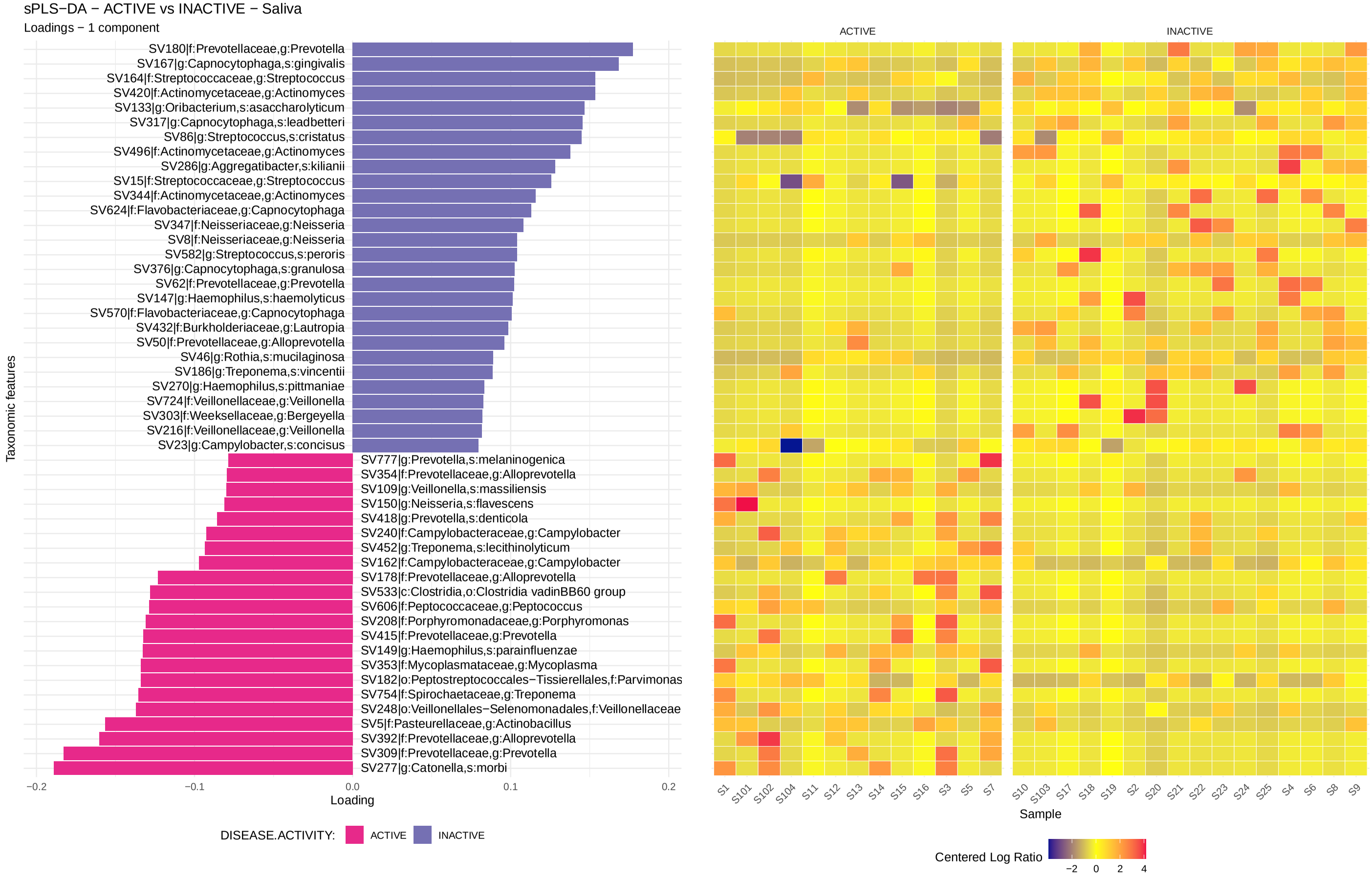


Figure 6.4: sPLS-DA analysis of saliva samples. Loading values (on the left) represent the discriminant taxa of the first component, associated with the clinical status. Bigger the loading absolute value, stronger the association. Heatmap (on the right) shows the CLR values of the discriminant taxa in all the samples.

