

Faculté des sciences

Applying multi-omic data integration methods to identify metabolites associated with dysbiotic vaginal microbiotas

Author: **Florent DUPRET**
Supervisor: **Laura SYMUL**
Readers: **Laurent GATTO, Céline BUGLI**
Academic year 2023–2024
Master [120] in statistics, biostatistics option

Acknowledgements

First and foremost, I would like to thank my promoter, Professor Laura Symul, for granting me the opportunity to work with her, for the many hours she has invested into this project, for inspiring motivation in me, for her many pieces of advice to improve my work, for adapting to my work style when necessary, and most importantly, I would like to thank her for her availability throughout the year.

Secondly, I would like to thank the readers of this thesis, Professor Laurent Gatto, and Professor Céline Bugli. I also want to acknowledge the help and enthusiasm of Professor Bernadette Govaerts.

Thirdly, I would like to thank all my professors and course assistants at Louvain School of Statistics, Biostatistics and Actuarial Sciences for providing me the mathematical background, critical thinking skills, strategic insights, and statistical expertise necessary to achieve this work.

Furthermore, I am grateful to Christine Vanstalle for proofreading my thesis.

Lastly, I would like to express my deepest gratitude to my parents for giving me the opportunity to pursue my passions, for their guidance, and for being my greatest supporters. Additionally, I want to acknowledge the unwavering support of my brothers, cousin, extended family, and friends.

Abstract

Bioinformatics and biostatistics is key to analyzing biologically sourced data across various fields. Within biostatistics, -omics data analysis focuses on processing and investigating data types such as transcriptomic, genomic, proteomic, and metabolomic. These analyses have significantly advanced microbiology and the study of microorganisms in their environment, enhancing our understanding of microbial activity's potential role in the health of the human body. In the specific context of the vaginal microbiota, research has revealed microbiota dominated by *Lactobacillus* spp. is considered healthy, while its replacement by undesirable non-*Lactobacillus* spp. has detrimental effects on the host. These negative impacts are suspected to be partially mediated by bacteria-produced metabolites. However, the relationships between metabolite abundances and microbiota composition have yet to be described on an "omic" scale considering the entire set of metabolites and bacteria found in the vagina. This work aims to address this gap by performing a joint analysis of microbiota composition and metabolite abundance quantified on the same samples from two cohorts of American women. To achieve this, metabolomic and metagenomic datasets were preprocessed and integrated. For the compositional data, the relative abundance of microbes was calculated, and dimension reduction was performed using topic models, which grouped bacterial species into subcommunities. Metabolite data preprocessing involved handling missing values, filtering, transformation, batch effect correction. Following preprocessing, an integrative analysis was conducted using Structuration de Tableaux à Trois Indices de la Statistique, Simultaneous Component Analysis, and Canonical correlation Analysis to identify associations between metabolites and microbial subcommunities. A comparative study was then performed to validate these associations. The most robust findings indicated that lactate and hippurate are associated with lactobacilli-dominated subcommunities, while putrescine and tyramine are linked to non-lactobacillus dominated subcommunities.

List of abbreviations

ANOVA Analysis of variance

ASV Amplicon sequence variants

BMI Body mass index

BV Bacterial vaginosis

CCA Canonical Correlation Analysis

CST Community state type

DNA Deoxyribonucleic acid

LC-MS Liquid chromatography mass-spectrometry/spectroscopy

LDA Latent Dirichlet Allocation

LLOD Lower limit of detection

LLOQ Lower limit of quantification

MAE MultiAssayExperiment - R object

MAR Missing at random

MCAR Missing completely at random

MNAR Missing not at random

PCA Principal Component Analysis ; "PC" = principal component

PCR Polymerase chain reaction

rRNA Ribosomal Ribonucleic acid

SCA Simultaneous Component Analysis

SE SummarizedExperiment - R object

STATIS Structuration des Tableaux A Trois Indices de la Statistique

UMD University of Maryland

Contents

1	Introduction	12
1.1	The vaginal ecosystem, its microbiota and bacterial vaginosis	13
1.2	Metabolomics and metagenomics	16
1.3	Project objectives	19
2	Corpus	20
2.1	Materials and methods	20
2.1.1	Population	20
2.1.2	Sample acquisition	20
2.1.3	16S rRNA sequencing count data generation	22
2.1.4	Metabolite data generation	24
2.1.5	Data preparation	26
2.1.6	Microbial composition data preprocessing	28
2.1.7	Metabolite data preprocessing	31
2.1.8	Metabolite and Compositional data integration	38
2.1.9	Comparative study	44
2.2	Results	45
2.2.1	Data preparation	45
2.2.2	Microbial composition data preprocessing	46
2.2.3	Metabolite data preprocessing	56
2.2.4	Metabolite and compositional data integration	74
2.2.5	Comparative study	86
2.3	Discussion	89
3	Conclusions	94
4	Appendixes	95
4.1	Appendix 1	95
4.2	Appendix 2	96
4.3	Appendix 3	100
4.4	Appendix 4	102

4.5	Appendix 5		104
4.6	Appendix 6		104
4.7	Appendix 7		106

List of Figures

1.1	Diagram of transition from healthy/normal lactobacilli-dominated microbiota to bacterial vaginosis, with <i>Gardenerella vaginalis</i> . (Huma, 2023)	15
1.2	"The workflow of omics platforms. Various omics approaches applied for the study of cancer immunotherapy were demonstrated with general workflows of next-generation sequencing and mass spectrometry-based omics" (Yoon et al., 2021)	16
1.3	Diagram of DNA transcription into RNA and RNA translation into protein; nature.com	18
2.1	Summary of pre-processing steps for metabolite table M and compositional table C represented as matrix manipulation. Rows are referred using i subscript, columns are referred using j subscript . .	38
2.2	R output structure of MAE object for Stanford (right) and UMD (left) cohort	46
2.3	Contingency table of identified microorganisms according to kingdom and phylum	48
2.4	Library size of sample's metagenomic data (ordered by increasing value) in Stanford (left) and UMD (right)	49
2.5	Compositional barplot of the 10 most abundant bacterial species of Stanford (left) and UMD (right) samples	49
2.6	Ordered proportion of <i>Lactobacillus</i> spp. in Stanford (left) and UMD (right) samples	50
2.7	Number of samples assigned to each sub-community state type . . .	51
2.8	(Left) Mean proportion of some <i>Lactobacillus</i> species of samples assigned to topic I, II, III and V; (Right) Proportion of <i>Lactobacillus</i> spp. in samples	51
2.9	Proportion of <i>Lactobacillus</i> spp. in the <i>redefined</i> topics I, II, III and V	52
2.10	Transport (left) and product (right) alignment results from non- <i>Lactobacillus</i> topics models 1 through 20	53

2.11	Number of path as a function of the number of topics K for non- <i>Lactobacillus</i> models using transport (left) and product (right) alignment	53
2.12	Coherence (left) and refinement (right) scores distribution associated to the number of topics K for non- <i>Lactobacillus</i> topics models	54
2.13	Beta plot for non- <i>Lactobacillus</i> models with 5 and 8 topics	55
2.14	Beta plot for combined <i>Lactobacillus</i> and non- <i>Lactobacillus</i> topics with a total of 10 topics / subcommunities	55
2.15	Ordered missing values per column and row in the raw metabolomic dataset of Stanford (left) and UMD (right)	56
2.16	Scatterplot of proportion of UMD and Stanford samples containing a given metabolite	57
2.17	Scatter plot of the mean log10 peak intensities of raw metabolomic dataset as a function of the proportion of missing values in Stanford (left) and UMD (right)	57
2.18	Scatterplot of mean metabolite abundance per sample versus proportion of missing values per sample in Stanford (red) and UMD (blue). <i>Right</i> : metabolites with missing frequency of 85% or above were filtered out. <i>Left</i> : metabolites with missing frequency of 85% or above and samples with missing frequency of 50% were filtered out	59
2.19	Boxplots of Stanford and UMD metabolite intensity for each sample using no scale (top), log10 scale (middle) and hyperbolic arc-sine scale (bottom)	60
2.20	Mean-variance plots of metabolites for Stanford (left - red) and UMD (right - blue) results	61
2.21	Mean-variance plots of metabolites for Stanford (left - red) and UMD (right - blue) results	61
2.22	Barplot of percentage of variance explained by the first 10 principal components of Stanford (left) and UMD (right) PCA on filtered, imputed, transformed dataset	63
2.23	PCA scores plot of Stanford (left) and UMD (right) colored by batch; data input was filtered, imputed, and transformed	64
2.24	PCA scores plot of Stanford (left) and UMD (right) colored according to 10 random subjects; data input was filtered, imputed, and transformed	64
2.25	PCA loadings plot of Stanford (left) and UMD (right); data input was filtered, imputed, and transformed	65
2.26	Metabolite intensity density function of batch 1 and 2 (density plot was drawn using a bandwidth of approximately 0.17)	65
2.27	Gestational age and subject age density functions of batch 1 and 2 .	66

2.28	Metabolite boxplot distribution before batch correction (above) and after correction (below)	69
2.29	Scatterplot of mean metabolite intensity per sample as a function of corresponding principal component' score in Stanford (left) and UMD (right) cohort	70
2.30	PCA loading's plot after size effect correction in Stanford (left) and UMD (right) using dimension 1 and 2	71
2.31	Lactate intensity as a function of <i>Lactobacillus</i> spp. proportion per sample at the different steps of the preprocessing: filtered data (top-left), transformed (top-right), imputed (bottom-left), batch effect removed (bottom-right)	72
2.32	Lactate intensity as a function of <i>Lactobacillus</i> spp. proportion per sample after correcting size effect	73
2.33	STATIS' interstructure plot of the first and second dimensions for Stanford (left) and UMD (right)	75
2.34	STATIS' variance decomposition barplot of the first 30 dimensions out of 141 for Stanford (left) and out of 174 for UMD (right)	76
2.35	STATIS' row coordinates on the compromise' first 4 dimensions for Stanford (left) and UMD (right), colored by the proportion of <i>Lactobacillus</i> spp.	77
2.36	STATIS' column coordinates (loadings) on the compromise' first 4 dimensions for Stanford (left) and UMD (right)	78
2.37	SCA' variance decomposition barplot of the first 10 dimensions for Stanford (left) and UMD (right)	79
2.38	SCA' scores plot using the first 4 principal components for Stanford (left) and UMD (right); 10 random subjects have been identified using different colors	80
2.39	SCA' loadings plot using the first 4 principal components for Stanford (left) and UMD (right)	81
2.40	Barplot of canonical correlations rounded to the nearest hundredth for Stanford (left) and UMD (right)	83
2.41	CCA's variable contribution to their respective block-specific variates 1, 2, 3, and 4 for Stanford (left) and UMD (right)	84
2.42	Colormap of associations identified between metabolites and microbial sub-communities per approach (CCA, SCA, STATIS); associations are colored according to whether they were identified in Stanford , UMD or both	88
4.1	Ordered proportion of <i>Lactobacillus iners</i> in Stanford (left) and UMD (right) samples	95

4.2	Ordered proportion of <i>Gardnerella vaginalis</i> in Stanford (left) and UMD (right) samples	96
4.3	Scatterplot of the log10 mean metabolite peak intensities per sample as a function of the women's age (in years) for Stanford (left) and UMD (right)	96
4.4	Scatterplot of the log10 mean metabolite peak intensities per sample as a function of Stanford's women's BMI (left) or UMD's women menstrual cycle length in days (right)	97
4.5	Scatterplot of the log10 mean metabolite peak intensities per sample as a function of Stanford's women's gestational age in days (left) or UMD's women's cycle day (right)	98
4.6	Scatterplot of the log10 mean metabolite peak intensities per sample as a function of UMD's women's vaginal pH	99
4.7	Scatterplot of the log10 mean metabolite peak intensities per sample as a function of all women's reproductive status	99
4.8	Scatterplot of the log10 mean metabolite peak intensities per sample as a function of UMD's women's cycle number	100
4.9	Eigenvalue table of Stanford (left) and UMD (right) principal component analysis on filtered, imputed, transformed metabolomic dataset	100
4.10	Scores plot of Stanford (left) colored by batch; PCA on filtered, imputed, transformed metabolomic dataset	101
4.11	Scores plot of Stanford (left) and UMD (right) colored according to 10 random subjects; PCA on filtered, imputed, transformed metabolomic dataset	101
4.12	Loadings plot of Stanford (left) and UMD (right); PCA on filtered, imputed, transformed metabolomic dataset	102
4.13	Eigenvalue and variance decomposition tables of Stanford (left) and UMD (right) PCA results for size effect corrected metabolomic datasets	102
4.14	Loading's plot (dimension 3 and 4) of Stanford (left) and UMD (right) PCA results for size effect corrected metabolomic datasets .	103
4.15	Scores's plot of Stanford (left) and UMD (right) PCA results for size effect corrected metabolomic datasets	103
4.16	SCA' scores plot using the first 4 principal components for Stanford (left) and UMD (right)	106

List of Tables

2.1	Summary of data structure: missing values, inconsistencies, and variability sources in the 16s rRNA count data	24
2.2	Summary of data structure: missing values, inconsistencies, and sources variability in the metabolite peak intensity data	26
2.3	Spearman's correlation coefficient estimate and pvalue between mean metabolite intensity and proportion of missing values in each cohort	59
2.4	Contingency table of samples in four different self-proclaimed races (Asian, Black, White, and Other) according to batch. The numbers in each cell correspond to the following format "number of samples (proportion of samples)"	66
2.5	Regulation tables of the linear models adjusted before batch correction (above) and after correction (below)	67
2.6	Spearman's correlation coefficient and associated p-value between the mean intensity of metabolites and the different sample attributes	68
2.7	UMD cohort only; one-way ANOVA test between mean metabolite intensity and categorical variables <code>Reprod_status</code> and <code>cycle_nb</code> . .	70
2.8	Spearman's correlation coefficient between mean metabolite log ₁₀ (intensity) and principal component' score	71
2.9	Correlation tests (estimates and p-values) between proportion of <i>Lactobacillus</i> and lactate intensity per sample	71
2.10	RV coefficient estimates between metabolite and subcommunity data blocks and associated pvalue using Monte Carlo test for both UMD and Stanford	74
2.11	STATIS block weights for Stanford and UMD	74
2.12	Stanford's 8 most contributing metabolites to CCA's first variate .	85
2.13	UMD's 8 most contributing metabolites to CCA's first variate . . .	86

Chapter 1

Introduction

Bioinformatics and biostatistics are key to analyzing clinically and biologically sourced data from many different fields. A branch of biostatistics, -omics data analyses, was designed to process, and investigate -omics data such as transcriptomic, genomic, proteomic, metabolomic, etc. Of particular interest for this project, such analyses have helped progress in the area of microbiology and the study of microorganisms in their environment. Moreover, it allows us to understand the potential role of microbial activity in the health of the human body. Here, we focused on the analysis of the human vaginal microbiota. In this context, bacterial vaginosis, a syndrome resulting from unbalanced levels of bacteria in the vagina, is associated with higher risk of preterm birth and acquisition of sexually transmitted infections (Callahan et al., 2017b) (Gosmann et al., 2017). Bacterial vaginosis (BV) is most common in sexually active populations, with a prevalence of 23 to 29%. While its origin and causes are still debated, several bacteria, such as *Gardnerella spp.*, *Prevotella spp.*, *Mobilincus spp.*, *Megaspahera spp.*, *Sneathia spp.* are consistently associated with it. The popular theory regarding BV is that the replacing of beneficial lactobacilli, which help maintain a healthy vaginal environment, by undesirable bacterial strains has detrimental effects on the host (WHO, 2023) (Mayo, 2023) (Borgogna et al., 2021) (Gosmann et al., 2017). These negative effects are suspected to be mediated, in part, by bacteria-produced metabolites. However, the relationships between metabolite abundance and microbiota composition have not yet been described on a "omic" scale (*i.e.*, considering the set of all metabolites and bacteria found in the vagina). This work aims at addressing this gap by performing a joint analysis of microbiota composition and metabolite abundance quantified on the same samples.

This involves processing, describing, then integrating two omics dataset: metabolomics data, quantifying vaginal metabolites which were obtained via Metabolon's quantitative semi-targeted mass-spectrometry, and metagenomics data, quantifying microbiota composition, obtained from the sequencing of the V4 region of the 16S

rRNA gene. Sequencing results were pre-processed such that we acquired amplicon sequence variants (ASVs) counts for each sample. These two datasets were quantified on the same samples, collected in the form of vaginal swabs, from pregnant women recruited at Stanford University (Stanford, USA) and non-pregnant women recruited at University of Alabama (Birmingham, USA).

In the rest of this introduction, we will first provide an overview of vaginal health and disease, focusing in particular on Bacterial Vaginosis, a syndrome characterized by a dysbiotic vaginal microbiota. Then, we will explain what "omics" data are in general, and focus on metabolomics and metagenomics, the two types of "omics" used in this thesis. Finally, we will clarify the statistical objectives of this project and hint at the means to achieve them, which will be described in details in the Methods section.

1.1 The vaginal ecosystem, its microbiota and bacterial vaginosis

To understand the *dysbiosis* linked to bacterial vaginosis, one needs to understand what physicians define as a "healthy balanced vaginal environment". It is well known that microbes, whether they are benign or unfriendly, colonize and live on and within living beings such as humans. In this case, the human is called a *host* and provides a niche or a habitat in which microbial life can reside. There are many examples of microbial populations living inside or onto human hosts and this community of life is called *microbiota*: gastrointestinal track, skin, oral mucosa, lung, vagina, and many more (contributors, 2024b). These organs or tissues become the environment in which the organisms live. Admittedly, the relationship between a host and its microbiota can be different from one environment to another, and can vary over time. There are three major categories of bacteria-host interactions : symbiotic relationship (where each partner benefits from the other), commensal relationship (partners share the same physical space but show no benefit or detriment to living together) and pathogenic relationship (one partner benefits to the detriment of the other) (Steinert et al., 2000).

A common interpretation of a "healthy vaginal flora" is one where the bacterial microbiota is "balanced" and maintains a symbiotic relationship with the human host in the sense that: the host provides a niche (the vaginal wall) and space to colonize and feed on, while the bacteria create a biologically and chemically stable environment. In this case, stability refers to conditions in which neither the bacterial community nor the vaginal wall deteriorate. This stable ecosystem would be able to grow beneficial bacteria and prevent pathogenic ones from colonizing the space. The mechanisms for host-bacteria interactions are not fully understood

as they are complex and plentiful (Per-Anders, 1991).

In the human vagina, research has provided evidence that bacteria species of the *Lactobacillus* genus play a major role in keeping the vaginal ecosystem in check and providing stability. Indeed, *Lactobacillus* spp. are the most numerous and the most prominent bacterial lifeforms in healthy microbiotas. In short, the way they benefit their host is by acidifying their environment through the production of lactic acid (also called *lactate*). By lowering the pH of their surroundings, these bacteria create a harsh acidic habitat where only specialized individuals can thrive (like other lactobacilli). The acidification and colonization of the vaginal wall by *Lactobacillus* spp. and other adapted bacteria leads to a balanced ecosystem that protects the host from infections, diseases and other pathogens, while the host supplies its inhabitants with nutrients (Kovachev, 2018). Part of our understanding of this microbiome comes from studying the bacterial composition and quantification of vaginal environmental compounds such as metabolites.

As of yet, we have strong evidence to believe bacterial vaginosis is the product of a pathogenic relationship where bacteria benefit from the host: the host provides a physical space for bacteria to live in but the microbes do not offer ecosystemic stability or protection against pathogens. However, how bacterial vaginosis initiates and how bacteria interact with host is still unclear.