some patients and for some ASVs, such as for patients S1, S11, and S3, SV208 - *Porphyromonas* CLR values were homogeneous across oesophageal segments even if the histological values were different between biopsies of the same patient. Except for a few rare cases (the top 5 most discriminant taxa), it was difficult to identify a microbial pattern common to multiple samples.

6.5 Discussion

The pathogenesis of EoE is still uncertain. Recent studies hypothesised a role of the oesophageal microbiome in both molecular pathogenesis and as a predisposing risk factor for disease development. However, prospective data, including multiple analyses not limited to the saliva or single-site oesophagus are lacking. Thus, we performed this prospective pilot study to characterise mainly the salivary and partially the oesophageal, and gastric microbiome in EoE and to correlate it with disease activity, with the final aim of discriminating a microbial signature (or a complex of signatures) between patients with EoE compared to patients with oesophageal symptoms due to a non-EoE condition.

Using a sPLS-DA we observed that in saliva samples, 23 ASVs associated with EoE and 27 ASVs associated with non-EoE were able to discriminate between EoE and non-EoE patients with a reasonably low classification error (CE=24%). We also validated the model on an additional small sample of patients, observing a 78.6% accuracy, 80% sensitivity, and 75% specificity. This represents a promising result considering the ease of collecting salivary samples from our patients as compared to the more cumbersome execution of upper endoscopy and suggests the potential utility of saliva microbiome assessment as a non-invasive disease marker to be confirmed in future larger studies. In contrast, the analysis of oesophageal microbiota samples did not identify a specific microbial pattern that distinguished between the study groups, in agreement with a recent study which observed that there were

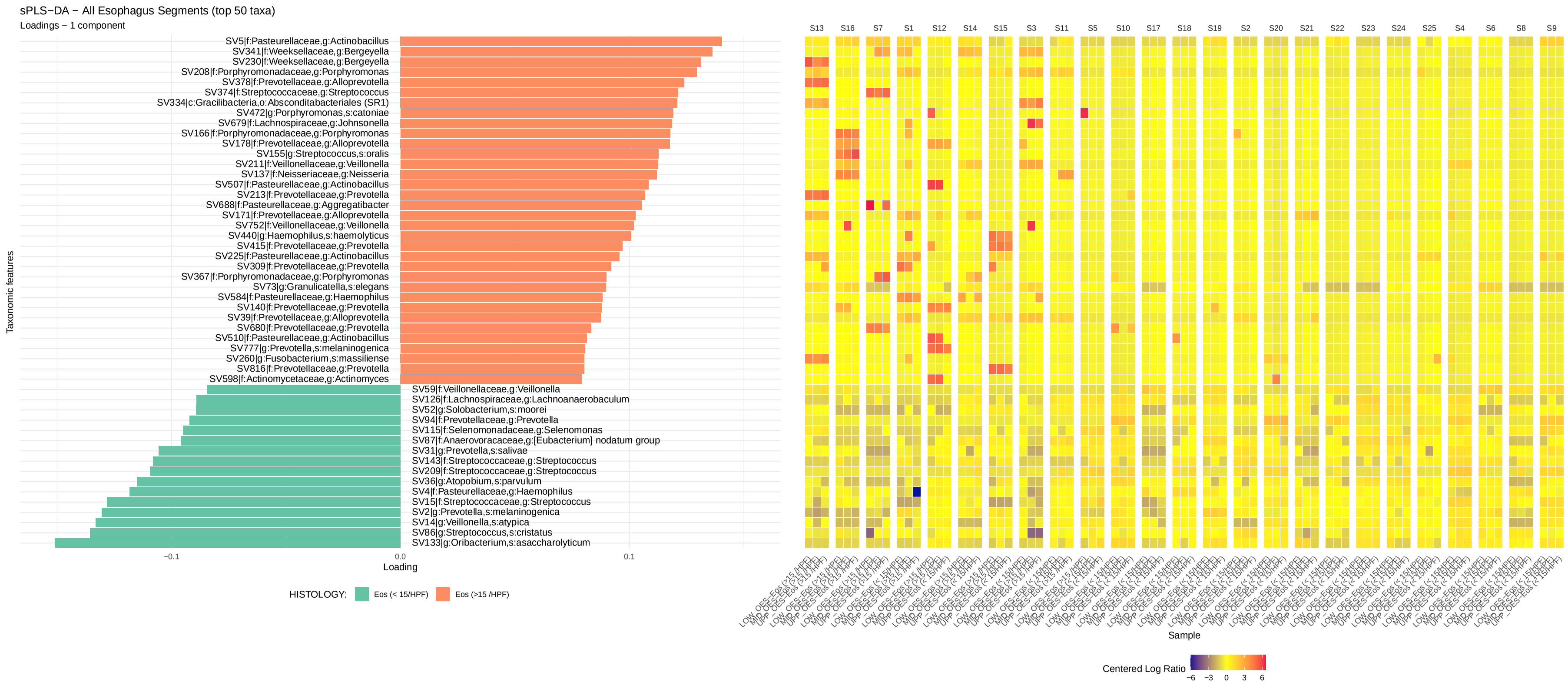


Figure 6.5: sPLS-DA analysis on all oesophagus samples (proximal, mid, and distal). The top 50 loading values (on the left) represent the discriminant taxa of the first component, associated with the histology. Bigger the loading absolute value, stronger the association. Heatmap (on the right) shows the CLR values of the discriminant taxa in all the samples.

no significant differences in the oesophageal microbiome between newly diagnosed EoE cases and non-EoE controls in adults, or within EoE cases based on clinical features [38]. However, it is true that the small number of samples available does not allow us to reach conclusive results on this issue. To the best of our knowledge, this is the first study comparing the salivary microbiome with the oesophageal microbiome examining multiple oesophageal biopsy sites and the gastric microbiome in patients with EoE. From an analysis divided by collection site, we highlighted a substantial difference between salivary and oesophageal microbiota, with greater intradiversity in saliva and gastric fundus than in the oesophagus (Fig. 6.1 a). This microbiological difference may be explained by PPI administration in the majority of our subjects with the consequent increase of intragastric pH and loss of barrier effect of the stomach. Moreover, the same results were observed in both EoE and non-EoE subjects, suggesting that this difference was not influenced by any pathological condition. On the other hand, we cannot exclude that this microbiological difference between saliva and oesophageal microbiome could be due to the presence of atopic pathologies presented by both non-EoE controls and EoE patients. Indeed, it has been reported that both eosinophils and basophils can kill bacteria, the former through a number of antimicrobial products including granule cationic proteins and defensins, and the latter through extracellular traps. These products could modify the local microbiota in atopic diseases where there is a significant infiltration of these granulocytes [39].