From a more medical and epidemiological point of view, bacterial vaginosis "is the most common vaginal infection among women of reproductive age and is associated with important adverse health outcomes" (Peebles et al., 2019). With a prevalence of 23% to 29%, this disease is believed to be more frequent in women with unprotected sexual practices and an increased number of sexual partners. Diagnostics are sometimes limited. One can look for potential symptoms such as "unusual discharge from the vagina, which may be thin and light grey or white", "a strong fishy or musty odour from the vagina", "itching or irritation around the vagina", and "burning during urination". Another diagnostic test is the Gram stain test, "which uses a light microscope to examine vaginal smears. If bacterial vaginosis is present, the test shows an abundance of gram-positive and gram-negative cocci, along with a decrease in gram-positive lactobacilli (Nugent criteria)". While bacterial vaginosis is common, it may lead to problems in pregnancy, higher risk of sexually transmitted infections, including HIV, and pelvic inflammatory disease. (WHO, 2023) (Verstraelen and Verhelst, 2009)

The previously mentioned diagnostics, along with other medical practices and research, traditionally use binary categories (healthy vs vaginosis). However, actual data shows that vaginal health lies on a multidimensional spectrum (Symul et al., 2023). Considering the molecular microbiota composition and relying on mixed-membership dimension reduction methods is more appropriate than categorization into distinct categories when attempting to describe the evolution between health

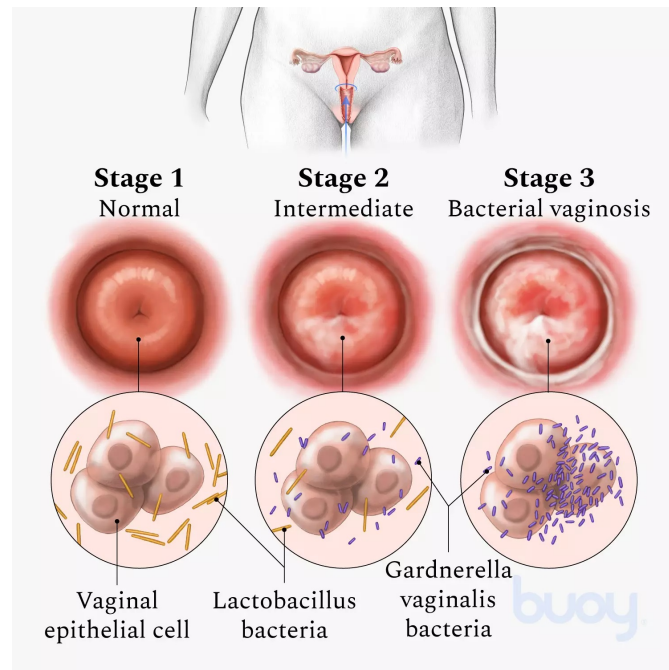


Figure 1.1: Diagram of transition from healthy/normal lactobacilli-dominated microbiota to bacterial vaginosis, with *Gardenerella vaginalis*. (Huma, 2023)

and disease.

Instead of considering distinctive classes, thinking of BV as a transition from one state to another is more appropriate. Indeed, Bacterial Vaginosis occurs when switching from a healthy *Lactobacillus*-dominated vaginal microbiota, to an unstable non-*Lactobacillus*-dominated community, like it is shown in figure 1.1.

In that case, one understands host-bacteria interactions on the cellular and molecular level is crucial to understanding the vaginosis syndrome. It has already been shown that bacterial levels and dominance do alter over time and are correlated with vaginal status. For example, *Lactobacillus* spp.'s dominance cycles and decreases during menses. Additionally, a paucity of lactobacilli is associated with bacterial vaginosis (BV) (Symul et al., 2023). Moreover, metabolite concentration is also correlated with menstrual cycles and bacterial levels (Borgogna et al., 2021). Thus, integrating the metabolomic and microbiota compositional data is necessary to understand the factors involved in BV.

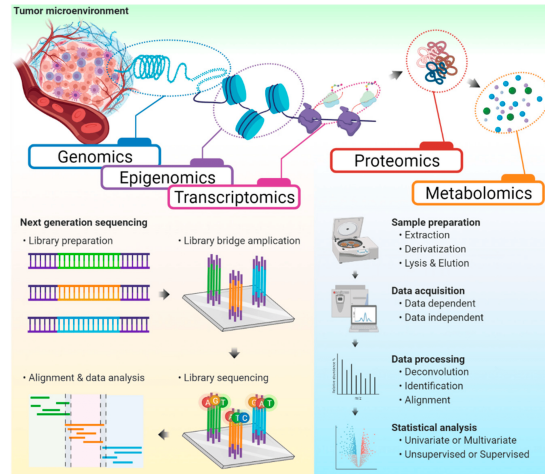


Figure 1.2: "The workflow of omics platforms. Various omics approaches applied for the study of cancer immunotherapy were demonstrated with general workflows of next-generation sequencing and mass spectrometry-based omics" (Yoon et al., 2021)

1.2 Metabolomics and metagenomics

Metabolomics, the study of metabolites, and metagenomics, the study of microbial composition in a given environment, are two examples of Omics data. -Omics data analyses cover a wide range of disciplines: metabolomics, transcriptomics, proteomics, lipidomics, toxicogenomics, metagenomics, etc, for which a general summary can be found in figure 1.2. What they have in common are the last five letters "-omics". The suffix "ome" refers to a totality of some sort (contributors, 2024c). For example, transcriptomics is the study of the totality of the transcriptome. In omics analyses, one will usually obtain data with a small sample size (few subjects) but many variables, where each variable corresponds to a molecule from its respective class (transcriptome, proteome, metabolome, etc.). Such data is usually complex and important in size. It calls for different measuring techniques and preparations. These few steps can require statistical expertise to yield clean structured datafiles, worthy of efficient statistical analyses.(Symul and Hoogstoel, 2024)

In metabolomics, one focuses on small or intermediate biochemicals involved in metabolic pathways. Inside and outside living cells, many pathways build up and break down organic compounds, metabolites, that are used by the cell. These molecules have different roles: precursors of other molecules, regulators of pathways and/or gene expression, acidic or basic properties, signaling molecules, etc. Lactate is an example of metabolite that can be found where lactic acid production takes

place. Through these different roles, they can help accommodate a bacteria's environment, communicate with the host's tissues, or be produced by the host and serve as nutrients for microorganisms. Therefore, these molecules can be of great interest when studying bacterial-host behaviors.

The way one analyses them is by collecting a biological sample, preparing it, assessing it using appropriate techniques and technology, and examining the output. Among several techniques, liquid-chromatography mass-spectrometry (LC-MS) is a popular way to detect and measure relative metabolite concentration. In LC-MS-based measurements, one gathers cell samples (either *in vivo* tissue samples or *in vitro* cell cultures) which are lysed. The metabolites are then isolated from the rest of the sample content, and separated using liquid chromatography to improve resolution. The samples then go through an "ionization" step: this important step will uniformly charge the metabolite molecules. The collected ionized metabolites are then placed in a mass-spectrometer. This machine separates the entities according to their mass-to-charge ratio – for example, by submitting them to a magnetic field – and records their signal intensity as they pass through a detector. As a result, for each passing chemical, one obtains a signal intensity and a mass-to-charge ratio. The combination of these two elements is called the metabolomic profile of a biochemical. The metabolomic profile is then compared to a database to identify the name of the metabolite. In this identification process, also referred to as *peak alignment*, each intensity peak is matched to a known metabolite from a database and a matching score is computed. The match with the highest score will usually be the chosen candidate. Based on the peak alignment approach, the database with which one works, the search keys, and the quality of the data, the scores and the overall matching process will differ. However, one can either perform "targeted" MS, where only a selection of known metabolites are being measured and the matching is more efficient, or "untargeted" MS, where metabolites are unknown at first and matched to a database blindly. (Zhou et al., 2012) (Dettmer et al., 2007) (Metabolon, d) (Metabolon, b)

The other type of data we will be looking at is *metagenomics*. Metagenomics is the study of genetic material recovered directly from environmental samples. It allows scientists to analyze the collective genome of all microorganisms present in a sample. This approach is particularly useful for studying communities of bacteria, viruses, fungi, and other microorganisms in their natural habitats, which are often difficult or impossible to culture in the lab. Several methods exist to quantify the set of genomes present: whole genome sequencing (also called shotgun), and targeted sequencing. The former is rather expensive and time-demanding, while the latter is easier and still very accurate. Targeted sequencing relies on the sequencing of a "universal" gene found throughout all microbes. In this project, the metagenomics data was generated through 16S rRNA targeted gene sequencing.

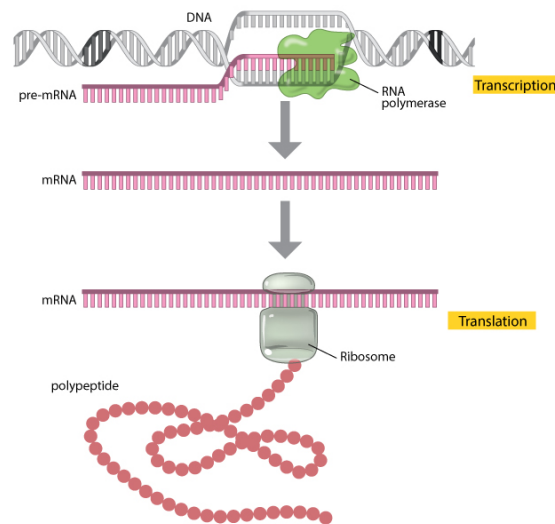


Figure 1.3: Diagram of DNA transcription into RNA and RNA translation into protein; nature.com

To understand why the 16S is a good candidate for this "universal gene", one needs to have a general intuition of gene expression. When a gene is *expressed*, the coding deoxyribonucleic acid (DNA) strand in a living cell is *transcribed* into a ribonucleic acid (RNA), see figure 1.3. RNA strands will usually be used to build proteins or to regulate other gene transcription in the nucleus. However, a special type of RNA, called *ribosomal RNA* (rRNA), will help assemble the ribosome: a macro-molecular machine performing biological protein synthesis in the cytoplasm of a cell. Every living being possesses ribosomes in their cells, including bacteria. In prokaryotes, such as bacteria, there are three types of rRNA: 23S, 5S and 16S. The 16S rRNA gene sequence can be divided further into "hyper-variable regions" denoted V1 through V9. (contributors, 2024a)

Now, biologists discovered mutations in the DNA sequence occur somewhat regularly in a species' life. Therefore, mutations in the coding sequence of rRNA occur. Such mutations may lead to different rRNA and by extension to different ribosomes. Mutations are also a key factor in species differentiation in bacteria and other lifeforms. Thus, generally speaking, one expects different bacterial species to have different rRNA because they experienced different mutating phases. The 16S gene is common in every bacteria, indicating we can use one single gene to compare bacteria, but has *enough* mutations in its hyper-variables regions to be able to distinguish between species. These two key-elements favored the development of bacterial phylogeny and composition of complex bacteria-dominated environments.

Essentially, quantification of the bacterial species in the samples was performed

through the following steps. First, the bacterial cells in the samples were lysed. Secondly, the cell content was extracted and placed in a thermal cycler for Polymerase Chain Reaction (PCR). In PCR, DNA or RNA sequences are *amplified* (*i.e.*, replicated many times through a series of cycles). The amplified sequences are then sent through a detector which will literally count the number of sequences going through for each type of sequence. And since the sequence of rRNA can be mapped out back to its species, one can infer the relative amount of bacteria species in a given sample based on RNA sequencing and PCR.

1.3 Project objectives

This work presents an applied multi-omic data integration where we intend to better understand and identify metabolites associated with bacterial vaginosis; and more generally with microbiota composition. As the samples have already been collected and the data generated, the main focus of this work encompasses the preparation and preprocessing of the metabolomics and metagenomics datafiles, followed by an association analysis of both types of data (*i.e.*, data integration). While metagenomics count data will inform us on the microbial composition of our given samples, metabolic intensity data will give us a broad view of the biochemical profile of the samples. Inference on how metabolites relate to bacteria population development will be drawn. Furthermore, a comparative study, aiming to find the resemblance and dissimilarity between the findings for the Stanford and UMD cohorts, is in order.

The statistical and strategical methods used for this thesis can be found in section 2.1 and include: data preparation, data preprocessing for both metabolomic and sequence count data, integrative data analysis for association between biochemical profile and vaginal bacterial composition, and a comparative study between Stanford and UMD discoveries. The preprocessing is a very important section as it is required to provide standardized, generalisable, complete, unbiased data cleared from batch effects and noise. The results are presented in section 2.2 to give proper information for a discussion of the ethology of *bacterial vaginosis* in section 2.3.

Chapter 2

Corpus

2.1 Materials and methods

2.1.1 Population

40 non-pregnant women from the University of Maryland (UMD) were recruited at the University of Alabama (UAB - Birmingham, USA) and 40 pregnant women were recruited either at the University of Alabama or at Stanford University (Stanford, USA). The following attributes were recorded for all 80 women: age, body mass index, status (either **pregnant** or **non pregnant**), self-declared race or ethnicity (either **asian**, **black**, **hispanic/latino**, **white** or **other**). Additionally, status-specific variables were recorded:

- For pregnant women: gestational age in days (*i.e.*, the number of days since conception)
- For non-pregnant women: vaginal pH, menstrual bleeding flow (on a scale from 0 to 3, with 0 being *none* and 3 being *heavy flow*), and inferred menstrual cycle day (from -18 to 7, where -18 corresponds to 18 days prior to menses, 0 is the first day of menses and 7 is 7 days after the first days of menses) (Symul et al., 2023)

Admittedly, one expects some women to have a *Lactobacillus*-dominated vaginal microbiota and some to have a non-*Lactobacillus*-dominated microbiota. However, since the sample collection was not part of a controlled experiment, the number of women in each category of microbiota is not known in advance.

2.1.2 Sample acquisition

The samples took the form of a self-collected vaginal swab, which were then placed in a vial containing a fixed volume of storage medium. The vials would then be

handed to either Stanford University or the University of Alabama. All samples from pregnant women were stored and prepared at Stanford University, while all samples from non-pregnant women were stored and prepared at UAB.

Swabs were self-collected: weekly for pregnant women and daily for non-pregnant women. Among all swabs from each participants, five samples per women were selected prior to data preparation, amounting to 400 samples ($5 \text{ samples} \times 2 \text{ cohorts} \times 40 \text{ women}$) to be processed.

The 400 biological samples were then divided into two *batches* for metabolite and rRNA measuring: a batch of 280 samples and a batch of 120 samples, called *batch 1* and *batch 2* respectively. All 120 samples from batch 2 and 80 samples from batch 1 originated from Stanford, while the remaining 200 samples from batch 1 originated from UMD. Batch 1 samples were analysed first: see section 2.1.3 and 2.1.4 for analysis details.

A few elements need to be addressed about this sample acquisition and preparation.

- Since the vaginal swabs were self-collected and that no specific measures were taken to standardize the collection process, we expect some heterogeneity in sample quality. Some swabs could be of low quality (*e.g.*, instead of the vaginal wall, swabs of the vaginal labia were collected) or contain low quantity of biological material. Such non-ideal cases might give rise to potential issues in the content analysis of these swabs. Each case must be considered, as well as how it would affect the data generation.
- Although swabs were placed in a fixed volume of storage medium, the variability in total biological material on the swab has an impact on the concentration of material. One challenge will be to discriminate between true biological differences in concentrations and artificial variability due to the amount of material collected on the swabs. Whether this technical variability is small or significant is worth investigating in section 2.1.6 and 2.1.7.
- Since storage and preparation was performed in two different places (*i.e.*, Stanford and UMD) with limited efforts or ability to standardize protocols, there might be differences between samples due to this storage-preparation (*e.g.*, Stanford and UMD might use different preparation medium, different storing facilities or fridges, etc.). Since we have no information regarding how each university handled their samples, we cannot rule out the possibility of a "UMD vs Stanford" difference. The motor of this difference could be called the *site effect*. This site effect is unfortunately confounded with a potential *pregnancy effect* as Stanford's participants were recruited when pregnant and UMD recruited only non-pregnant participants. Thus, the differences observed between sites might also be due to being pregnant or non-pregnant

(the *pregnancy effect*). It is therefore not possible to distinguish whether the observed differences are due to pregnancy or due to university. This creates what one calls a *confusion effect*: the effect of the university site is confounded with the effect of pregnancy. In statistics, no tools, if only they existed, can help resolve this issue. Therefore, in the following sections, all preprocessing steps and analysis will be separated according to university (either Stanford or UMD). Only a cautious general comparative study, see section 2.2.5, can be made to identify whether or not the results are concordant in each facility.

- Finally, one can speculate the batch might also have an effect on the data. When analysing sample content in different batches, samples could be handled differently or the measuring machine might be calibrated differently, etc. All these elements contribute to a potential batch effect which must be considered when preprocessing the data, see section 2.1.6 and 2.1.7.

2.1.3 16S rRNA sequencing count data generation

Let us go over the means by which the vaginal microbiota composition was quantified and how this data generation might help us plan the processing of the data.

As mentioned previously in section 1.2, 16S rRNA sequencing has multiple steps. In the first step, for the 400 samples organized by batch, the bacterial cells were lysed and their DNA and RNA were extracted. The genetic material is amplified using Polymerase Chain Reaction. The amplified sequences are then sent through a sequencer (Illumina HiSeq/MiSeq platforms) which detects and counts all different types of sequences coming through. The next step in order is the mapping of the 16S sequences to their corresponding bacterial strain. However, sequencing technologies can introduce errors. To deal with this issue, the amplicon sequence variants (ASV) method was applied: sophisticated algorithms are used to correct the errors and accurately determine the original sequences present in the sample. ASV methodology, compared to the more traditional Operational Taxonomic Units (OTUs) methodology, yields higher resolution as it can distinguish two different sequences based on a *single nucleotide* (the elemental molecule of a genetic strand). The result of the 16S rRNA sequencing using ASV methodology is an accurate, consistent, and non-standardized contingency table of ASV sequences mapped to their species for each sample. (Symul et al., 2023) (Callahan et al., 2017a) (DADA2)

Notes on this contingency table must be addressed. First, the amplification process of the PCR was performed sample wise. Therefore, the count data after sequencing is proportional to the amount of genetic material at the sample level. This aspect has two implications: 1) even if samples had low quantities of biological material, the sequencing results would still give relevant information on the sample's microbial

composition; 2) one can standardise the contingency table through sample-wise scaling to make comparisons between samples (see section 2.1.6).

Additionally, the ASV method yields high resolution results in terms of error proportion. Hence, the number of inconsistent values will be low. In this context, *inconsistent value* refers to any and all value which is not correct: either a product of technical errors, a product of technological errors, or an outlier. The only potentially significant errors or inconsistent values which could arise from this measurement would be due to low-quality samples: there is a risk low-quality samples will not efficiently portray the subject's true compositional profile.

Now that we understand how the data was generated, we can detail its characteristics and plan the preprocessing phase more efficiently. Firstly, count data has no missing values: either the sequence is present (because the corresponding species producing the rRNA sequence is present in the sample) and the sequencer detects the number of sequences, or the sequence is absent (because there are no species producing this rRNA in the sample) and the sequencer counts 0 sequence. Secondly, there are very few inconsistent values in the contingency table. As mentioned before, the high resolution of the detection method makes few errors; and the low-quality samples will yield counts that just won't represent what we want to learn about. Third and finally, the data variability and correlated values can be decomposed into relevant and non-relevant information:

- Samples from the same woman will probably be correlated.
- Samples from different women can be correlated. When women have truly similar, dissimilar, and different independent microbial composition, one expects the correlation to be positive, negative, and null respectively.
- Samples from the same batch can be correlated
- Women with similar attributes (like age, BMI, gestational age, menstrual cycle status, etc.) might have correlated compositional profiles. In that case, it is important to avoid confounding effects (*i.e.*, not to confuse the compositional effect with the attribute effect).
- Part of the variability cannot be explained for different reasons: random correlations might occur, we don't have the full knowledge of the quantification method parameters, etc.

We can summarize the structure of the contingency table and the correlation sources in table 2.1.

Now that the structure of the datatable has been examined, we can plan our preprocessing more efficiently. Since the goal of the main analysis is to describe and quantify the *within and between women* variability, we need to remove or correct

Missing values	none
Inconsistent values	very few
Sources of variability	within women between women within batch due to pregnancy and cohort between women with similar attributes Random correlations due to errors (very few)

Table 2.1: Summary of data structure: missing values, inconsistencies, and variability sources in the 16s rRNA count data

for the other correlation sources. Indeed, we need to standardize the count data, to correct potential batch effect (if standardisation does not remove it first), and check for correlations between compositional profile and attributes. In addition to this, we will perform a mixed membership topic analysis to reduce the dimensions of the datatable. All of these steps are detailed in section 2.1.6.

2.1.4 Metabolite data generation

Similarly to microbial composition profiles, the metabolic profiles of the 400 samples were generated in two batches. Metabolic profiles were generated by an external company, Metabolon (Metabolon, a), for analysis via "semi-targeted" liquid chromatography mass-spectroscopy. In this semi-targeted detection method, most metabolites were identified. However, as we will see in section 2.2.3, some metabolites were not recorded in the reference database and therefore have been given a name starting with X-, such as X-99999. (Metabolon, b) (Metabolon, d)

The raw data generated by MS is a collection of spectra for each sample. Based on the sample quality and quantity, the raw data will be affected. Low-quality samples (or samples which were prepared and stored poorly) will most probably yield inconsistent spectra. Then, one can hope to identify these inconsistent values in the descriptive analysis of the metabolites. Low-quantity samples and non-optimal spectrometer calibration will generate missing values. Indeed, spectrometers can only detect molecular concentrations above a certain threshold at a certain calibration level. Consequently, if a sample contains small amounts of biological material amounts, even if it portrays well the sample's metabolomic profile, the molecular concentration will be low and the spectrometer might not detect it. In contrast to the preprocessing of RNA data, biological material concentration will be a great concern in the preprocessing of metabolomics data. To confirm whether or not the calibration was optimal and the sample concentration was high enough, we

need to check for missing value patterns and 'missing value - metabolite abundance' relationship.