In this study, we also tried to compare the composition of the salivary, gastric, and oesophageal microbiome in active and inactive EoE. The analysis of oesophageal microbiota samples observed a clear microbial pattern able to discriminate between active and inactive EoE (CE=8%), while the performances in identifying active and inactive-EoE of salivary and gastric fundus microbiota patterns were less precise (CE=48% and 40%, respectively). Thus, our findings suggest that salivary samples seem less practical to be used for segregating EoE patients according to their disease activity, due to the fact that a large group of 151 discriminant ASVs was found in saliva samples between active and inactive-EoE patients. Considering the top 50, 22 ASVs were associated with active-EoE and 28 were associated with inactive-EoE. Similarly, a recent study has tried to correlate the modification of the salivary microbiome to disease activity, both in terms of endoscopic activity according to the EREFS score and histologic activity according to the EoEHSS score [15]. Hiremath *et al.* found a higher abundance of *Haemophilus* in patients with active EoE and higher EREFS and EoEHSS scores associated with this bacteria [15]. On the other hand, we observed that a microbial signature characterising the salivary microbiota of active patients (*Catonella morbi*) was also abundant in some gastric biopsy samples. *Catonella morbi* is a non-motile, non-spore forming, obligate anaerobic Gram-negative rod that ferments carbohydrates and produces major amounts of acetic acid and smaller amounts of formic and lactic acids. *Catonella morbi* is a normal inhabitant of the oral cavity and has been suggested to be associated with marginal periodontitis. This signature has been also associated with different disorders, including endodontic lesions and coronary heart disease, and oral squamous cell carcinoma [40]. *Catonella morbi* is not the only microbial signature characterising the salivary microbiota in EoE patients to be associated with periodontal diseases. Indeed, in the EoE salivary microbiome, at least two well-characterised signatures (*Prevotella oris* and *Alloprevotella tannerae*), and other genera (*Prevotella*, *Selenomonas*, and *Phorphyromonas*) were associated with oral cavity diseases [41].

At the oesophageal level, we showed a Bacteroidota predominance (*Porphyromonas*, *Alloprevotella*, and *Bergeyella*) in active-EoE patients, which is in contrast with other studies, while patients with the inactive disease showed an undifferentiated presence of Firmicutes, Bacteroidota, and Proteobacteria (Additional file 1) [13, 15, 18, 42]. An additional sPLS-DA analysis was performed to verify whether different microbial signatures were present on the surface of oesophageal biopsies characterised by <15 eos/HPF compared to biopsies characterised by $\geq$15 eos/HPF. We showed members of the *Actinobacillus*, *Bergeyella*, *Porphyromonas*, and *Alloprevotella* genera were positively associated with biological samples with eos/HPF $\geq$15. These are Gram-negative microbial signatures associated mainly with the oral cavity (*Porphyromonas* and *Alloprevotella*) or with the respiratory tract (*Actinobacillus*) and sometimes associated with endocarditis (*Actinobacillus*, *Bergeyella*). On the other hand, bacteria associated with eos/HPF $<$15 histologies as *Oribacterium asaccharolyticum* and *Streptococcus cristatus* Gram-positive or *Veillonella atypica*, *Prevotella melaninogenica* Gram-negative are species differently associated with the healthy oral microbiota. A limitation of the study is related to relatively small sample size and the low number of biopsy samples collected from non-EoE controls. This prevented us from clearly evaluating biopsy microbial signatures as possible discriminating signatures. However, we opted for this approach because our initial preliminary analysis on a few EoE subjects and non-EoE controls did not show relevant differences for the oesophageal and gastric microbiome and therefore, we decided to focus more on salivary evaluation. Our decision was also supported by a recent meta-analysis underlining the importance of oral microbiome assessment to predict in the future various oesophageal diseases via oral samples that can be easily obtained as compared to oesophageal samples [43]. Another limitation is represented by the lack of metabolomic analysis, which could have provided more data on the role of the oesophageal microbiome on EoE. A further limitation includes the lack of control of factors that could influence the microbiome composition as well as the demographic differences observed between our EoE patients and non-EoE controls, including diet, drugs, and gender. However, previous studies showed that diet, gender, and PPI have no or limited effect on the salivary microbiota composition [44], whereas data on topical steroids are lacking. On the other hand, previous studies suggested that drugs like PPIs and topical steroids may have a role in changing gastro-oesophageal microbiome composition [42]. Then again, some points of strengths should be emphasised. This pilot study had a prospective design, which allowed us to collect all the patients' data and control for confounding factors. Moreover, we collected samples of different types and locations from the same subjects, providing a more clear and comprehensive analysis of the microbiome characteristics of the upper GI tract, both in disease and healthy state, whereas previous studies focused on salivary or oesophageal microbiome only. Finally, we correlated the microbiome characteristics with clinical features to increase our understanding of the complex interaction between the upper GI tract microbiome and EoE. Another point of strength should be emphasised: the ease of saliva sampling. Saliva is easy and non-invasive to collect and offers an attractive biofluid for diagnosis and prognostic value. Alterations in salivary microbial ecology are linked to increasing numbers of oral and systemic disease states [45]. Emergent knowledge of the salivary microbiome alongside that of the gut microbiome may offer significant potential for applications in precision or P4 medicine (predictive, preventative, personalised, participatory). The gold standard in the diagnosis of EoE will remain OGDs for many years to come. However, in the near future, our preliminary data suggest that the analysis of the salivary microbiota will help for a better management of patients with oesophageal dysfunction leading to a more rapid and efficient screening of the population to refer for endoscopy in order to confirm the diagnosis of EoE.

In conclusion, our data confirmed that microbial signatures of *Actinobacillus* and *Haemophilus* characterise the salivary microbiota of patients with EoE compared to control patients [17]. Additionally, the discriminant analysis allowed us to characterise a plethora of bacteria in the saliva (as many as 23 positive signatures and 27 negative microbial signatures for EoE patients) whose interaction could be involved in EoE pathogenesis. Moreover, in this pilot study, the validation of our machine learning model, allowed us to reach a sensitivity of 80% and a specificity of 75% for EoE diagnosis. Thus, the metabarcoding analysis of saliva samples in combination with classification methods based on machine learning approaches could become a valid, cheap, non-invasive discriminating test between EoE and non-EoE patients.

References

1. Dellon, E. S. & Hirano, I. Epidemiology and Natural History of Eosinophilic Esophagitis. *Gastroenterology* **154,** 319–332.e3 (2018).
2. Dellon, E. S. *et al.* ACG clinical guideline: Evidenced based approach to the diagnosis and management of esophageal eosinophilia and eosinophilic esophagitis (EoE). *American Journal of Gastroenterology* **108,** 679–692 (2013).
3. Torrijos, E. G. *et al.* Eosinophilic esophagitis: Review and update. *Frontiers in Medicine* **5** (2018).
4. Soon, I. S., Butzner, J. D., Kaplan, G. G. & Debruyn, J. C. C. Incidence and prevalence of eosinophilic esophagitis in children. *Journal of pediatric gastroenterology and nutrition* **57,** 72–80 (2013).
5. O'Shea, K. M. *et al.* Pathophysiology of Eosinophilic Esophagitis. *Gastroenterology* **154,** 333–345 (2018).
6. Lim, E. J., Lu, T. X., Blanchard, C. & Rothenberg, M. E. Epigenetic regulation of the IL-13-induced human eotaxin-3 gene by CREB-binding protein-mediated histone 3 acetylation. *Journal of Biological Chemistry* **286,** 13193–13204 (2011).
7. Jensen, E. T., Kuhl, J. T., Martin, L. J., Rothenberg, M. E. & Dellon, E. S. Prenatal, intrapartum, and postnatal factors are associated with pediatric eosinophilic esophagitis. *Journal of Allergy and Clinical Immunology* **141,** 214–222 (2018).
8. Jensen, E. T. & Bertelsen, R. J. Assessing early life factors for eosinophilic esophagitis: lessons from other allergic diseases. *Curr Treat Options Gastroenterol* **14,** 39–50 (2016).
9. Arias, Á., González-Cervera, J., Tenias, J. M. & Lucendo, A. J. Efficacy of dietary interventions for inducing histologic remission in patients with eosinophilic esophagitis: A systematic review and meta-analysis. *Gastroenterology* **146,** 1639–1648 (2014).
10. Donia, M. S. & Fischbach, M. A. Small molecules from the human microbiota. *Science* **349** (2015).

11. Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L. & Gordon, J. I. Human nutrition, the gut microbiome and the immune system. *Nature* **474,** 327–336 (2011).
12. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489,** 220–230 (2012).
13. May, M. & Abrams, J. A. Emerging insights into the esophageal microbiome. *Current Treatment Options in Gastroenterology* **16,** 72–85 (2018).
14. Benitez, A. J. *et al.* Inflammation-associated microbiota in pediatric eosinophilic esophagitis. *Microbiome* **3** (2015).
15. Hiremath, G. *et al.* The salivary microbiome is altered in children with eosinophilic esophagitis and correlates with disease activity. *Clinical and Translational Gastroenterology* **10** (2019).
16. Dellon, E. S. The Esophageal Microbiome in Eosinophilic Esophagitis. *Gastroenterology* **151,** 364–365 (2016).
17. Harris, J. K. *et al.* Esophageal microbiome in eosinophilic esophagitis. *PLoS ONE* **10** (2015).
18. Amir, I., Konikoff, F. M., Oppenheim, M., Gophna, U. & Half, E. E. Gastric microbiota is altered in oesophagitis and Barrett's oesophagus and further modified by proton pump inhibitors. *Environmental microbiology* **16,** 2905–2914 (2014).
19. Bik, E. M. *et al.* Molecular analysis of the bacterial microbiota in the human stomach. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 732–737 (2006).
20. Hunt, R. H. & Yaghoobi, M. The Esophageal and Gastric Microbiome in Health and Disease. *Gastroenterology clinics of North America* **46,** 121–141 (2017).
21. Ianiro, G., Molina-Infante, J. & Gasbarrini, A. Gastric Microbiota. *Helicobacter* **20,** 68–71 (2015).

22. Andersson, A. F. *et al.* Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS ONE* **3** (2008).
23. Dellon, E. S. *et al.* Updated International Consensus Diagnostic Criteria for Eosinophilic Esophagitis: Proceedings of the AGREE Conference. *Gastroenterology* **155,** 1022–1033.e10 (2018).
24. Dellon, E. S. & Gupta, S. K. A Conceptual Approach to Understanding Treatment Response in Eosinophilic Esophagitis. *Clinical Gastroenterology and Hepatology* **17,** 2149–2160 (2019).
25. Reed, C. C. *et al.* Optimal Histologic Cutpoints for Treatment Response in Patients With Eosinophilic Esophagitis: Analysis of Data From a Prospective Cohort Study. *Clinical Gastroenterology and Hepatology* **16,** 226–233.e2 (2018).
26. Hirano, I. *et al.* Endoscopic assessment of the oesophageal features of eosinophilic oesophagitis: Validation of a novel classification and grading system. *Gut* **62,** 489–495 (2013).
27. Collins, M. H. *et al.* Newly developed and validated eosinophilic esophagitis histology scoring system and evidence that it outperforms peak eosinophil count for disease diagnosis and monitoring. *Diseases of the Esophagus* **30** (2017).
28. Takahashi, S., Tomita, J., Nishioka, K., Hisada, T. & Nishijima, M. Development of a prokaryotic universal primer for simultaneous analysis of Bacteria and Archaea using next-generation sequencing. *PLoS ONE* **9** (2014).
29. Illumina. 16S Metagenomic Sequencing Library Preparation. *16S Metagenomic Sequencing Library Preparation* (2013).
30. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* **13,** 581–583 (2016).
31. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic acids research* **41,** D590–D596 (D1 2013).