We know for a fact metabolomic spectra have already been processed once. Unfortunately, we do not have much information regarding this processing. We only know the spectra were converted into metabolite peak intensities, measured in m/z or 'mass to charge', through a *peak alignment* step (Metabolon, c). Potentially, this processing step might have removed or created effects. Then, one can hope that the removed effects were not the ones of interest and that the created effects will be evaluated in the descriptive analysis. The peak alignment – which allows identification of the metabolites in the sample – could also make errors and generate inconsistent or missing values. Once again, these values need to be spotted and dealt with adequately.

Let us decompose the correlations and variability amongst the metabolomics datasets:

- Samples from the same woman could be positively correlated
- Samples from different women can be correlated: either women have similar metabolic profiles, "opposite" profiles or independent profiles
- Samples from the same batch might be correlated (potential batch effect)
- Women with similar attributes (age, BMI, gestational age, etc.) might show correlated profiles. In that case, it is crucial to identify and control for any confounding effects
- Variability due to material concentration in the vial could happen (this is called a *size effect*)
- Random correlations might occur due to spectrometry measuring errors and processing errors

Let us summarise the structure of the metabolomics dataset in table 2.2

In the preprocessing of our data, just like the RNA data, we want to single out the within and between women correlations and to remove any undesired effects. The missing values must be characterized and dealt with, either through filtration or through imputation, and inconsistent values must be addressed. We also need to remove batch and size effect and check for correlations between metabolite abundance and attributes.

Missing values	due to sample quantity, spectrometer calibration and processing errors
Inconsistent values	due to sample quality, measuring errors and processing errors
Sources of variability	within women between women within batch between women with similar attributes due to size effect random correlations due to errors

Table 2.2: Summary of data structure: missing values, inconsistencies, and sources variability in the metabolite peak intensity data

2.1.5 Data preparation

All following mentioned data manipulation, processing, graphical outputs, etc., were done using R version 4.3.3, and the coding scripts of these analyses can be accessed on the following Github repository: github.com/dupretf/2024_BSTA2M_Thesis_DF.

Data preparation involved importing the raw datafiles, selecting the information of interest and organizing it into a central structured object called a `MultiAssayExperiment` object (Ramos et al., 2017).

Data importation

The first step was to import the following datafiles:

- `mae_7_with_MC.Rds`,
- `Gates_metabolomics_batch_accounting.csv`,
- `gates_metabolomics_022420_stanford.csv`, and
- `gates_metabolomics_061219_UMD.csv`

Once imported in R, they were named `res`, `batch`, `stan` and `umd` respectively. Since the data originates from a larger research project than my master’s thesis, we already know a lot about it and what it contains. Therefore, with my promoter, we were able to identify relevant pieces of data for my project, omitting to consider other pieces. In that sense, some of the following procedures might be based on given instructions or prior information and need no justification.

The `res` and `batch` object both contain sample-related data. `res` also contains a `SummarizedExperiment` object with the 16S rRNA count data for the vaginal

microbiota composition of both UMD and Stanford cohort members (Morgan et al., 2024). Finally, objects `stan` and `umd` contain the metabolomic data of Stanford and UMD's samples respectively.

Data preparation and creation of the `MultiAssayExperiment` objects

A `MultiAssayExperiment` object allows to combine data from different assays in the form of `SummarizedExperiment` objects together with sample-related information common to all assays. The latter is stored as a `data.frame` in the `colData` slot of the `MultiAssayExperiment` object.

Since the analyses for the Stanford and the UMD cohorts will be performed in parallel, we chose to create two `MultiAssayExperiment` objects: one for each cohort.

Onto the preparation of sample-specific information datatable, we used both `res` and `batch`. The `batch` data was slightly pre-processed: the sample ID column was shortened, and a new column combining the cohort and the batch components into a single string of characters was created. This was the first element of the sample data (future `colData` slot). All other sample-related information was contained in the `res MultiAssayExperiment` (MAE) object. From this file, the following variables of interest to this project were selected: *SampleID*, *Subject*, *Status*, *Cohort*, *Race*, *BMI*, *Age*, *GestationalAge_days*, *Bleeding*, *PH*, *cycle_nb*, *cycle_nb_m*, *cycle_length*, *cycleday*, *MC_ok* and *Reprod_status*. After combining the variables with the batch data into a new dataframe called `coldata`, this newly created object was separated according to cohort: one column data object for Stanford, called `MB_stan_col` and one for UMD, called `MB_umd_col`.

The metabolomics datasets were tackled next: the metabolite-specific metadata such as associated KEGG pathways was isolated from the main metabolomic peak intensity data and renamed `MB_stan_row` and `MB_umd_row` accordingly. Two extra rows from the Stanford metabolomics data were discarded as they did not contain any data (completely empty rows). These main data tables, renamed `MB_stan_assay` and `MB_umd_assay`, were isolated from the rest. The assay data (peak raw intensities), metabolite-specific metadata and sample-specific metadata were collected into a `SummarizedExperiment` object called `MB_stan` and `MB_umd` accordingly. The row names and column names were also corrected to match those of the other assays.

As a second-to-last step, the 16S rRNA gene counts were prepared: only the variables of interest were selected in the column data. It was then separated according to cohort allocation into the object `VM_stan` and `VM_umd`.

With all assays ready, the two `MultiAssayExperiment` objects were created. They both comprised the microbiota composition data (16S rRNA gene counts – `VM_stan` or `VM_umd`) and metabolite data (peak intensities – `MB_stan` or `MB_umd`)

with their respective row and column metadata elements. These MAE objects will be the central data source for the subsequent data preprocessing steps and analyses.

2.1.6 Microbial composition data preprocessing

A quick inspection of the library size (*i.e.*, sample wise total sum of counts) was performed using a barplot, colored by batch. The second and third steps consisted of exploring relative abundance of species across samples using a compositional barplot, and applying a topic model for sample soft-classification.

Relative abundance

Microbiota data is often viewed as compositional: the absolute abundances matter less than the dominance of bacteria. Absolute abundances are also typically not well estimated by 16S rRNA sequencing as total library size often results from technical amplification rather than from true bacterial abundance. Consequently, raw counts are often transformed into relative abundances indicating the proportion of each species in each samples. Relative abundances are computed as

$$Y^{rel} = [Y_{ij}^{rel}]_{i=1,\dots,n;j=1,\dots,q} = \left[\frac{Y_{ij}}{R_j} \right]_{i=1,\dots,n;j=1,\dots,q}$$

where Y_{ij} is the absolute count for the i^{th} ASV in the j^{th} sample, $R_j = \sum_{i=1}^n X_{ij}$ is the library size of the j^{th} sample, and Y^{rel} is the new relative abundance table.

A compositional barplot was then constructed based on the 10 most abundant species of each cohort.

Throughout this project, a focus on the genus *Lactobacillus* and the lactic acid metabolite (lactate) was made as *Lactobacillus* is considered as a protective genus for vaginal health and is one of the sole producer of lactic acid. We thus expect strong correlations between *Lactobacillus* proportions and lactic acid concentrations in our samples, regardless of pregnancy status, batch, or cohort. Therefore these correlations will often be displayed and computed throughout the various processing and data transformation steps.

Topics models

In addition to transforming raw counts into relative abundances, we augmented our original data with an assay attempting to reduce the dimensionality of the relative abundance data as there were a total of 3590 species/ASV detected in our 2×200 sample dataset. Even after aggregating at the taxa level or aggregating ASVs with low total counts across sample, the dimension (number of feature) remained large

compared to the sample size (Armstrong et al., 2022). Traditionally, microbiota composition is summarized using classification techniques.

Diagnosis is an example of classification: a patient is either diagnosed with a pathology or is healthy. When considering microbiota composition, researchers have used clustering-based classifications methods applied to the proportions of microbial species in each sample. If there are K categories in this classification system, the k^{th} class *centroid* is defined by the proportions p_{jk} of species $j = 1, \dots, q$. For example, *class I* corresponds to microbiomes dominated by *Lactobacillus crispatus*. *Class IV* corresponds to microbiota not dominated by *Lactobacillus* spp. but composed of a wide variety of bacterial species (Kaambo et al., 2018) (Ravel et al., 2011) (De Seta et al., 2019). In the case of bacterial vaginosis, these classes are called community state types (CST). Additionally, there are variants of the five main classes (CST I through V) called sub-community state types (subCST) which take the name *CST I-A*, *CST III-B*, etc.

This classification can be used to infer transition probabilities from a healthy microbiome (CST I) to microbiome prone for bacterial vaginosis (CST IV). Although it is better than a binary system (*e.g.*, either healthy or BV), it is sub-optimal as analyses have shown that women’s vaginal microbiota cannot be defined and classified into a fixed number K of classes.

Instead, one relatively novel idea was to express the microbiota composition in terms of *proportions* of subcommunities. Maybe, a given sample is not 100% of one unique class, but rather 90% of a first "class" or subcommunity, 7% of another one, and 3% of a last one. This type of system is called a *mixed membership model*: samples are allowed to be a mixture of multiple subcommunities. Such system is an example of soft-classification.

This model relies on the assumption that such subcommunities (*i.e.*, synergistic consortia of bacteria) exist. The reality is likely more fuzzy. It has to be understood that mixed membership models attempt to find an optimal compromise between useful (*i.e.*, interpretable) dimension reduction and accuracy of the representation of the true biological composition of a women’s vagina. A recent study showed that mixed memberships provide a better description of microbiota samples than classification-based methods do and better allow to predict transitions from one state to another. (Symul et al., 2023)

Subcommunity composition and mixed memberships are here modelled using Latent Dirichlet Allocation (LDA, commonly referred to as "topic models"). Although Stanford and UMD cohorts have distinct characteristics and samples were processed and stored differently across sites, we made the assumptions that, if they exist, subcommunities would be conserved across pregnancy status and geographies and that potential biases induced by technical differences would be negligible compared to the biological synergies existing between bacteria. Based on

this assumption, topic models were fit on the *combined* UMD and Stanford 16S rRNA data.

In LDA model, one assumes the $n \times q$ matrix Y can be decomposed into two matrices $\gamma(n \times K)$ and $B(K \times q)$ such that $Y = \gamma' B$. This scenario applies to a matrix Y with $i = 1, \dots, n$ samples and $j = 1, \dots, q$ ASV sequences in columns that we wish to summarize into $k = 1, \dots, K$ distinctive subcommunities. The columns $[\beta_k]_{k=1, \dots, K}$ of matrix B represent the proportion of each sequence (*i.e.*, microbial species) in subcommunity k , while $[\gamma_i]_{i=1, \dots, n}$ gives the weight of sample i to the subcommunities $k = 1, \dots, K$. Through this modelling, one understands the samples $i = 1, \dots, n$ are perceived as mixtures of the subcommunities. (Fukuyama et al., 2022)

In terms of distribution, let $R_i = \sum_{j=1}^q Y_{ij}$ be the sample wise sum of counts of sequences, each Y_i is drawn independently according to:

$$\begin{aligned} x_i | \gamma_i &\sim \text{Mult}(R_i, B\gamma_i) \\ \gamma_i &\sim \text{Dir}(\lambda_\gamma \cdot 1_K) \end{aligned}$$

where λ_γ and λ_β are hyperparameters: "Large λ_γ and λ_β result in Dirichlet distributions that place more mass near the uniform distribution. Small λ_γ and λ_β place more mass on edges and corners of the simplex, resulting in sparser γ_i or β_k , respectively." (Fukuyama et al., 2022)

Regarding the third hyperparameter, the number of topics / subcommunities K , it can be chosen using cross-validation and/or using three diagnostics scores following the alignment: number of paths in the overall alignment, the distribution of topics coherence scores, and of topics refinement scores. The number of paths, let us call it h , is the number of distinctive subcommunities found throughout the models. This number is to be compared with the number of topics K in the model: h is either equal to or below K . Only when $h < K$ can one infer that two (or more) topics are very similar in composition and that the number of topics can be restrained to h .

One may also find the right K by choosing the model which maximizes the coherence scores. Coherence measures the consistency of a topic across a range of values of K . It is used to evaluate the "similarity between a given topic k and the other topics on the same path" (Fukuyama et al., 2022).

The refinement score can also be used to identify a good number of topics K . This score measures the divergence of a given topic in the next model. If a parent topic gives rise to only one descendant topic, the refinement is maximized and the parent topic is a true subcommunity. On the contrary, if a parent topic splits up into many divergent descendant topics, the score is minimized and there is no reason to believe the parent topic was relevant. Therefore the best number of topics K is a compromise between high refinement scores and low amounts of

topics, which translates to the start of a plateau in the refinement-vs-number of topics graph. (Fukuyama et al., 2022).

Constrained topics models

Previous studies where topic models have been applied to vaginal microbiota data have shown that the most common *Lactobacillus* species (*Lactobacillus crispatus*, *iners*, *gasseri*, or *jensenii*) tended to constitute sparse topics in which each species dominates one of the topics (Fukuyama et al., 2022) (Symul et al., 2023). However, rare non-*Lactobacillus* species were also sometimes present in small proportion in these topics, preventing a clear interpretation of topic proportions in samples from individuals containing these rare non-*Lactobacillus* spp. To prevent misleading readers into thinking that a participant may have a lactobacilli-dominated microbiota when their metagenomics data shows a non-*Lactobacillus* dominance, we resorted to additional constraints on the topic composition matrix (B/β).

The approach taken here was to first define the *Lactobacillus*-dominated topics using the *Lactobacillus*-assigned ASV counts, then the non-*Lactobacillus*-dominated topics using the counts for the remaining ASV. Results are presented in section 2.2.2.

Practically speaking, the topics model analysis was done using the `alto` package (Sankaran et al., 2024), a wrapper around the `lda` package, and which allows to examine how subcommunities relate to one another across models with different number of subcommunities.

We end this section by noting that subcommunities composition must still be interpreted with caution and one needs to remember that their generalizability is limited to the population of interest and might not be relevant for other populations. For information, all participants lived in the United States, and the age range (19–38) only covers a small subset of the general population.

2.1.7 Metabolite data preprocessing

Understanding what is inside a dataset and how one could transform it to obtain a clean and practical object is one of the first steps of analyzing data. In this section, we focus on the descriptive analysis and pre-processing of the metabolite datasets. As mentioned in section 2.1.4, the goal of the preprocessing of metabolite data is to remove any undesired effects while preserving the "true" within and between women data variability. For that to happen, we need to identify these nuisance factors, and, if they exist, attempt to remove them

The preprocessing of the metabolite datasets was a challenging task as there were many aspects of this data that needed attention and consideration: missing

values, variable distribution, mean-variance relationship, correlations between mean metabolite intensity per sample and subject attributes, batch effect, and size effect.

Missing values

The metabolite datasets contained a large proportion of missing values such that we could not simply exclude a few samples or metabolites for downstream analyses.

This large proportion of missing values required that we understand how these missing values were generated and how we should handle them. Missing values typically belong to one of three categories: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In MCAR data, missingness is independent of observed and unobserved data. In MAR, missingness is related to observed data but independent of unobserved data. In MNAR, missingness is related to unobserved data. (Mack et al., 2018)

Based on the quantification technique used here, a metabolite might be missing because its concentration was lower than the limit of detection or because it was mis-identified. To attempt to infer the overall contribution of both of these sources of missingness in our data, several diagnostic analyses and plots were achieved. Specifically, a scatter-plot comparing proportion of missing values per metabolite in both UMD and Stanford cohorts was drawn. The idea behind this plot was to test whether some metabolites were missing more often than others reflecting an overall low concentration or identification challenges for these often-missing metabolites. This was also inferred from a plot expressing the mean metabolite abundance as a function of missing proportion. If a trend exists, then one could hypothesize that most metabolites are missing because their concentration was lower than the quantification limit.

Based on these analyses' results, and thinking ahead, several methods could be used to handle missing values. If some metabolites were more frequently missing than others, it was deemed reasonable to filter those out and exclude them for downstream analyses. If, after this filtering, some samples still had high proportions of missing data, it was also deemed reasonable to exclude these samples.

For the remaining missing values, the decision to impute them or not was based on the results of the analyses described above. *Imputation* is the substitution of missing data by plausible values, had they been observed. Different imputation methods exist, some more complicated than others, some more efficient. When imputing data, there is a risk of creating effects that would never exist. Thus, one always needs to be cautious as no perfect imputation is possible. In our case, if analyses concluded that metabolites with higher proportion of missing values also had lower peak intensities, suggesting that missing values were mostly attributable to low concentrations rather than mis-identification, our approach was to consider imputation to a low value.

Metabolite distribution

Investigating variable distribution and considering which transformation of the data to use is crucial to any statistical analysis. Indeed, for downstream analysis, one usually needs variables with certain properties like identical, normal, or homoskedastic distribution.

Moreover, in some analysis, it is advised to use variables which have the same relative *weight* to prevent bias. This weight is proportionnal to the variable's variance: in methods like principal component analysis, variables with higher variance will be given greater weight than those with lower variance.

Peak intensities from mass spectrometry measurements do not inherently possess these distributional properties nor share similar variance. Hence, identifying suitable transformations for the metabolite abundances is necessary.

In data preprocessing, there are two key types of transformations that have distinct meanings and consequences: *scaling* and *transformation*. Scaling involves adjusting the features' values by dividing each value by a feature-specific scaling factor. This process gives each feature a similar variance, thereby making them comparable in magnitude and reducing potential biases in downstream analyses, but does not affect their distribution. On the other hand, transformation typically refers to applying a mathematical formula to each cell of the data table. This could involve logarithmic, square root, or other non-linear transformations that can normalize and thus modify data distributions, stabilize variances, or make relationships between variables more linear. Both scaling and transformation are crucial for preparing data for analysis, but they serve different purposes and are applied in different contexts based on the specific needs of the dataset. (van den Berg et al., 2006)

We drew boxplots or histograms to easily study distribution, and we investigated metabolite mean-variance relationship using mean-variance plots - a scatterplot with the variance of each variable as a function of the variable's mean. If the variance was independent of the mean (*i.e.*, the plot did not show a clear trend), we concluded that our data was compatible with the homoskedastic assumptions. Otherwise, we concluded that a variance-stabilizing transformation was necessary.

Here are the different scaling procedures and transformations considered in this project:

Scaling:

- Autoscaling: for a metabolite i of sample j , its autoscaled value y_{ij} derived from its original value x_{ij} is given by $y_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$ where $s_i = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_i)^2}$ is the standard deviation and $\bar{x}_i = \frac{1}{n} \sum_{i=1}^n x_{ij}$ is the arithmetic mean

- Pareto scaling: is similar to the autoscaling process but the scaling factor is different: $y_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$
- Range scaling: the scaling factor is computed as the difference between the maximum metabolite value $x_{i_{max}}$ and the minimum metabolite value $x_{i_{min}}$: $y_{ij} = \frac{x_{ij} - \bar{x}_i}{(x_{i_{max}} - x_{i_{min}})}$
- Vast scaling: uses the standard deviation s_i times the coefficient of variation ($cv = \frac{s_i}{\bar{x}_i}$) as scaling factor: $y_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i} \cdot \frac{\bar{x}_i}{s_i}$
- Level scaling: uses a location parameter rather than a dispersion parameter as scaling factor. In this case, we will use the mean: $y_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$

Transformation:

- Log10-transformation: applying the $\log_{10}(x_{ij})$ mathematical formula to every metabolite value
- Hyperbolic arcsine transformation: for a specified value x_{ij} , $\sinh(x) = \frac{e^{x_{ij}} - e^{-x_{ij}}}{2}$ where h represents hyperbolic and e is the constant equal to approximately 2.718; the inverse of the function is $\sinh^{-1}(x_{ij})$ or the $\operatorname{arcsinh}(x_{ij})$.

Principal component analysis

As a first-line exploratory analysis, a principal component analysis was performed to identify underlying data structures. We investigated the eigenvalues, the scores, and the loadings, to identify any batch or size effects in both cohorts.

Batch effect

Stanford biological samples were sent for treatment and measurements in two different batches, referred to as *batch1* and *batch2*. Such a procedure could lead to a batch effect.

A batch effect is a difference of distribution between the different levels of the batch variable. In our study, this batch effect would translate to a difference of metabolite abundance distribution between samples from batch 1 and samples from batch 2. It might occur if, for example, the chemical buffer used for spectrometry was different when analysing batch 1 samples compared to batch 2 samples, or if the spectrometer wasn't calibrated the same way. Thus, the sources of batch differences are probably systematic, artificial, and technical errors. A batch effect would influence the downstream results and interpretations. (Yu et al., 2023) (Liu et al., 2020)

Besides, one must also check that the samples from batch 1 and 2 come from the same population of women. Indeed, by chance, some population attributes between the two batch-associated cohorts might differ. For example, samples from batch 2 could accidentally belong to women of a distinctive age group compared to women in batch 1. In that case, the batch would be confounded with age and by correcting for the batch effect without accounting for these differences in age, we may introduce age-related biases in our dataset.

For that reason, and in addition to the results of the PCA, we used Bioconductor's *limma* package to assess the importance of the batch effect (Ritchie et al., 2015) (Liu et al., 2020) and analysed all potential differences between the two Stanford's batch-associated cohorts. We checked for contrasts in mean abundance of metabolites, distribution of missing values, proportion of racial groups, age distribution, and gestational-age distribution. Density plots, proportions, chi-square test, and non-parametric tests of distribution equality were computed for this analysis.

limma package was also used to correct for batch effect by taking advantage of the `removeBatchEffect()` and `lmFit()` functions. On the one hand, the `removeBatchEffect()` function will remove any systematic variation due to batches or any given covariates. Thanks to last section's analysis on batch and batch correlated variables, we will be able to provide these covariates. This function can implement a design matrix to take into account other parameters, like batch correlated variables.

On the other hand, the `lmFit()` function will fit a linear model. It can implement potential covariates in the process through a design matrix. The idea is to use the linear model to test or inspect whether the batch correction worked or not. Let us take the same design matrix for both the batch effect correction and linear model.

On a side note, one advantage of the `removeBatchEffect()` function is that it works on data tables which may contain missing values (see appendix 4.5 to learn more about batch correction packages and *limma* approach). At this stage of the project, we will use the latest processed dataset which is filtered, imputed and transformed. Future works on this data could focus on the batch effect removal of differently preprocessed data or differently ordered processing steps. Indeed, one could work on the initial unscaled filtered dataset, or try the hyperbolic arcsine scaled filtered dataset, or another transformation.

To compare before and after batch effect correction, a boxplot of the samples' distribution and *limma* linear model table of differential regulation are computed.

Size effect

As mentioned earlier, the total amount of biological material collected on swab might be highly variable. This variability in total concentration may induce spurious correlations between all metabolites.

Unfortunately, since that quantity has not been measured, it is not possible to discriminate between potential biological reasons for increased overall metabolite concentrations and technical reasons (*i.e.*, amount collected on swab).

We remind the reader that samples were collected by participants who were asked to take a vaginal swab on their own and put the swab in a vial filled with a liquid collection medium. The vial would then be treated and their content transferred in spectrometer for measuring. Of course, the volume of that collection medium is supposedly identical for every swab (and the variability associated with that medium volume is assumed to be insignificant). However, the amount of biological material on the swab could potentially differ greatly between individuals and between time points. Therefore, the *ratio between swab material and medium volume* could vary considerably. This variability in material concentration influences the characteristics of the data. If this variance is great enough, then one could see that even a very frequent metabolite, such as lactate, could vary widely from one sample to another only because the material-medium ratio is considerably different. This applies to every metabolite.

This conjoint variability in metabolite concentration is often referred to as a *size effect*, and is directly linked to the overall metabolite abundance.

In an attempt to understand more about this size effect, an analysis of the relationships between host attributes and the overall metabolite abundances was carried out. Specifically, we tested the correlation between mean biochemical peak intensity and menstrual cycle phase, cycle length, gestational age, BMI, age, and pH.

Another important variable to check in our case is the proportion of *Lactobacillus* species per sample. Indeed, one might argue that the concentration of metabolites in a sample is correlated to the composition of the vaginal microbiota, especially the presence of characteristically dominant *Lactobacillus* spp. A correlation between mean metabolite peak intensity and *Lactobacillus* proportion would indicate that *Lactobacillus* have a strong influence on the vaginal metabolic profile. This could be a direct effect (*e.g.*, *Lactobacillus* could be strong or weak producer of metabolites beyond their lactic acid production or induce an elevated or decreased production of metabolites by host epithelial cells), or an indirect effect mediated by, for example, the amount of cervicovaginal mucus produced (and thus collected on swab) in *Lactobacillus* vs non-*Lactobacillus*-dominated environments.

Furthermore, we performed an ANOVA test for equal metabolite abundance across levels of the categorical variables 'reproductive status' and 'cycle number'.

Besides the link with host attributes, the size effect, if it exists, is correlated to the first principal component: the dimension capturing the greatest data variability. If we believe that the origin of that effect is technical rather than biological, one needs to adjust for this size effect to make the samples more comparable. To do

so, we performed a principal component analysis of the datatable of Stanford and UMD using function `prcomp`. Afterwards, the first component / vector of scores are manually set to 0 as the idea was to eliminate the variability contained in PC 1. Then, an approximate data matrix is computed by multiplying the new scores matrix T by the loadings matrix P such that $X = T'P$.

Since the scores and loadings are computed on a standardized dataset, the next step in order is to revert to a "non-standardized" dataset: the biochemical-wise standard deviation from the 'old' data is multiplied to the PC1-removed data and the biochemical-wise mean is added to the PC1-removed data. Mathematically speaking, the principal components were computed on centered and scaled data $X^{cs} = [X_{ij}^{cs}]_{i=1,\dots,n;j=1,\dots,p}$ where

$$X_{ij}^{cs} = \frac{X_{ij} - \bar{X}_i}{\sqrt{\sigma_i^2}}$$

and $\bar{X}_i$ is the metabolite mean and σ_i^2 is the metabolite variance. Therefore, the score matrix was $T_{X^{cs}}$ and the approximated PC1-removed data was $\hat{X}^{cs}$. To obtain comparable data, one needs to compute

$$\hat{X} = [\hat{X}_{ij}]_{i=1,\dots,n;j=1,\dots,p} = \hat{X}_{ij}^{cs} \cdot \sqrt{\sigma_i^2} + \bar{X}_i \text{ for } i = 1, \dots, n; j = 1, \dots, p$$

One can verify whether the size effect has been properly removed in data $\hat{X}$ by performing a PC analysis, computing the eigenvalues, the percentage of variance explained, drawing a loadings' plot, etc. Additionally, this PCA is used to verify the batch effect removal in the scores' plot.

Lactate abundance vs *Lactobacillus* proportion

Since it has been well established that *Lactobacillus* bacteria are the main producer of lactate in the vagina, we used the correlation between lactate abundances and *Lactobacillus* proportion as a "quality control" at each steps of the pre-processing (raw data, filtered data, transformed data, batch effect removed data, etc.).

Scatter-plots of the lactate intensity as a function of *Lactobacillus* proportion were plotted, with a dotted linear regression line and confidence interval to give a broad idea of the correlation. A table with the estimated Spearman correlations and their respective p-values is also computed.

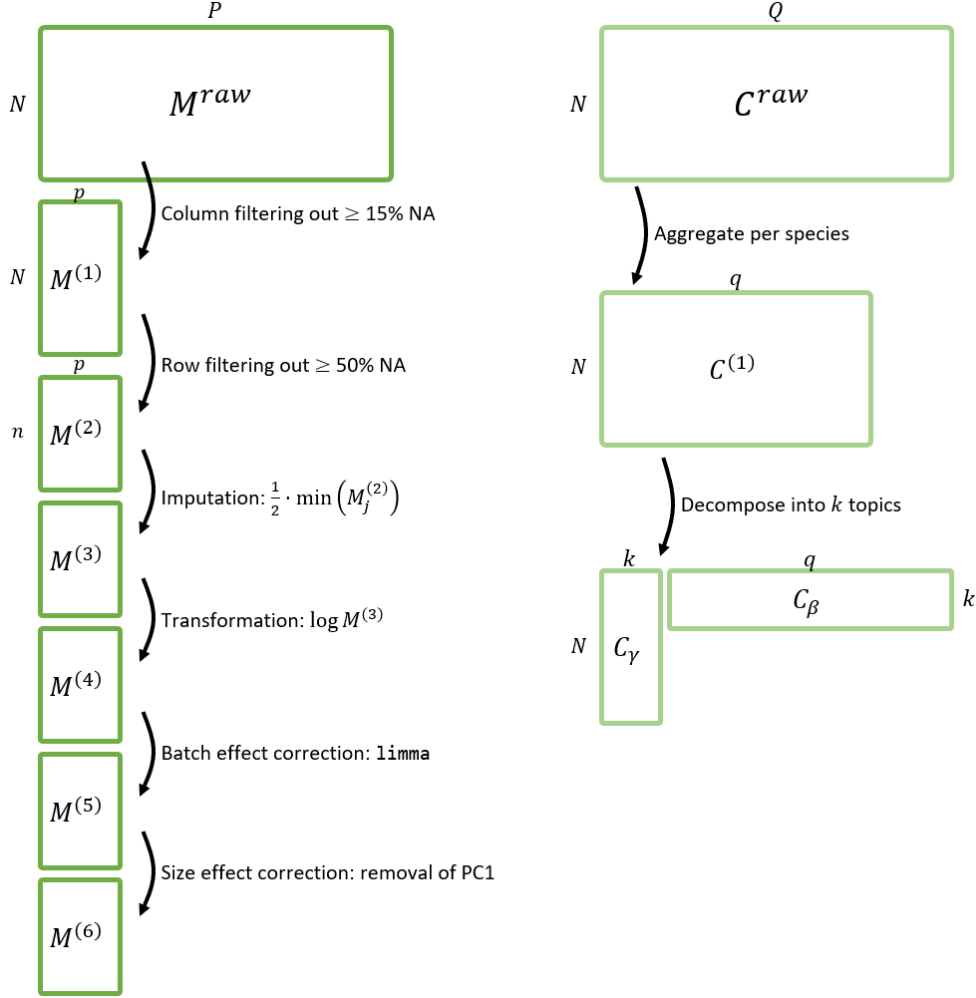


Figure 2.1: Summary of pre-processing steps for metabolite table M and compositional table C represented as matrix manipulation. Rows are referred using i subscript, columns are referred using j subscript

2.1.8 Metabolite and Compositional data integration

To quantify and visualize the associations between metabolomic profile and microbial composition datatables, several methods have been proposed. Here, we chose to measure the RV coefficient between the two blocks and to perform a STATIS analysis (L'Hermier des Plantes, Structuration des tableaux a trois indices de la statistique), Simultaneous Component Analysis (SCA) (Levin, 1966), and Canonical Correlation Analysis (CCA) (Hotelling, 1936). Once again, analysis were performed separately for Stanford and UMD samples.

First, let us detail the notation used to explain the three types of analysis:

- Let the number of datatables be $k = 1, \dots, K$. In our case, $K = 2$ because we have the metabolite table (let us call it $X_{k=1}$) and the compositional table (let us call it $X_{k=2}$). In the literature (and the following paragraphs), tables X_k are either referred to as "tables", "blocks", or "block matrices".
- The dimensions of the block matrices are $n_k \times p_k$
- The following analyses rely on "common-row" blocks with different variables, meaning the observations of the blocks must be identical (different features quantified for the same samples). In our case, both metabolite and composition tables share the same samples (from the same 80 women). Therefore, we can write $n_1 = n_2 = n$ and simplify the dimensions of the blocks to $n \times p_k$

We note that, because we filtered samples in the metabolite preprocessing, we must match the remaining samples in the compositional data, resulting in the loss of some information in both tables.

RV coefficient

The RV coefficient (Robert and Escoufier, 1976) measures the degree of similarity between a $(n \times p_1)$ matrix X_1 and a $(n \times p_2)$ matrix X_2 using the variance-covariance matrix of these block matrices. Assuming X_1 and X_2 are mean-centered blocks, this $(p_1 \times p_2)$ variance-covariance matrix is defined as:

$$\Sigma_{12} = E(X_1' X_2) = \frac{1}{n-1} X_1' X_2$$

Then, the RV coefficient between X_1 and X_2 is defined as:

$$RV(X_1, X_2) = \frac{\text{tr}(\Sigma_{12} \cdot \Sigma_{21})}{\sqrt{\text{tr}\Sigma_{11}^2 \cdot \text{tr}\Sigma_{22}^2}}$$

The RV coefficient takes a value between 0 and 1. If two blocks X_1 and X_2 are completely independent, then $RV(X_1, X_2) = 0$. Conversely, if X_1 and X_2 are identical, then $RV(X_1, X_2) = 1$.

There are multiple ways to compute an RV coefficient in R. In `ade4` package, we can use the function `RV.rtest()` which performs a Monte-Carlo test. In the Monte-Carlo simulations, the rows of each datatables are permuted randomly and the total inertia of the tables is computed. This inertia is the sum of all eigenvalues of the co-inertia analysis (*i.e.*, an analysis aiming to find pairs of linear combinations U and V of the datasets' variables which maximize covariance:

$cov(U_i, V_i) = cor(U_i, V_i)\sqrt{var(U_i)var(V_i)}$, "that is the sum of the covariances between successive pairs of axes". Through the permutation, the simulation may build the distribution of the inertia. The pvalue of the test is given by the frequency of observing the inertia value of the original datasets. (Thioulouse et al., 1995)

STATIS

The STATIS analysis, or *Structuration des Tableaux A Trois Indices de la Statistique* is a method founded by L'Hermier Des Plantes H. (1976), with the help of his promoter Escoufier Y., to initially study the existing associations between K datatables with the following characteristics: n common observations of p common variables, measured K different times. It was then adapted to common-row tables with n rows and $[p_k]_{k=1,\dots,K}$ columns. The main idea was to infer the *trajectory* or *situations* of these n individuals over time. To make such inference, L'Hermier Des Plantes thought of making a *compromise* of all K situations. (L'Hermier des Plantes and Thiébaud, 1977)

In a way, STATIS method is similar to principal component analysis, in the sense that it finds latent (compromised) variables with the highest variability across *situations*. However, a novel but key element to this PCA is the consideration of the weight given to each situation. The STATIS method can be considered in three steps:

First step (Interstructure)

- Compute the $n \times n$ similarity matrix $S_k = X_k X_k'$ for $k = 1, \dots, K$. This matrix measures the similarities between individuals (*i.e.*, samples in this project). We used the centered and scaled block matrices X_k so that S_k describes a sort of a "correlation" between individuals. To give the reader some intuition, this previous matrix S_k is homologous to the correlation matrix of a datatable X_k (if X_k is centered and scaled) which transcribes variables relationships. However, instead of describing *variability* amongst variables, S_k describes the *similarity* between individuals.
- Compute the $n^2 \times K$ matrix S by concatenating the vectorized matrices $[S_k]_{k=1,\dots,K}$ such that $S = (\text{vec}(S_1) \text{ vec}(S_2) \dots \text{vec}(S_K))$ where

$$\text{vec}(S_k) = \begin{pmatrix} [S_k]_{11} \\ [S_k]_{12} \\ \vdots \\ [S_k]_{nn} \end{pmatrix}$$

- Compute the $K \times K$ matrix RV between the K blocks such that:

$$[RV]_{kl} = \frac{[\Sigma]_{kl}}{\sqrt{[\Sigma]_{kk}}\sqrt{[\Sigma]_{ll}}} \quad k = 1, \dots, K; l = 1, \dots, K$$

- ...where $\Sigma = S'S$. The element of the k^{th} row and the l^{th} column of matrix RV gives the RV coefficient between blocks k and l : a coefficient value of 0 corresponds to entirely independent blocks and 1 corresponds to identical blocks.
- Perform eigendecomposition of RV such that

$$RV \cdot W = \lambda \cdot W$$

- ...where W is the $K \times K$ matrix of eigenvectors and λ is the $K \times 1$ vector of eigenvalues. The subspace created by the eigenvectors is referred to as the *interstructure*. By multiplying the eigenvectors by the square root of the eigenvalues, one obtains the $(K \times K)$ standardised contribution matrix Q (i.e., $Q' = W'\sqrt{\lambda}$) of the K blocks to this interstructure. We then use the coordinates of the blocks in Q to infer the similarity between two or more tables $[X_k]_{k=1, \dots, K}$. If the relative position of the samples is similar in both tables X_k and X_l , then the tables are considered similar and the vectors $O\vec{Q}_k$ and $O\vec{Q}_l$ will be collinear on the interstructure space. The vectors $[O\vec{Q}_k]_{k=1, \dots, K}$ are usually studied in 2-dimensions (because comparisons are easier in 2 dimensions and because the first two components hold the greatest percentage of explained variance), referred to as s_1 and s_2 , but additional dimensions may need to be considered.
- Additional notes on the interstructure can be found in appendix 4.6.

Second step (Compromise)

- Compute the $n \times n$ compromise matrix $\mathcal{C}$ using: 1) the absolute value of the diagonal elements $[\Sigma]_{kk}$ of matrix Σ , and 2) the previously defined first eigenvector $W^{(1)} = (W_1^{(1)} \dots W_K^{(1)})'$ of the RV matrix such that:

$$\mathcal{C} = \sum_{k=1}^K \frac{|W_k^{(1)}|}{\sqrt{[\Sigma]_{kk}}} S_k$$

- One can interpret the *best compromise* between all blocks as the weighted average, using weights from the first eigenvector of RV , of all similarity

matrices $[S_k]_{k=1,\dots,K}$. The reason why the first eigenvector is used and not any other vector is because the first one is associated with the greatest eigenvalue, and therefore translates the latent dimension with the greatest variability across blocks.

Third step (Intrastructure)

- Perform PCA (using eigenvector decomposition method) on compromise matrix $\mathcal{C}$.
- The number of eigenvalues of $\mathcal{C}$ is $A = \min(n, \sum_k p_k)$. Because $\mathcal{C}$ is a compromise between the K blocks, the variance decomposition portrayed by the eigenvalues combines the variance associated with variables from all blocks.
- The PCA's $(n \times A)$ scores matrix T of the compromise $\mathcal{C}$ translates a "compromised position" of each sample from the K datatables in a subspace called the *intrastructure*.
- The $(p_k \times A)$ loadings P_k of each block are computed as $X'_k \cdot T$ and describe the contribution of the variables to the principle components (L'Hermier des Plantes and Thiébaud, 1977) (Kiers, 1989) (Lavit, 1988) (Fabien et al., 2020)

According to Van Deun et al. (2009), the STATIS model can be written as an optimization problem of

$$\arg \min_{T, P_k} \sum_{k=1}^K W_k^{(1)} \cdot \|X_k - T P'_k\|^2$$

where $W_k^{(1)}$ is the k^{th} element of the first eigenvector $W^{(1)}$ of the RV matrix, X_k is the k^{th} block matrix, and T and P_k are the scores matrix and loadings matrix, respectively, of the principal component decomposition of X_k (such that $T = X_k P_k$). One may notice that the loadings matrix, unlike the scores matrix, is block-dependent. It is because the n individuals are the same across blocks but the variables are not. (Bougeard and Dray, 2018)

We used the `statist` function of `ade4` package to implement the STATIS analysis. By default, this function takes the first two dimensions, denoted c_1 and c_2 to compare individuals across tables in the intrastructure sub-space, but 2 additional dimensions were considered in this project. We plotted the following elements: the interstructure, a barplot of the variance decomposition of the compromise axis, a score plot on the first 4 axis, and the intrastructure.

Simultaneous component analysis

Simultaneous component analysis (SCA), proposed by Levin (1966), is very similar to the STATIS method. It was originally designed to compare different groups of individuals on which one had measured the same set of variables. However, it has been adapted to study datasets of different sets of variables on the same group of individuals. (Liland, 2023)

The idea is still to perform a PCA on a weighted average of all similarity matrices $[S_k]_{k=1,\dots,K}$, but in SCA, the weights are evenly distributed. (De Roover et al., 2012a) (De Roover et al., 2012b) (Kiers, 1989) (Levin, 1966). In that sense, the objective of the SCA is to optimize the following criterion:

$$\arg \min_{T, P_k} \sum_{k=1}^K \|X_k - TP'_k\|^2$$

One may compare it with the optimization problem of the STATIS method and notice that the weights $W^{(1)}$ are missing in SCA. That is simply because each table is given the same weight in SCA. (Van Deun et al., 2009) (Smilde et al., 2005) (Kassama et al., 2010)

In his 1966 paper, Levin emphasizes the centering and scaling of the input blocks $[X_k]_{k=1,\dots,K}$ prior to computing the similarity matrices S_k , compared to the STATIS method which did not highlight the need for the scaling step. However, we used the centered and scaled data blocks for STATIS and SCA analysis, therefore any difference due to input tables have been avoided.

One can then identify two sources of differences between STATIS and SCA results. Firstly, the weights given to each block is different. Thirdly, the PCA performed on the compromised matrix is rotation-invariant in the sense that the principal components could be oriented differently, without changing the information they contain. For that last reason, for instance, scores from STATIS could potentially be rotated to find similar scores in SCA.

The SCA was performed using `sca` function from `multiblock` package and different plots were analysed: a barplot of the variance decomposition of the principal components, a scores plot on the first 4 axis, and loadings plot of each data block.