32. Wright, E. S. Using DECIPHER v2.0 to analyze big biological sequence data in R. *R Journal* **8,** 352–359 (2016).
33. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5** (2010).
34. McMurdie, P. J. & Holmes, S. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **8** (2013).
35. Oksanen, J. *et al. vegan: Community Ecology Package* `https://CRAN.R-project.org/package=vegan`.
36. Lê Cao, K.-A. *et al.* MixMC: A Multivariate Statistical Framework to Gain Insight into Microbial Communities. *PLoS ONE* **11** (2016).
37. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology* **13** (2017).
38. Johnson, J. *et al.* Lack of association of the esophageal microbiome in adults with eosinophilic esophagitis compared with non-eosinophilic esophagitis controls. *Journal of Gastrointestinal and Liver Diseases* **30,** 17–24 (2021).
39. Muir, A. B., Benitez, A. J., Dods, K., Spergel, J. M. & Fillon, S. A. Microbiome and its impact on gastrointestinal atopy. *Allergy: European Journal of Allergy and Clinical Immunology* **71,** 1256–1263 (2016).
40. Zhao, H. *et al.* Variations in oral microbiota associated with oral cancer. *Scientific Reports* **7** (2017).
41. Liljestrand, J. M. *et al.* Association of endodontic lesions with coronary artery disease. *Journal of dental research* **95,** 1358–1365 (2016).
42. Benitez, A. J. *et al.* Effect of topical swallowed steroids on the bacterial and fungal esophageal microbiota in eosinophilic esophagitis. *Allergy: European Journal of Allergy and Clinical Immunology* **76,** 1549–1552 (2021).

43. Park, C. H. & Lee, S. K. Exploring esophageal microbiomes in esophageal diseases: A systematic review. *Journal of Neurogastroenterology and Motility* **26,** 171–179 (2020).
44. Kaakoush, N. O. The microbiome in oesophageal disease – a year in review. *Microbiota Heal Dis* **2** (2020).
45. Acharya, A. *et al.* Salivary microbiome in non-oral disease: A summary of evidence and commentary. *Archives of Oral Biology* **83,** 169–173 (2017).

Chapter 7

Conclusions

This thesis has tried to retrace my PhD journey from my first attempt to extensively explore differential abundance analysis tools to some real data analyses in microbiome research. Ultimately, the goal has been to better understand and manage the increasing complexity of microbiome data.

On the one hand, the benchmarking research and its application (benchdamic) is an attempt to put the basis for developing new approaches from a solid starting point where the criticisms and limitations of current methodology become clearer. Indeed, the wide availability of DAA methods and the consequent choice of the best DAA tool for the dataset under analysis is still an open question. Keeping benchdamic up to date, by adding DAA tools as they are released, would allow practitioners to choose the right method for their data and developers to assess the relative merits of a new method compared to those already available. As an inspection window, benchdamic could become a valid tool to pursue the scope of an easier access to valuable, but sometimes underestimated, methodological and theoretical evaluations. To this regard, the development of appropriate visualisation tools will increase ability to interpret results, especially for non-experts. Sharing the package with the research community is another priority that I have been taking into account. Firstly, by enhancing the content published in the Bioconductor platform and, secondly, by presenting the package as a short talk and instructor-led live demo at the conferences. Taken together, this

allowed me to highlight weak spots and receive valuable feedbacks.
On the other hand, the research about the probiotics in healthy individuals and the other, about the Eosinophilic Oesophagitis, represent two of the most significant contributions I have made during this three years where I had the chance to collaborate in several research projects. In conclusion, both researches have shown that exploratory data analysis and appropriate statistical tools are crucial for obtaining meaningful information from microbiome data, despite small sample sizes. Several ideas for future analytical approaches have been generated, such as exploring mixed effects regression models in greater depth and improving the visual presentation of their results in longitudinal studies. Moreover, the encountered limitations have given rise to opportunities for future developments, including data integration and the enlargement of datasets using available sources in literature. By continuing to explore new avenues of research and development, we can gain a deeper understanding of the complex world of microbiome data and its implications not only for human health but also for the health of other living organisms and ecosystems. As Walt Disney famously said, "we keep moving forward, opening new doors, and doing new things, because we're curious and curiosity keeps leading us down new paths."