Canonical Correlation Analysis

Canonical correlation analysis (CCA), proposed by Hotelling (1936), is the exploration of pairs of vectors $U = X_1a$ and $V = X_2b$ which maximize the correlation $\rho = \text{corr}(U, V)$ such that:

$$\arg \max_{a,b} \text{corr}(X_1a, X_2b)$$

where X_k is $(n \times p_k)$, a is $(p_1 \times 1)$, and b is $(p_2 \times 1)$. Basically, we look for coefficients $a = (a_1, a_2, \dots, a_{p_1})$ and $b = (b_1, b_2, \dots, b_{p_2})$ to form linear combinations $U = X_1 a$ and $V = X_2 b$ of the variables of the two datasets X_1 and X_2 which maximize a given correlation measure between the two response U and V , also named *variates*.

CCA may find more than one couple of linear combinations U_1 and V_1 and further search for additional variates $[U_c]_{c=1, \dots, C}$ and $[V_c]_{c=1, \dots, C}$ which also maximize the correlation ρ_c . These new variates were constrained to be perpendicular to the previous computed variates for orthogonality purposes.

But one must be aware of the hypothesis behind CCA. Firstly, the CCA model relies on the assumption of linear relationships between the variables of the datasets. If such assumption cannot be checked, the results might be biased. Secondly, the number of rows in the blocks must be greater than the number of columns: even though omics data analyses involve large datasets with the number of variables p usually greater than the number of observations n , the standard canonical correlation analysis relies solely on datasets with dimensions $n \times p$ where $n > p$. Sparse CCA has been introduced to deal with situations where the number of features is larger than the number of samples (Parkhomenko et al., 2007), but this is outside the scope of this project and was not implemented in the results' section. Thirdly, CCA can only be applied to full-rank data blocks: if one or more variables from either block can be written as a linear combinations of the other variables, typical CCA implementations will yield error messages.

For our data, we assumed the linear relationship is not an issue and checked that the dimensions were correct. However, for the compositional datasets, we had to discard one variable. Since the compositional data describes the proportion of topics (in columns) for each sample (in rows), the row-wise sum is always 100%. In that case, the row-wise value of any one-column x_{iJ} can be calculated based on the values of the remaining columns such that $x_{iJ} = 1 - \sum_{j=1}^{J-1} x_{ij}$. Therefore, there will always be one too many variable. To solve the issue, we discarded *IV-O.b* because *IV-O.a* and *IV-O.b* were not very well defined topics, with *IV-O.b* being the subcommunity gathering all species that did not belong in any other topic.

A standard CCA was performed using `cc()` (González et al., 2008) (Bruin, 2011) and a few plots were drawn: a barplot of the canonical correlations ρ_c and loadings plot on the first four variate axis. In addition to this, to improve interpretation, the eight most contributing metabolites to the first variate, in absolute value, of each dataset were identified along with their respective coefficient.

2.1.9 Comparative study

A qualitative and observational comparative study was initiated to balance the the occurrence of the integrative analysis results in Stanford and UMD cohort.

Because the cohorts suffer from an unsolvable confusion effect (recruitment site effect confounded with pregnancy effect), their respective results cannot be compiled together without making mistakes. For example, one cannot say the difference in associations found in Stanford and UMD is due to pregnancy (because it might also be a product of recruitment site). Hence, we cautiously compared the outcomes of the analysis made on Stanford and UMD’s data and tried to give robustness to some discoveries.

To compile the information from the integration analysis and the identification of associations between metabolites and subcommunities, we used a colormap (see results in section 2.2.5). Each cell of the colormap is meant to represent the association found in either Stanford (colored in red), UMD (colored in red), or in both (colored in purple). However, the white cells in this graph do not automatically correspond to an absence of association. Though an association might exist, it was not always visible enough to contribute to the graph. Moreover, one has to take into account that the graph does not tell the direction of the associations (*i.e.*, positive or negative).

2.2 Results

2.2.1 Data preparation

The prepared data, in the form of MAE objects, contains different elements:

- MB = pre-processed peak intensities of metabolites samples measured using semi-targeted mass spectrometry (MS). This `SummarizedExperiment` contains an assay with 200 columns and 853 rows in Stanford, and 200 columns and 853 rows in UMD. There are approximately 5 samples from 40 pregnant women in Stanford and 5 samples from 40 non-pregnant women in UMD.
 - Its `rowData` has 853 columns (for both Stanford and UMD) expressing the name of the biochemicals / metabolites, the super-pathway to which they are part of, the sub-pathway to which they are part of, the compound’s mass and its KEGG (Kyoto Encyclopedia of Genes and Genomes) identification code.
 - The `colData` with batch related information (batch, sample ID, cohort) and clinical data for each cohort member like gestational age, age, BMI, vaginal pH, cycle day, etc., for a total of 18 variables in Stanford and 18 in UMD.
- VM = 16S rRNA sequence counts of 3590 (3590) identified microbes from the vaginal ecosystem of Stanford patients (UMD patients respectively). These

```

## A MultiAssayExperiment object of 2 listed
## experiments with user-defined names and respective classes.
## Containing an ExperimentList class object of length 2:
## [1] MB: SummarizedExperiment with 853 rows and 200 columns
## [2] VM: SummarizedExperiment with 3590 rows and 193 columns
## Functionality:
## experiments() - obtain the ExperimentList instance
## colData() - the primary/phenotype DataFrame
## sampleMap() - the sample coordination DataFrame
## '$', '[', '[' - extract colData columns, subset, or experiment
## *Format() - convert into a long or wide DataFrame
## assays() - convert ExperimentList to a SimpleList of matrices
## exportClass() - save data to flat files

## A MultiAssayExperiment object of 2 listed
## experiments with user-defined names and respective classes.
## Containing an ExperimentList class object of length 2:
## [1] MB: SummarizedExperiment with 853 rows and 200 columns
## [2] VM: SummarizedExperiment with 3590 rows and 200 columns
## Functionality:
## experiments() - obtain the ExperimentList instance
## colData() - the primary/phenotype DataFrame
## sampleMap() - the sample coordination DataFrame
## '$', '[', '[' - extract colData columns, subset, or experiment
## *Format() - convert into a long or wide DataFrame
## assays() - convert ExperimentList to a SimpleList of matrices
## exportClass() - save data to flat files

```

Figure 2.2: R output structure of MAE object for Stanford (right) and UMD (left) cohort

counts come from 193 (200) samples of the same 40 pregnant women in Stanford (or the same 40 non-pregnant women in UMD).

- This object also contains row data of dimension 3590×15 in Stanford and 3590×15 in UMD.
- It also contains its specific column data of dimension 193×6 (200×6) with information specific to the RNASeq measurements (like the library size).

2.2.2 Microbial composition data preprocessing

This section presents the results of exploratory and descriptive analyses, and pre-processing steps applied to the microbiota data, as described in the previous sections 2.1.6 and 2.1.7.

We remind the reader that microbiota data presented here corresponds to RNA sequencing count data of 16S rRNA. There were 193 samples in Stanford, 200 samples in UMD, and 3590 sequences in for both Stanford and UMD RNAseq.

Row metadata

The row metadata associated with the microbiota set is composed of fifteen columns and is identical in Stanford and UMD datasets:

- **ASV**: a portion of the species-specific 16S rRNA Amplicon Sequence Variant
- **ASV_assay_key**: the database key for identifying the species
- **ASV_username**: a shorter name for the assay key
- **total_counts**: ASV total count (the sum of all counts across samples)

- **Lowest_Level_Available:** Lowest taxonomic level available for the identified sequence
- **Lowest_Level_Info:** Lowest taxonomic name (corresponding to the lowest taxonomic level) according to the database
- **Kingdom, Phylum, Class, Order, Family, Genus, Species:** information regarding the phylogeny of the identified species. Takes a blank value "" if unknown.
- **Species_full:** sub-species identification according to the database; takes the value NA if inexistent or unknown.
- **strain_nb:** number of species from the same strain. For example, if we had identified 3 *Lactobacillus* spp. then all of them would have a strain number of 3.

To give a broad example, table 2.3 shows a contingency table of identified microorganisms of all samples according to phylogenetic kingdom (domain) and phylum levels.

Total library size

The *library size* for each sample, given the name **SampleSum** in our data, is presented in the form of a barplot in figure 2.4. For Stanford, it is clear the library size is batch-dependent. However, this was not a problem as we used the relative abundance rather than the absolute abundance in ulterior steps.

Relative abundances

As mentioned before, since microbiota is often viewed as compositional, relative abundance is a natural transformation of the data.

Relative abundance has the advantage of simplifying interpretations. For example, noting that the proportion of *Lactobacillus crispatus* is correlated with tyramine concentration might be helpful for associating metabolites and microbiota composition. Indeed, this is the objective of the main study: to identify and understand the correlations and associations between metabolic profile and microbial composition.

The relative abundance table for Stanford and UMD can be visualized via the barplot of figure 2.5.

	Archaea	Bacteria	Eukaryota
	0	3	0
Acidobacteria	0	3	0
Actinobacteria	0	475	0
Armatimonadetes	0	2	0
Bacteroidetes	0	662	0
Chlamydiae	0	2	0
Chloroflexi	0	6	0
Cyanobacteria	0	16	0
Deinococcus-Thermus	0	18	0
Dependentiae	0	1	0
Elusimicrobia	0	1	0
Epsilonbacteraeota	0	30	0
Euryarchaeota	5	0	0
Fibrobacteres	0	1	0
Firmicutes	0	1605	0
Fusobacteria	0	58	0
Gemmatimonadetes	0	1	0
Parabasalia	0	0	5
Patescibacteria	0	10	0
Planctomycetes	0	12	0
Proteobacteria	0	602	0
Spirochaetes	0	15	0
Synergistetes	0	12	0
Tenericutes	0	33	0
Verrucomicrobia	0	11	0
WS1	0	1	0

Figure 2.3: Contingency table of identified microorganisms according to kingdom and phylum

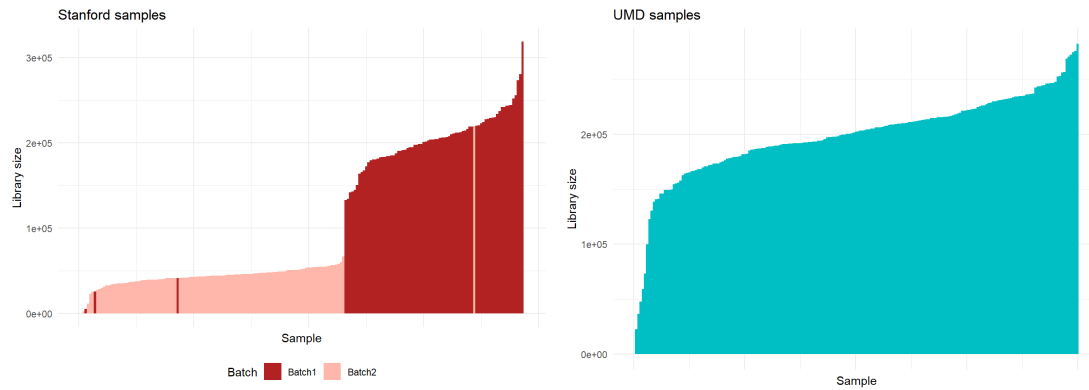


Figure 2.4: Library size of sample's metagenomic data (ordered by increasing value) in Stanford (left) and UMD (right)

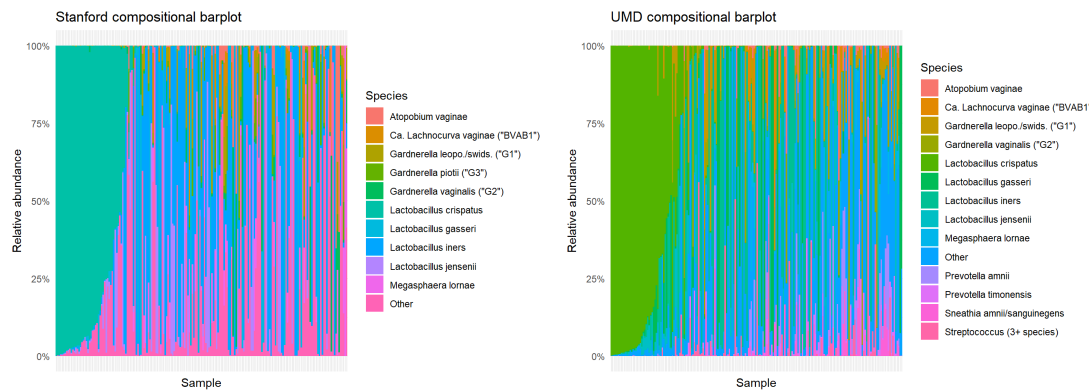


Figure 2.5: Compositional barplot of the 10 most abundant bacterial species of Stanford (left) and UMD (right) samples

***Lactobacillus* proportions**

Bacterial genus *Lactobacillus* is quite important in this project as it is key to healthy vaginal microbiota. *Lactobacillus* spp. dominated vaginal microbiota usually prevent diseases, infections and other unwanted circumstances.

Broadly speaking, the popular understanding of the mechanism through which *Lactobacillus* spp. monopolize the microbiota and "keep it in check" is by producing lactic acid (or "lactate"). This acid raises the pH in the environment, thus making it unbearable for most non-*Lactobacillus* bacterial species. High pH prevents many bacteria from functioning correctly and multiplying. Consequently, one can expect a great positive correlation between lactate concentration and *Lactobacillus* spp. presence, and a negative correlation between lactate and non-*Lactobacillus* species, see figure 2.6. One could also go into detail and look for the proportion of other

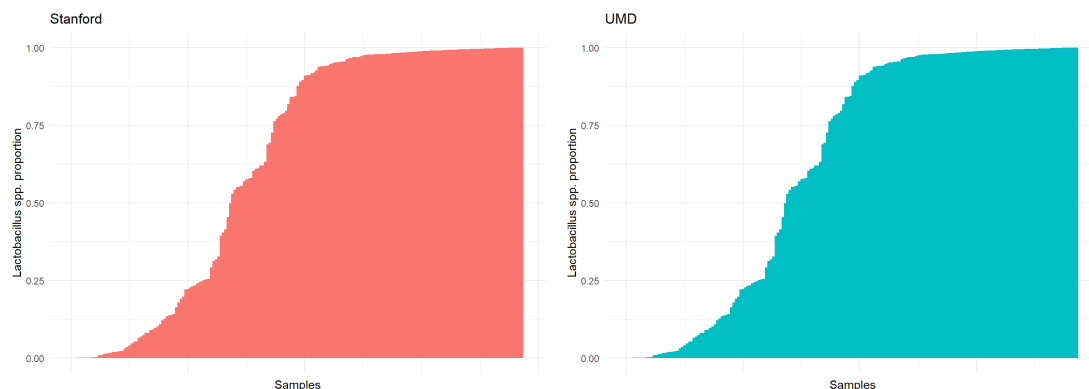


Figure 2.6: Ordered proportion of *Lactobacillus* spp. in Stanford (left) and UMD (right) samples

microbes like *Lactobacillus iners* and *Gardnerella vaginalis*, see figure 4.1 and 4.2 in appendix 4.1.

Topics models

A constrained topics model was used to reduce dimension of the metagenomics data. It is based on the classification of vaginal microbiota into classes called *community state types* (CST) that represent communities of microbial species. Figure 2.7 displays the allocation of the samples to these classes.

However, topics model aim to summarize the samples as proportions, rather than fixed allocation, of these CST's. The first part of this model focused on the *Lactobacillus*-characterized communities whereas the second part detailed the topics of non-*Lactobacillus*.

***Lactobacillus* topics** The need for a constrained topic model is obvious when one looks at right hand side of figure 2.8 showing the proportion *Lactobacillus* spp. in the different samples. Clearly, many samples contain high proportions of the first four species *L. crispatus*, *L. iners*, *L. jensenii* and *L. gasseri*.

Furthermore, these four species could be associated to four different topics. Indeed, as we can see in the left hand side of figure 2.8, each previously mentioned bacteria is present in high proportion in CST I, II, III, and V. Therefore, it makes sense to redefine CST I as dominated by *L. crispatus*, CST II as dominated by *L. gasseri*, III with *L. iners*, and V with *L. jensenii*.

This new definition of the topics is presented in figure 2.9: topics I, II, III, and V were set to be composed of exclusively *L. crispatus*, *L. gasseri*, *L. iners*, and *L. jensenii* respectively, while topic VI is left to be composed of all other *Lactobacillus*

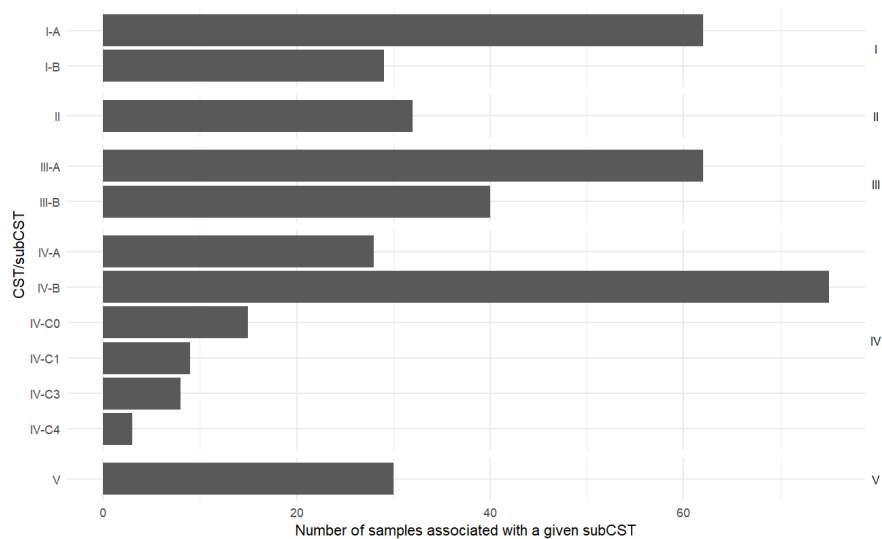


Figure 2.7: Number of samples assigned to each sub-community state type

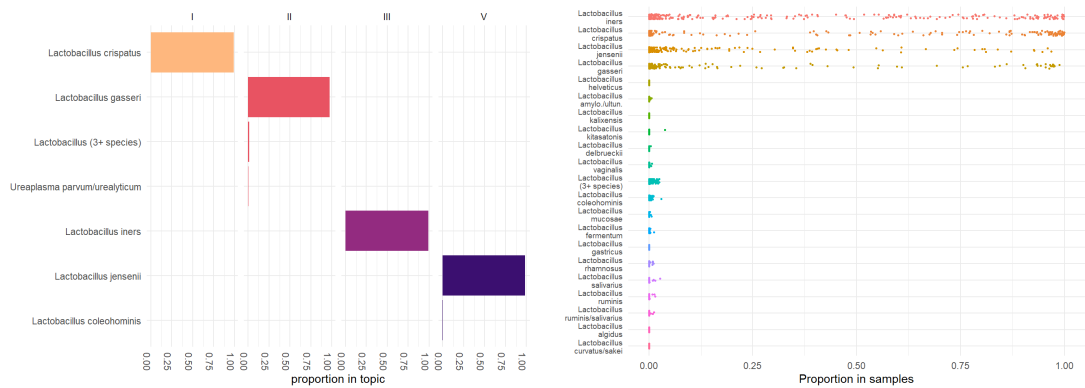


Figure 2.8: (Left) Mean proportion of some *Lactobacillus* species of samples assigned to topic I, II, III and V; (Right) Proportion of *Lactobacillus* spp. in samples

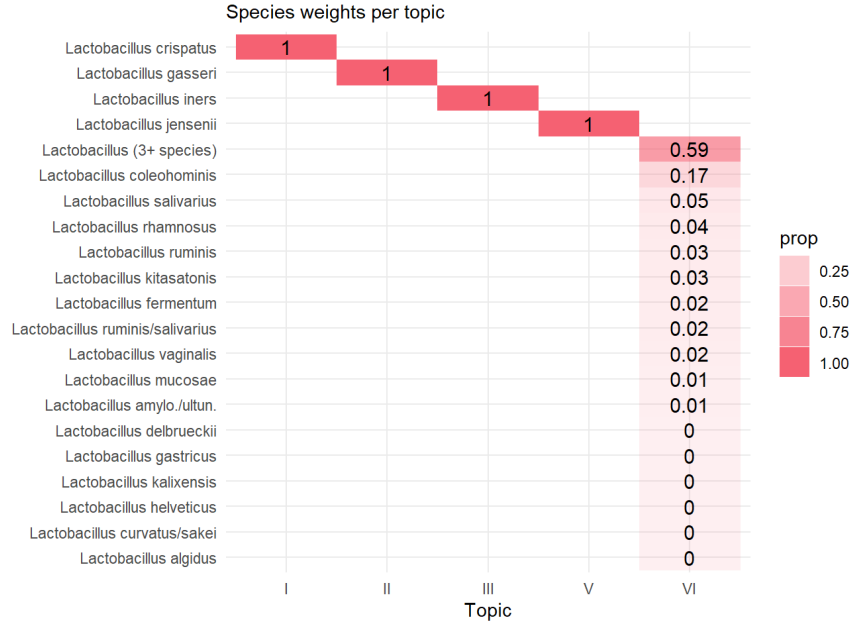


Figure 2.9: Proportion of *Lactobacillus* spp. in the *redefined* topics I, II, III and V

spp. not previously mentioned.

Non-*Lactobacillus* topics After treating the *Lactobacillus*-dominated subcommunities, we focused solely on non-*Lactobacillus* topics. Models with up to 20 topics were fitted.

The transport and product alignment of figure 2.10 help us understand the consistency of the topics. From the transport alignment plot, the brown-green, bright pink and light blue paths are the most consistent. Though it breaks up at model 7, the purple path reassembles itself at model 9 and stays rather constant afterwards. The red path is also very consistent. However, the turquoise path in model 6 looks more like an outlier with no organized connections to the other paths. Based on the transport alignment, 5 topics is probably a good solution.

In the product alignment plot, the distinction is less obvious but the pink and orange paths are consistent up to model 7-8. Though the colors are not the same, it is likely that the brown-green and pink paths from the transport alignment correspond to the orange and pink path of the product alignment respectively. Furthermore, the turquoise and blue paths of the product alignment plot are generally separated from one another. In conclusion, a reasonable number of topics is 4 or 5.

One may also rely on the plots of figure 2.11 to choose the number of topics. In these plots, a plateau appears in the product alignment after 5 topics, and there is

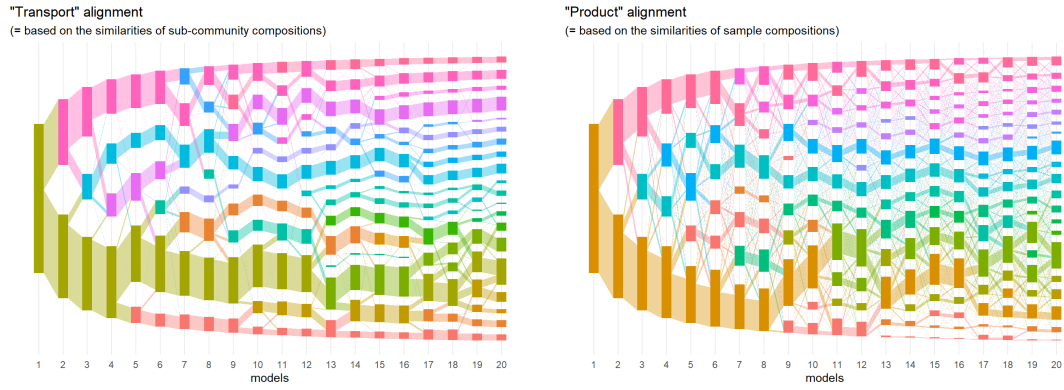


Figure 2.10: Transport (left) and product (right) alignment results from non-*Lactobacillus* topics models 1 through 20

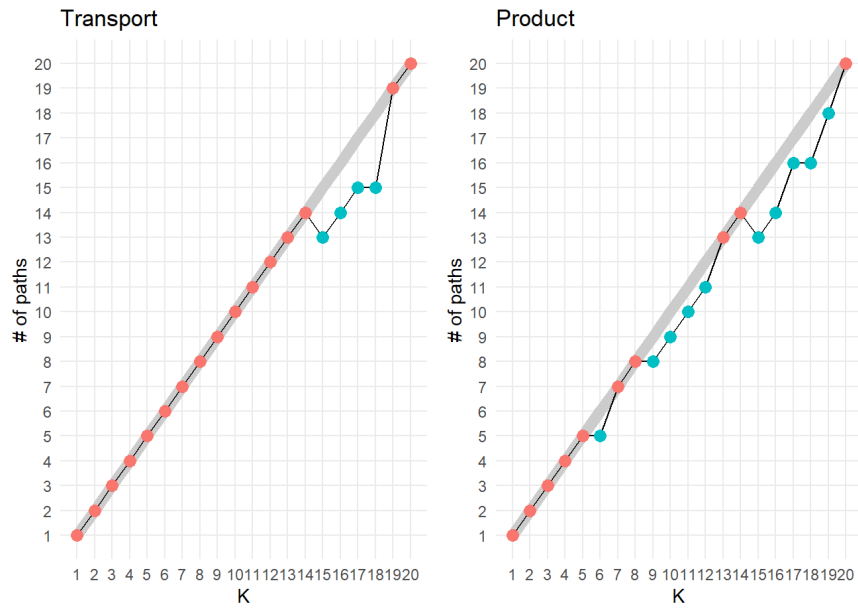


Figure 2.11: Number of path as a function of the number of topics K for non-*Lactobacillus* models using transport (left) and product (right) alignment

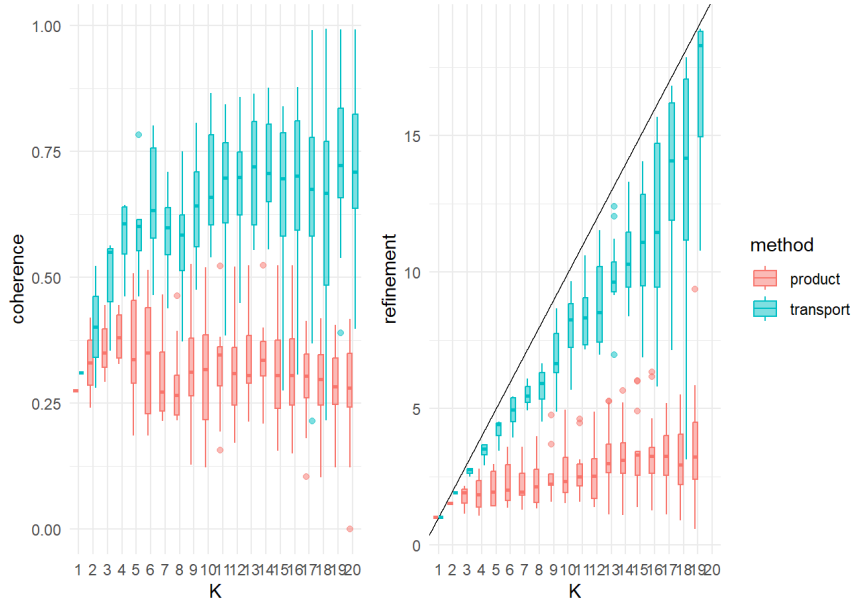


Figure 2.12: Coherence (left) and refinement (right) scores distribution associated to the number of topics K for non-*Lactobacillus* topics models

a second plateau at 8 topics. Then, both alignments reveal a decreasing curve at 14 topics. Therefore, it makes sense to choose a number of non-*Lactobacillus* topics below or equal to 8.

Metrics like the refinement and coherence score of figure 2.12 are also part of the diagnostics to find the number of topics K . At 5 topics, the median coherence score is already high and higher than 8 topics, suggesting 5 topics instead of 8. A plateau in the refinement score is relatively visible early on at 4 or 5 topics in the product alignment. Hence, evidence point at the number of topics $K = 5$ to be relatively efficient and advantageous. This means that these five bacterial subcommunities are both diverse enough to encompass the whole of the identified microbes and restrictive enough to make allegedly biologically distinct communities.

Finally, the beta plot of figure 2.13 shows the relative quantity of bacterial species in each subcommunity when considering either 5 or 8 topics for non-*Lactobacillus*. It would seem like some topics are very similar between 5 and 8 : for instance, the darker green, brownish, pink, and red topics are quite similar. Thus, 8 topics is not as advantageous as 5 is. Hence, the beta plot confirms the choice of 5 topics to be a good compromise between dimension reduction and information loss.

After combining the 5 *Lactobacillus* topics with the 5 non-*Lactobacillus* topics, the beta plot of the 10 topics in total is drawn and printed in figure 2.14.

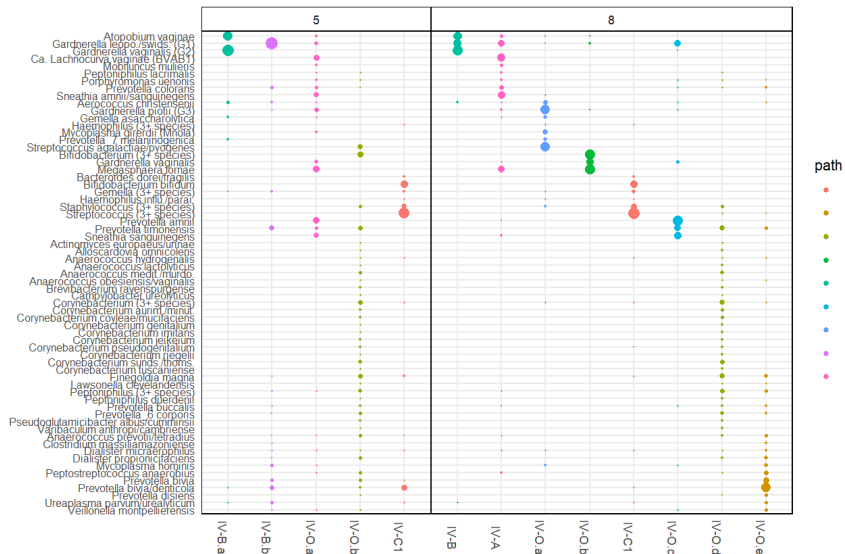


Figure 2.13: Beta plot for non-*Lactobacillus* models with 5 and 8 topics

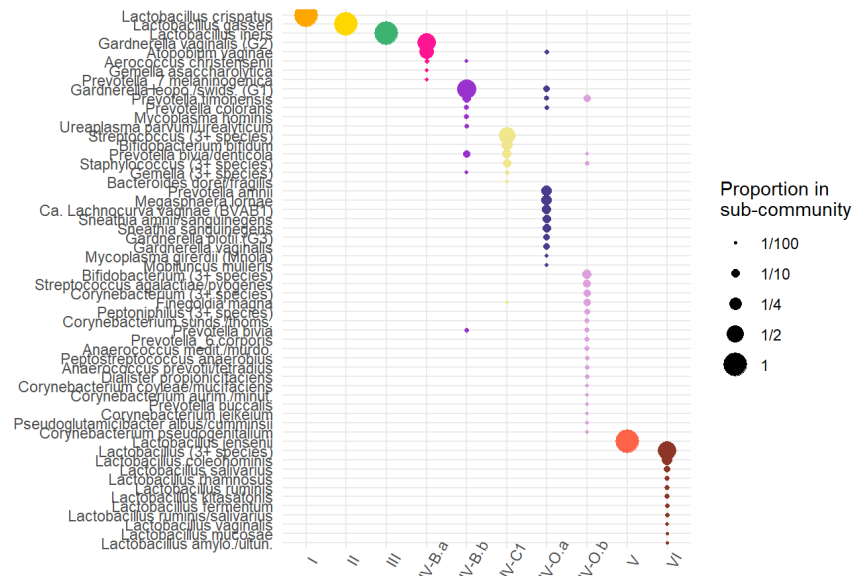


Figure 2.14: Beta plot for combined *Lactobacillus* and non-*Lactobacillus* topics with a total of 10 topics / subcommunities

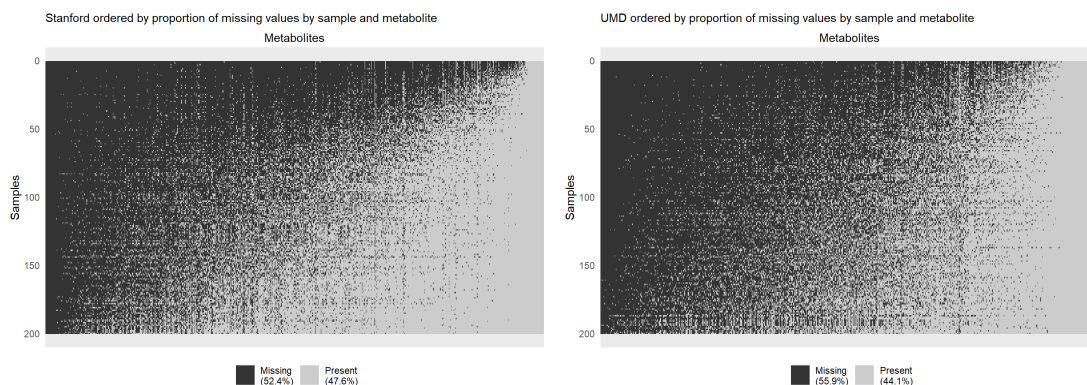


Figure 2.15: Ordered missing values per column and row in the raw metabolomic dataset of Stanford (left) and UMD (right)

2.2.3 Metabolite data preprocessing

Missing values

Figure 2.15 helps visualize and quantify the proportion of missing values in the metabolomics data of Stanford and UMD. It also shows more than half the data is missing: 52% in Stanford and 56% in UMD.

Then, we explored the possibility of metabolites missing due to their biological properties (like their baseline concentration). Figure 2.16 shows a positive correlation (Spearman's $\rho = 0.872$) between the proportion of missing metabolites in both UMD and Stanford dataset. This confirms the missingness is not completely random and likely originates from low concentrations rather than from technical artifacts. Additionally, figure 2.17 presents a negative correlation between the mean abundance of the metabolites and proportion of samples in which a metabolite is missing in both Stanford and UMD datasets (Spearman correlation coefficient of -0.651 and -0.529 respectively). Together these results lead us to conclude that the majority of missing data is linked to the biological abundance of the metabolites: the more abundant a metabolite is, the likelier it is to be detected by the spectrometer, the more probable to have a reading. Therefore, most missing data is not random and depends on the unobserved LLOD.

One way to deal with most of these missing values is to filter out undesired biochemicals. If we were to keep only the metabolites for which the proportion of missing values is greater than a threshold, say 25% for instance, then we would keep only the ones for which we have more information regarding their potential missingness pattern. Now this suggestion is not as proposterous as one would think. Because the variability across samples is vast, as we will see in the next section, we know for a fact that some vaginal swabs were poorly collected or stored, thus

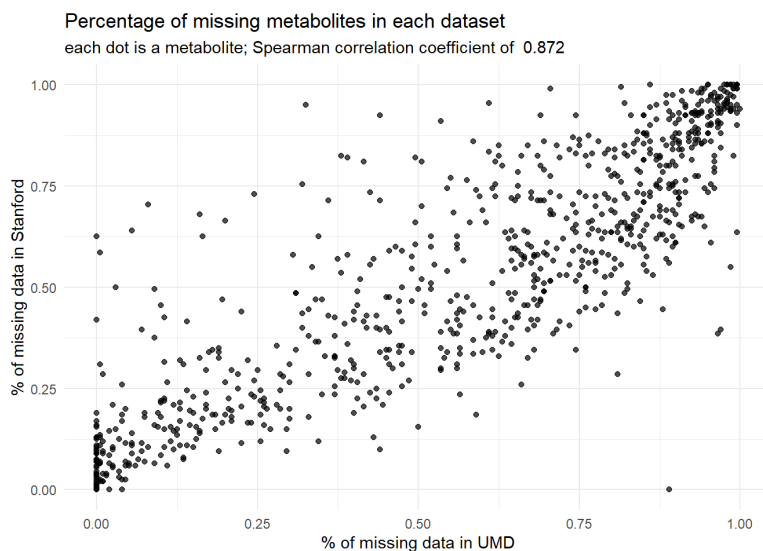


Figure 2.16: Scatterplot of proportion of UMD and Stanford samples containing a given metabolite

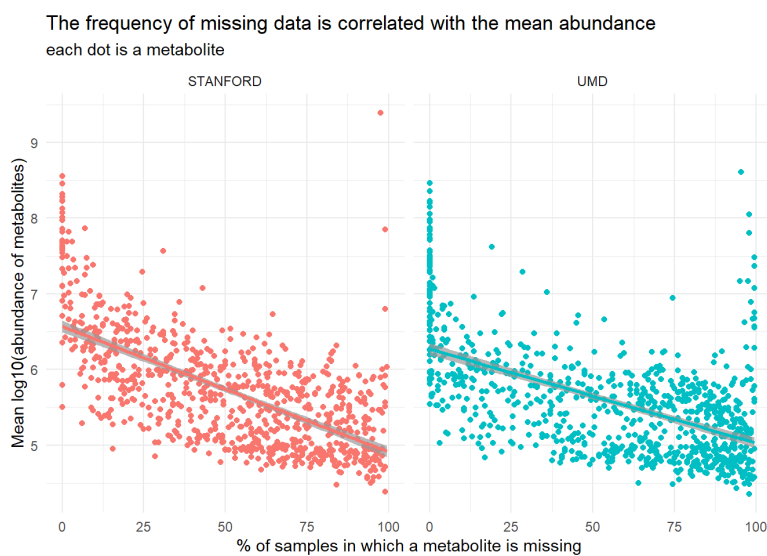


Figure 2.17: Scatter plot of the mean log₁₀ peak intensities of raw metabolomic dataset as a function of the proportion of missing values in Stanford (left) and UMD (right)

resulting in samples of low quality and low biological material quantity. These samples are more likely to contain missing values, like we saw in figure 2.16. Instead of desperately trying to find the best imputation method for these largely missing portions of the data, we discard them. If we had imputed these portions, we would potentially create an unwanted effect that will influence the end results. Implicitly, we select the "best" fragment of the samples and increase our chances to find a relevant biomarker or biomarkers.

Nonetheless, we note that by initiating this filtering process, we could possibly omit a rare discovery. There might be important metabolites of naturally low concentration that we will discard and never know how meaningful they are.

We have to keep in mind that the example of the 25% threshold value for proportion of missingness for a given metabolite is rather arbitrary and different values to balance stringency and loss of information might be as adequate. One could perform a sensitivity analysis and repeat all downstream analyses to estimate the impact of that threshold on the biological conclusions. This was outside the scope of this study but another researcher could use the code developed here to perform such a sensitivity analysis.

Here, we opted for preserving all metabolites with less than 15% of missing values. By doing so, we restrict our dataset to 132 biochemicals for Stanford and 165 for UMD. From this filtering process, there is a 89.39 percent match between the selected metabolites (*i.e.*, the percent match being the number of matching metabolites of the smallest dataset – Stanford in this case – we find in the largest dataset – UMD in this case –)

Further filtering must be done to improve the data quality: even though there is a strong negative correlation (Spearman's $\rho = -0.720$ and -0.545 for Stanford and UMD respectively) between the proportion of missing values in a sample and the mean abundance of the 132 and 165 "most common" metabolites in Stanford and UMD datasets respectively (remember, the "most common" metabolites refer to metabolites with less than 15% of missing values), some samples still comprise more than 50% of missing values, as we can see in figure 2.18. The quality of these samples might be too low and observed values in these samples might not be reliable. Consequently, in addition to filtering out metabolites for downstream analyses, samples with more than 50% of missing values (after metabolite filtering) were discarded from the analysis. The underlying correlation between mean metabolite abundance and proportion of missing values after filtering process is still significant, see table 2.3.

Variable distribution

After filtering features and samples with high proportions of missing values, we next evaluated the effects of different transformation on metabolites' abundance

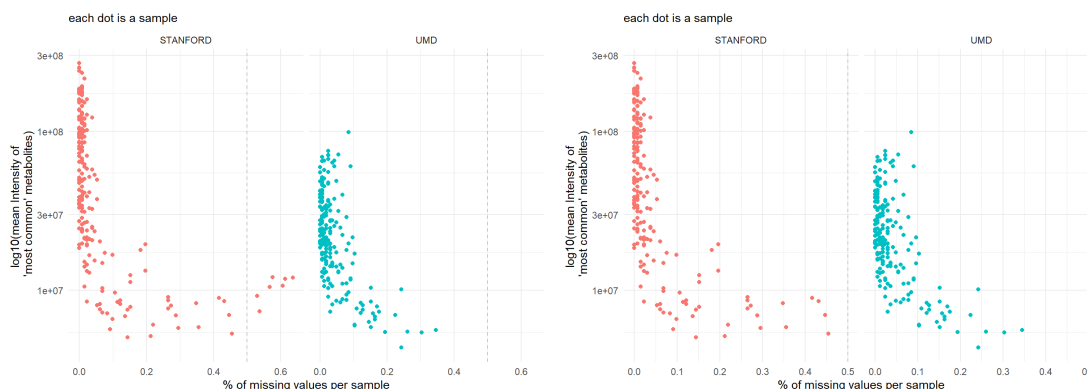


Figure 2.18: Scatterplot of mean metabolite abundance per sample versus proportion of missing values per sample in Stanford (red) and UMD (blue). *Right*: metabolites with missing frequency of 85% or above were filtered out. *Left*: metabolites with missing frequency of 85% or above and samples with missing frequency of 50% were filtered out

	Estimate	P-value
Stanford	-0.697	2.409e-29
UMD	-0.546	6.824e-17

Table 2.3: Spearman’s correlation coefficient estimate and pvalue between mean metabolite intensity and proportion of missing values in each cohort

distributions with the goal of finding a transformation that would alleviate any heteroskedasticity-related issues in downstream analyses.

A first visual inspection of the boxplots in figure 2.19 indicate that the log10-transformed peak intensities might improve variance stabilization. The asinh transformation also appears beneficial in terms of variance. The log10-transformation is very popular in metabolomics and suitable for many other concentration/intensity data. It is particularly adequate in this case as spectrometers usually use the inverse of the log10 (*i.e.*, the $f(x) = 10^x$ function) during measurements. This finding is in agreement with the data collection process of Metabolon’s spectrometry which uses the log scale in the measuring procedure of the samples (Metabolon, a).

After log10-transforming the metabolites’ distribution, using Benjamini-Hochberg’s correction method, there were 38 non-significant Shapiro-Wilk’s normal tests out of 132 for Stanford, and 71 out of 165 for UMD. It meant there is approximately 28.79% tests for which we could not reject the null hypothesis of normality (*i.e.*, H_0 : vector comes from a normal distribution and H_1 : vector does NOT come from a normal distribution) for Stanford, and 43.03% for UMD. If normality was a goal in this analysis, we could not say the log10 transformation really help in normalizing



Figure 2.19: Boxplots of Stanford and UMD metabolite intensity for each sample using no scale (top), log10 scale (middle) and hyperbolic arc-sine scale (bottom)

the data.

From the mean-variance plots in figure 2.20 and 2.21, one can say that a scaling or transformation of the initial datasets is useful. Indeed, the goal was to find a data format for which the metabolite's mean was not or poorly related to its variance. Through the previous graphs of Stanford and UMD cohorts, one can see:

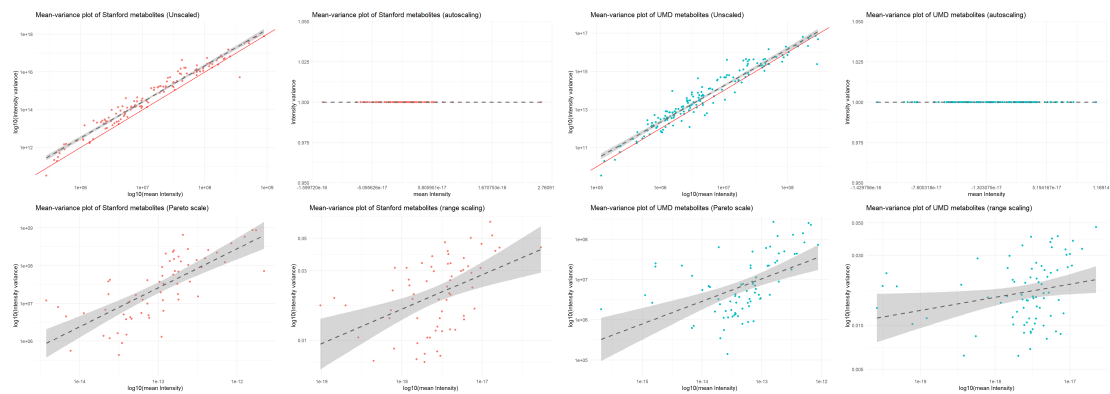


Figure 2.20: Mean-variance plots of metabolites for Stanford (left - red) and UMD (right - blue) results

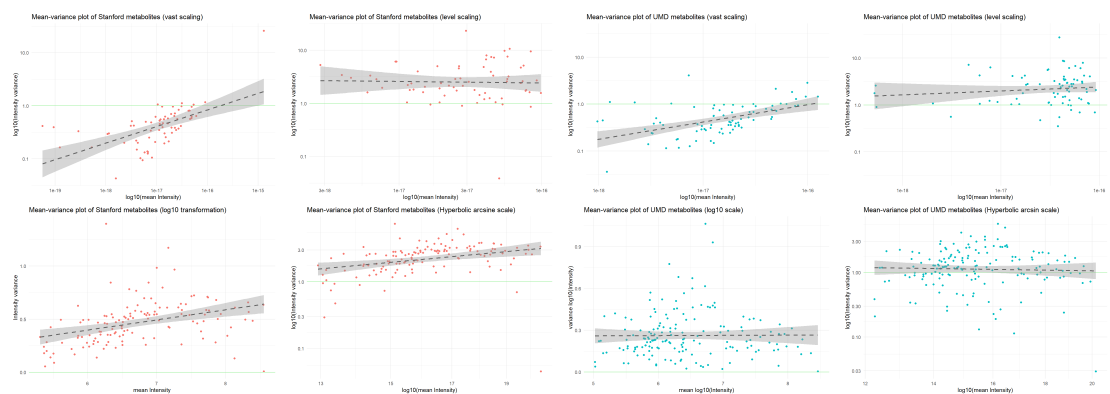


Figure 2.21: Mean-variance plots of metabolites for Stanford (left - red) and UMD (right - blue) results

- The unscaled data has a clear non-null relationship between mean and variance: the red line on the graph is the $y = 2x$ function, thus indicating the mean-variance followed a similar function with another intercept.
- The autoscaling completely resolves the mean-variance problem as metabolites gather on $y = 1$ curve (because dividing all columns by their respective standard deviation sets their respective variance to 1). However, two important drawbacks from this standardization we need to consider is that 1) it assumes the variables follow a normal distribution, and 2) it is sensitive to outliers. Since we cannot safely confirm the first assumption and saw that some metabolite values were very large, it is unreasonable to prize this scaling.
- Pareto scaling, range scaling and vast scaling do not really help overall: the relationship between mean and variance is still unquestionably positive.
- Level scaling helps decrease the correlation between mean and variance since the dashed line is almost horizontal.
- $\log_{10}$ and hyperbolic arc-sine transformations do help stabilize the variance across metabolites. Furthermore, they help decrease mean-variance correlation in UMD cohort but not Stanford.

Overall, the level scaling and both $\log_{10}$ and asinh transformations showed partial problem resolution. However, the $\log_{10}$ is probably the one with the most benefits in terms of variance stabilisation, interpretability and mean-variance relationship. Therefore, it will be used in the later sections of this project.

Principal Component Analysis

Once data were transformed to attenuate heteroskedasticity and skewness, we performed a principal component analysis to identify potential latent low-dimensional structure in the data.

In both PCAs (PCAs of Stanford's and UMD's datasets), the first component is much larger than the remaining ones (58% in Stanford and 41.73% in UMD), see figure 2.22. This could potentially be an effect introduced during data collection or measurement. It will be investigated further and handled in a subsequent section.

In general, when exploring the scores plot, one looks for structure among samples. In our case, one can speculate a division between the samples of either batches for Stanford. Moreover, since 5 samples were collected longitudinally for each subject, one could expect small clusters of individual points to correlate with within subject variance. Both of these effects are confirmed. Figure 2.23 shows how principal components 3 and 4 discriminate between batch 1 and batch 2 samples.

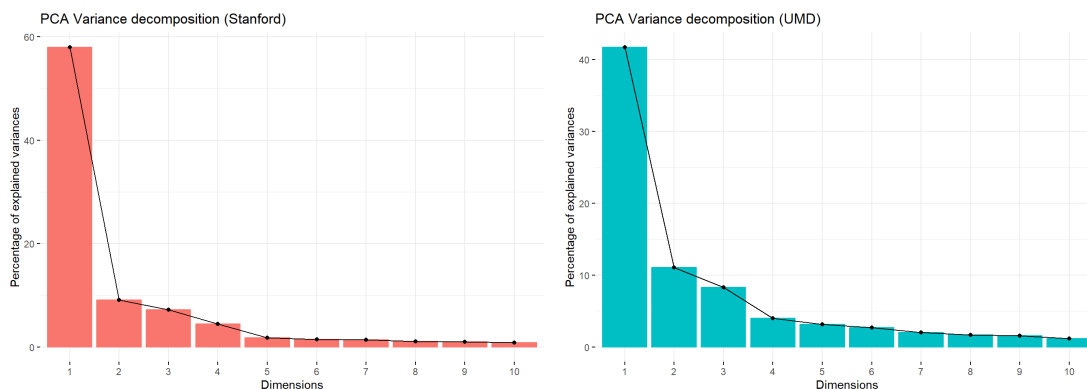


Figure 2.22: Barplot of percentage of variance explained by the first 10 principal components of Stanford (left) and UMD (right) PCA on filtered, imputed, transformed dataset

Figure 2.24 expresses how some samples from the same subject tend to form small clusters and others don't. This translates a subject effect in the data.

The previously mentioned size effect visually manifests itself by the negative correlation of all metabolites with the first principal component in the loadings'plot of figure 2.25.

Additional tables and figures of this PCA are available in appendix 4.3

Batch analysis

Stanford data was analysed in two different batches (batch 1 and batch 2). A comparison between density functions of attributes according to batch was performed. All following analysis were made on the filtered and transformed data as described in previous sections. After filtering, there were 79 Stanford samples in batch 1 and 114 Stanford samples in batch 2. The subject samples were not mixed, meaning all samples from a given patient belonged either to Batch 1 or Batch 2.

Figure 2.26 shows a slight difference between the metabolite intensity density of the two sub-cohorts, suggesting a difference in mean abundance. This observation is supported by Mann-Whitney (MW) and Kolmogorov-Smirnov's (KS) non parametric tests of equal distribution, with pvalue equal to $1.717 \cdot 10^{-6}$ for MW and $1.439 \cdot 10^{-5}$ for KS. Such low p-values suggest rejecting the null hypothesis of matching distributions. Therefore, one cannot assume the sub-cohorts have the same metabolite abundance distribution.

A visual inspection of figure 2.27 does not support differential distributions of gestational age or participant's age by batch. Through table 2.4, we cannot say the race is balanced out between the batches. Additionally, with a χ^2 test pvalue of 0.015, we reject the null hypothesis of proportional categories. We conclude there



Figure 2.23: PCA scores plot of Stanford (left) and UMD (right) colored by batch; data input was filtered, imputed, and transformed

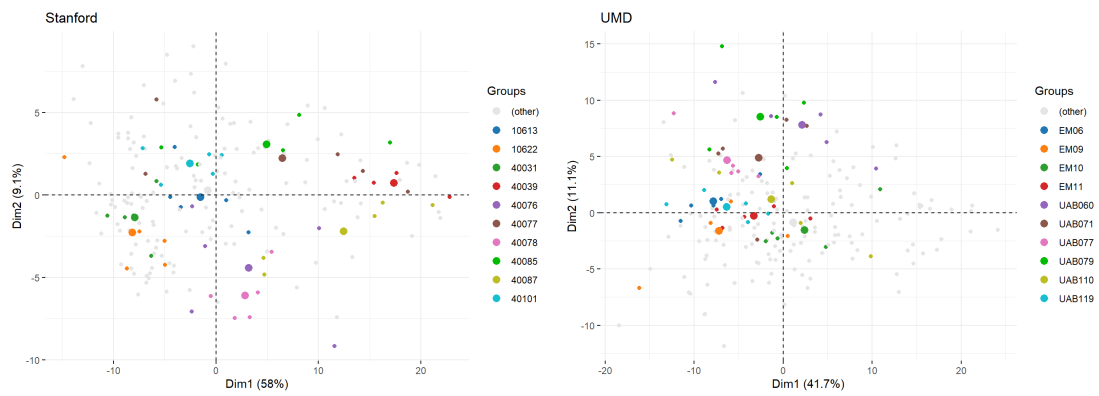


Figure 2.24: PCA scores plot of Stanford (left) and UMD (right) colored according to 10 random subjects; data input was filtered, imputed, and transformed

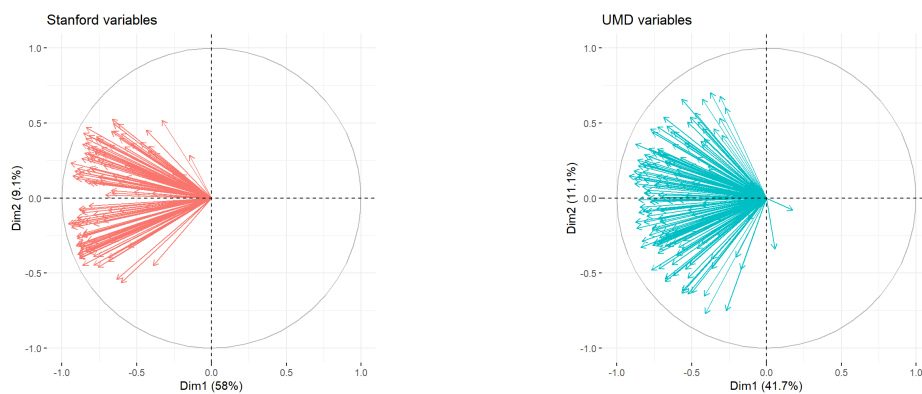


Figure 2.25: PCA loadings plot of Stanford (left) and UMD (right); data input was filtered, imputed, and transformed

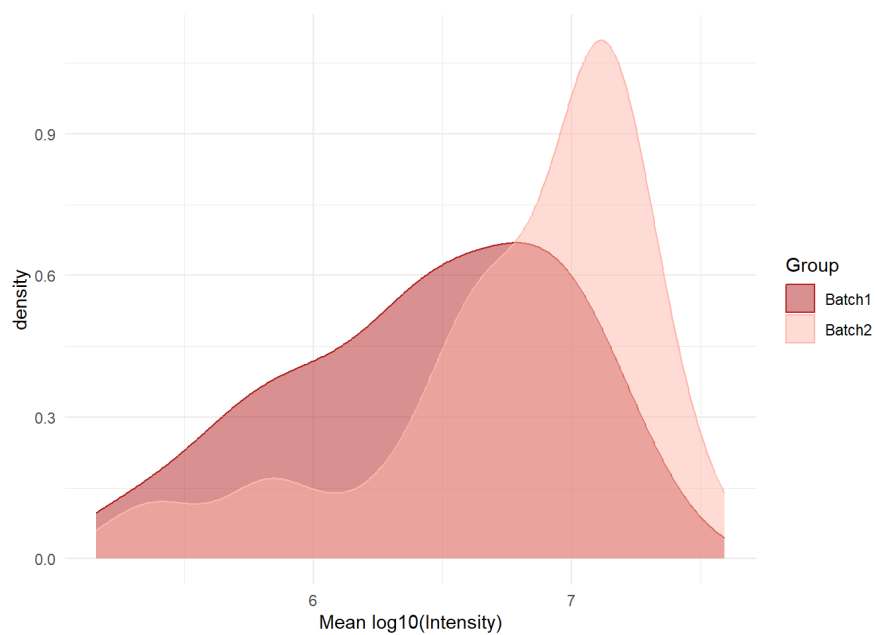


Figure 2.26: Metabolite intensity density function of batch 1 and 2 (density plot was drawn using a bandwidth of approximately 0.17)

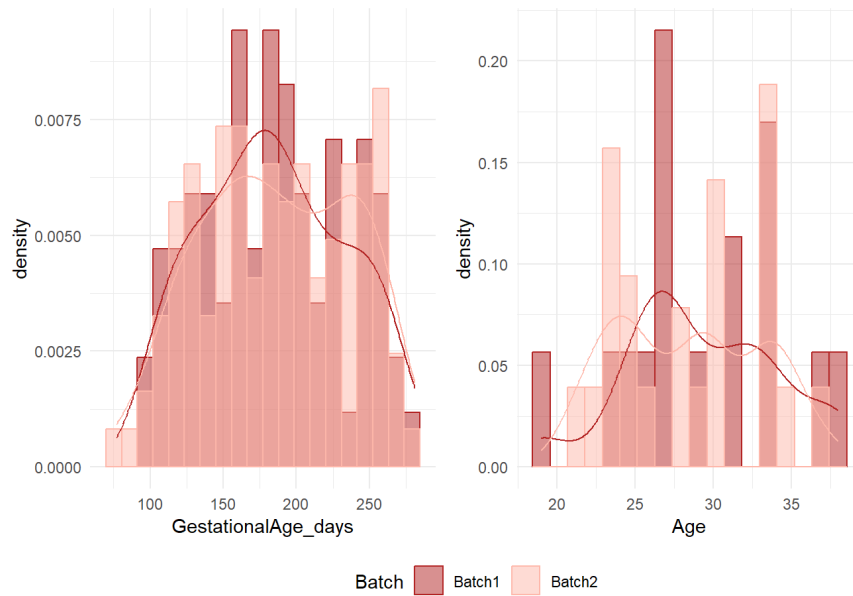


Figure 2.27: Gestational age and subject age density functions of batch 1 and 2

is a difference of race proportions across batches.

Group	Asian	Black	Other	White
Batch1	10 (67%)	34 (31%)	15 (50%)	20 (50%)
Batch2	5 (33%)	74 (69%)	15 (50%)	20 (50%)

Table 2.4: Contingency table of samples in four different self-proclaimed races (Asian, Black, White, and Other) according to batch. The numbers in each cell correspond to the following format "number of samples (proportion of samples)"

In conclusion, the distribution of metabolites mean abundance and the race can be considered as confounding with the batch effect. Ideally, one should take into account these variables while correcting the batch effect.

Batch correction

Let us compute the linear model before and after batch effect correction using the `lmFit()` function and examine how the different covariates "impact" the distribution of metabolites:

To interpret these tables 2.5, one has to know they were initially designed for gene expression datasets. The rows correspond to three categories of gene regulation: “down regulated”, “Not significantly regulated” and “up regulated”.

	Intercept	Batch2	Age	RaceBlack	RaceWhite	GestationalAge_days	Prop_Lacto	
Down	0	4	0	0	0	0	0	9
NotSig	0	26	132	132	132	132	132	45
Up	132	102	0	0	0	0	0	78
	Intercept	Batch2	Age	RaceBlack	RaceWhite	GestationalAge_days	Prop_Lacto	
Down	0	0	0	0	0	0	0	9
NotSig	0	132	132	132	132	132	132	45
Up	132	0	0	0	0	0	0	78

Table 2.5: Regulation tables of the linear models adjusted before batch correction (above) and after correction (below)

In the case of metabolomics data, one could interpret those as “low abundance”, “mean abundance” and “high abundance” respectively. The numbers in the cells correspond to the number of features (metabolites) that fall in each regulation category. Therefore, the table informs us on the variability induced by each variable of the design. If the variable introduces low variability, then one would expect all biochemicals to be in one single “abundance” category for that variable (either **down**, **NotSig** or **up**). Instead, if the variable is correlated to the abundance of the metabolites, then one could expect more variability and metabolites spread in two or more categories. To construct the “regulation tables”, I used three functions from limma package. The first one is `lmFit()` and, as explained previously, it fits a linear model for each feature (each metabolite). Then, I used `eBayes()` function which, given a linear model from `lmFit`, computes t-statistics, moderated F-statistics, and log-odds of differential expression by empirical Bayes method. The results of this function can be interpreted by the function `decideTests()` that can identify which metabolites are significantly differentially ‘expressed’ for each contrast from a fit object containing p-values and test statistics. This function also adjust for multiplicity testing by default, using Benjamini-Hochberg method, and a significance level of 0.05. Therefore, the combination of these function really test the significance of the linear coefficients for the fitted model and takes into account the sign of this coefficient if it is significant.

For example, according to the above table in figure 2.5, the Age effect is seemingly null in Stanford cohort since all biochemicals fall in the same category: the linear model did not detect much variability in peak intensity across the range of age values. This conclusion is quite different from the previous section 2.1.7 on correlated attributes which showed us a significant correlation between age and peak intensity.

The linear model also shows a batch effect: metabolites are partitioned across the three regulation categories, thus indicating variability due to batch. Similar partitioning exists for the lactobacilli proportion variable. However, the gestational age and the race do not seem to induce much variability.

After batch effect removal, the variability linked to the batch seems to be low

or absent according to the tables 2.5.

Size effect

The total amount of biological material collected on the swabs was highly variable from one sample to another and led to the identification of size effect. Unfortunately, we do not have the tools (*i.e.*, the metabolite true concentration) to determine the actual metabolite abundance in each sample. Besides, it is not possible to discriminate between potential biological reasons for increased overall metabolite concentrations and technical reasons (*i.e.*, swabbing quality, sample preservation, etc.).

To gain information on this size effect, we established the significance of the correlation between mean metabolite intensity and host attributes (age, BMI, reproductive, status, etc.), and corrected this size effect by cancelling the first principal component of the metabolite dataset (because the first component captures the greatest variability in the data).

All correlations' estimates and associated p-value (from a nullity test) are gathered in table 2.6 and can be visualized in appendix 4.2. Furthermore, ANOVA tests results for the categorical variables "Reprod_status" and "cycle_nb" are presented in table 2.7 and cannot indicate a significant difference of mean log10 metabolite abundance between the levels of the categorical variables. The biggest Spearman correlation coefficient in table 2.6, in absolute value, is 0.32 and corresponds to the trend of *Lactobacillus* proportion we have already identified. This suggests a small positive correlation between the mean intensity of metabolites and proportion of *Lactobacillus* spp. in the Stanford cohort. In addition, the correlation between mean intensity and age in Stanford, and correlation between mean and pH in UMD are significantly (almost significantly) different than 0. In this case, one needs to control for these effects or to interpret results cautiously.

	Estimate		Pvalue	
	Stanford	UMD	Stanford	UMD
Age	0.2161	0.1799	0.002544	0.01079
BMI	-0.1832		0.01186	
GestationalAge_days	-0.1079		0.1353	
Prop_Lacto	0.3281	0.276	4.094e-06	8.106e-05
pH		-0.1545		0.03234
cycle_length		0.06509		0.5492
cycleday		-0.06045		0.4895

Table 2.6: Spearman's correlation coefficient and associated p-value between the mean intensity of metabolites and the different sample attributes

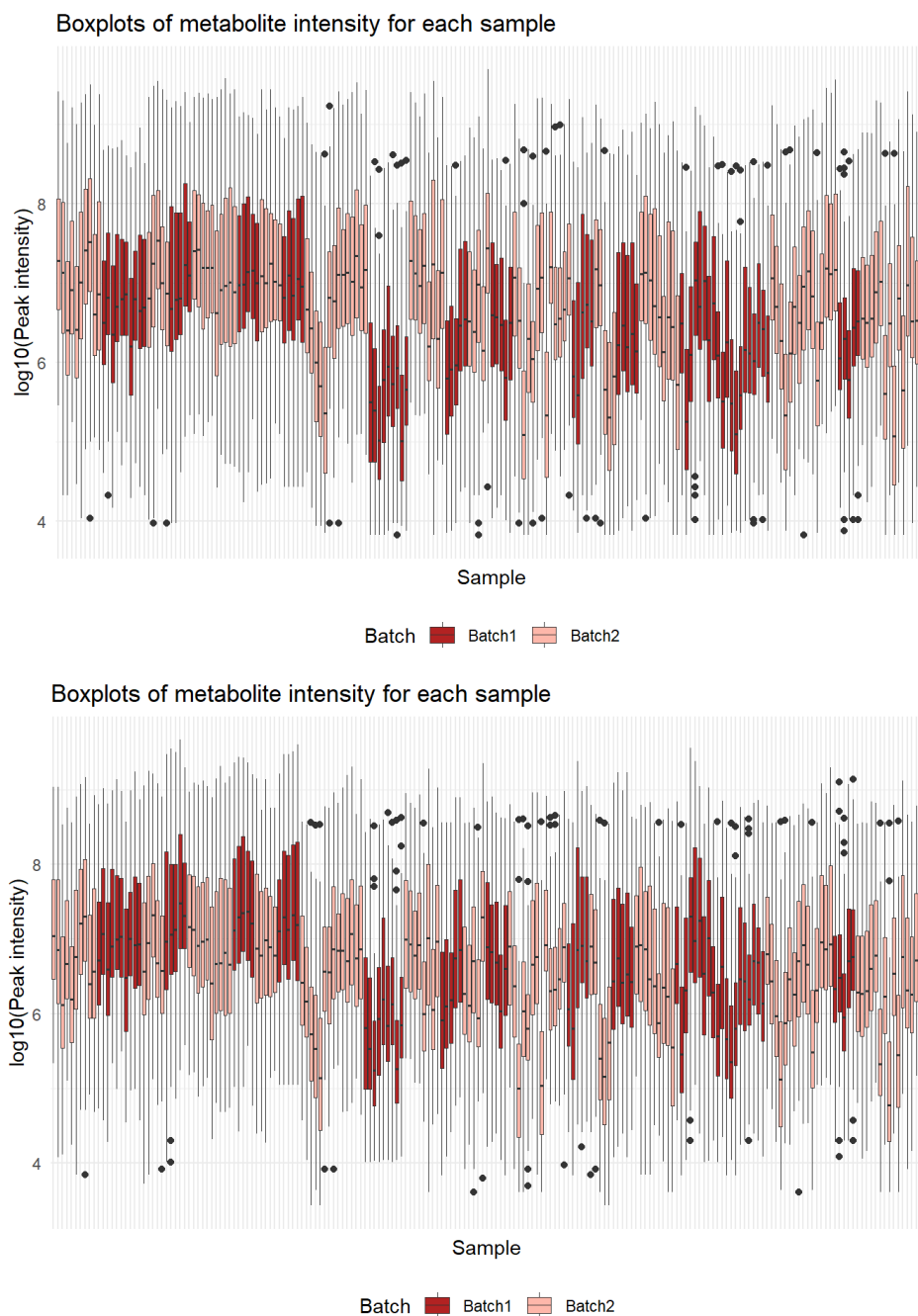


Figure 2.28: Metabolite boxplot distribution before batch correction (above) and after correction (below)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Reprod_status	4	42816	10704	3.529	0.008495
cycle_nb	1	87.01	87.01	0.02869	0.8657
Residuals	174	527715	3033		

Table 2.7: UMD cohort only; one-way ANOVA test between mean metabolite intensity and categorical variables `Reprod_status` and `cycle_nb`

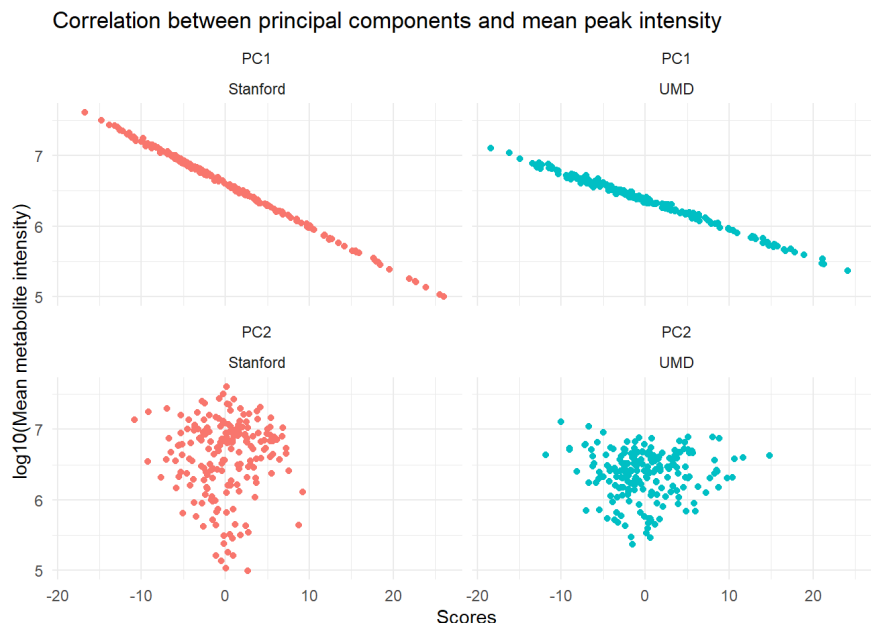


Figure 2.29: Scatterplot of mean metabolite intensity per sample as a function of corresponding principal component' score in Stanford (left) and UMD (right) cohort

Furthermore, figure 2.25 in the previous section 2.2.3 revealed that the first principal component captured the greatest variability in the metabolite dataset, which was indeed the manifestation of the size effect. In figure 2.29, we plotted the metabolite intensity as a function of the first PC' scores and quantified the correlation between these quantities in table 2.8. The graph and table confirm the great correlation between metabolite intensity and the first PC.

After correcting the size effect by cancelling the first principal component, one can plot the loading's plot for the first two dimensions and see how it differs from the previous graph 2.25. From figure 2.30 we can conclude the size has been properly removed. Additional figures from the PCA results can be visualized in appendix 4.4.

	Stanford	UMD
PC1	-0.999	-0.9946
PC2	0.04305	0.0003585

Table 2.8: Spearman’s correlation coefficient between mean metabolite $\log_{10}(\text{intensity})$ and principal component’ score

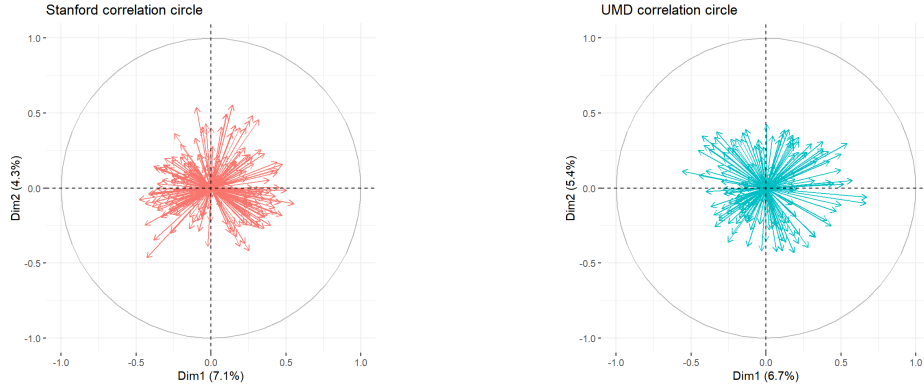


Figure 2.30: PCA loading’s plot after size effect correction in Stanford (left) and UMD (right) using dimension 1 and 2

Lactate abundance vs *Lactobacillus* proportion

Data status	Estimate		Pvalue	
	Stanford	UMD	Stanford	UMD
Filtered	0.436	0.589	7.1e-10	0
Filtered and log10 transformed	0.436	0.589	7.1e-10	0
Filtered, imputed, log10 transformed	0.459	0.589	2.46e-11	0
Filtered, imputed, log10 transformed, batch effect removed	0.585		6.26e-19	
Filtered, imputed, log10 transformed, (batch effect removed), size effect removed	0.218	0.287	0.00255	4.13e-05

Table 2.9: Correlation tests (estimates and p-values) between proportion of *Lactobacillus* and lactate intensity per sample

What we can conclude on figure 2.31 and table 2.9 is that, on the one hand, there is a significant positive correlation between the abundance of lactate and the proportion of *Lactobacillus*. This correlation is expected since species from *Lactobacillus* genus are well known for their lactic acid production. Thus, the

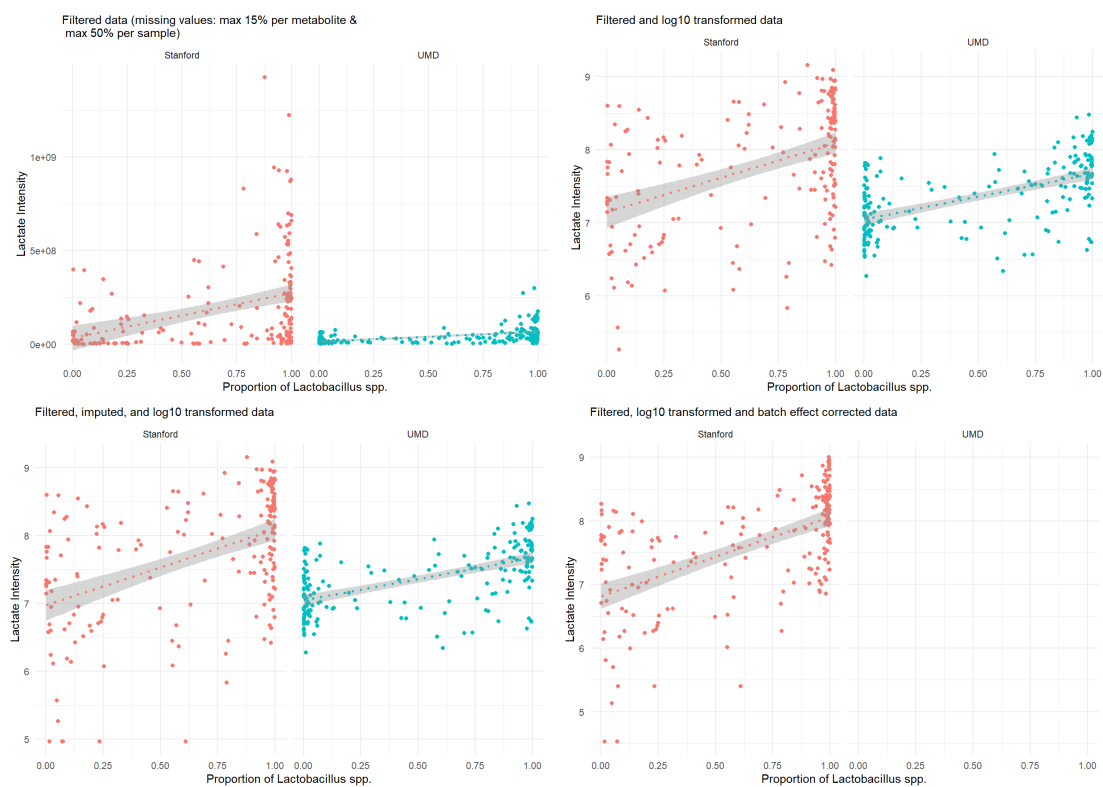


Figure 2.31: Lactate intensity as a function of *Lactobacillus* spp. proportion per sample at the different steps of the preprocessing: filtered data (top-left), transformed (top-right), imputed (bottom-left), batch effect removed (bottom-right)

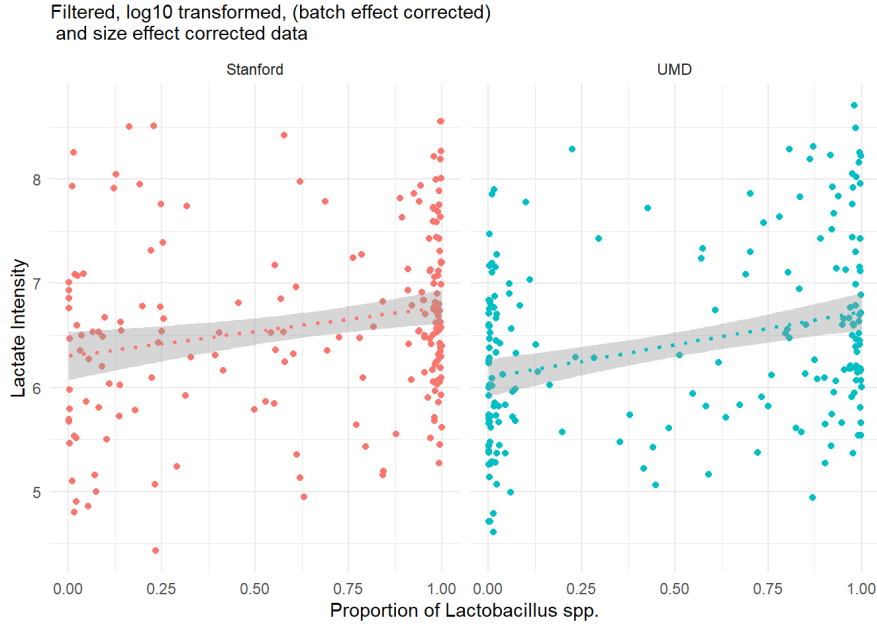


Figure 2.32: Lactate intensity as a function of *Lactobacillus* spp. proportion per sample after correcting size effect

higher the proportion of lactobacilli, the more producing bacteria, the higher the production. On the other hand, the correlation is identical whether the log10 transformation is applied or not.

Furthermore, the imputation did not alter the correlation too much, which is evidence that the method used was efficient. Then, if the batch effect is removed, the correlation changes slightly and increases. Finally, when the size effect is removed, Stanford's correlation becomes non-significant and UMD's correlation decreases a little (figure 2.32). This means that, when correcting the size effect, we also removed some correlation between the metabolite abundance and the proportion of bacterias. However, as mentioned in section 2.1.7, we want to keep this previous source of variation.

If we were to keep the size effect corrected data, then we would loose signals that is in our interest. This is especially visible for Stanford, and we extrapolate the same concept to UMD's results. Therefore, in the future, we must not use the size effect corrected data, but instead, rely on the filtered-imputed-transformed-batch-effect-removed data to increase our chances of detecting important signals.

2.2.4 Metabolite and compositional data integration

The following results pertain to the preprocessed data of Stanford and UMD cohort. After filtering the samples in order to have a match between the metabolite and compositional data blocks, 191 samples remained in Stanford and 200 in UMD. In terms of dimensions, there are 132 metabolites in Stanford and 165 in UMD. Since the compositional data was processed using the combined data, there are 10 subcommunities in Stanford and UMD topics block.

RV coefficient

The RV coefficient for Stanford and UMD data were computed and tested. The results summarized in table 2.10 show that both RV coefficients are significantly greater than 0 ($H_0 : RV = 0$ vs $H_1 : RV > 0$):

Cohort	Estimate	P-value
Stanford	0.1855	2e-04
UMD	0.1823	2e-04

Table 2.10: RV coefficient estimates between metabolite and subcommunity data blocks and associated pvalue using Monte Carlo test for both UMD and Stanford

STATIS analysis results

The first step of STATIS analysis was to check whether or not the data blocks were given the same weight $\left[\frac{|W_k^{(1)}|}{\sqrt{\Sigma_{kk}}}\right]_{k=1,2}$. The different components are summarized in table 2.11 which shows how the different weights for all blocks were computed. It is clear that weighting is greater for metabolite blocks than subcommunity blocks. This large difference of block weights is due to the number of variables in datasets: sets with more variables have greater importance according to STATIS approach.

	Stanford		UMD	
	Metabolite	Subcommunity	Metabolite	Subcommunity
$ W^{(1)} $	0.7071	0.7071	0.7071	0.7071
$\sqrt{\Sigma}$	0.006362	0.6645	0.003258	0.691
$\frac{ W^{(1)} }{\sqrt{\Sigma}}$	111.2	1.064	217	1.023

Table 2.11: STATIS block weights for Stanford and UMD

The closeness between blocks was visually checked using the interstructure plot of figure 2.33. Both table vectors are highly correlated with the first dimension,

indicating a certain proximity or "similarity" between the two tables. Moreover, the magnitude of the vectors (their norm) is 1, meaning they both contribute 100% to the compromise. This value of 100% is not accidental: when the number of data blocks to compare is equal to 2, the dimension of $\mathcal{S}$ is 2×2 and therefore there can only be a maximum of 2 components. Hence, the total variance is captured by the first and only two components.

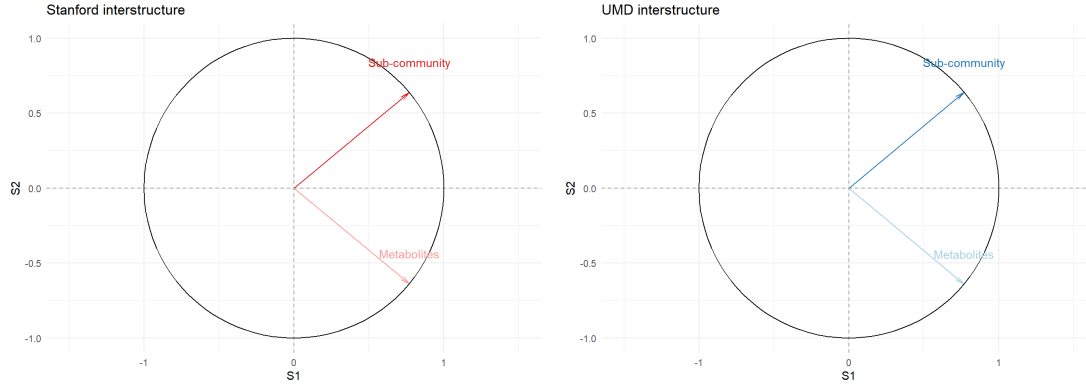


Figure 2.33: STATIS' interstructure plot of the first and second dimensions for Stanford (left) and UMD (right)

When looking at the variance decomposition of the compromise subspace in figure 2.34, the first principal component explains 24.5% of the variance in Stanford and 20.1% in UMD. The first two components capture 36.5% of the variance in Stanford and 33.1% in UMD. Furthermore, we require 7 components to reach 75% of variance explained in Stanford, but 9 components in UMD. Both Stanford and UMD's plots indicate the importance of the first dozen of dimensions in the explanation of the variance. Coincidentally, this number of important principal components is similar to the number of subcommunities (which was 10 in total).

Figure 2.35 shows the sample's relative position in the first four compromise dimensions for Stanford and UMD. Though the point cloud in dimensions 1 and 2 is rather homogenous, for both Stanford and UMD, this cloud forms a rough triangle in dimension 3 and 4. Moreover, the samples' position is linked to their inherent proportion of *Lactobacillus*. In Stanford and UMD compromise plot of dimension 1 and 2, there is a gradient of samples based on this lactobacilli proportion. In dimensions 3 and 4 of the same cohorts, the samples with great proportions of lactobacilli are more or less on in the outer regions of the triangle, and samples with lower lactobacilli proportions are centered within the inner regions.

The origin of these gradients is linked to the compositional data: the proportions of microbial species and subcommunities prevent hard-classification of samples and allow for smoother transition from one subcommunity to another. The fact

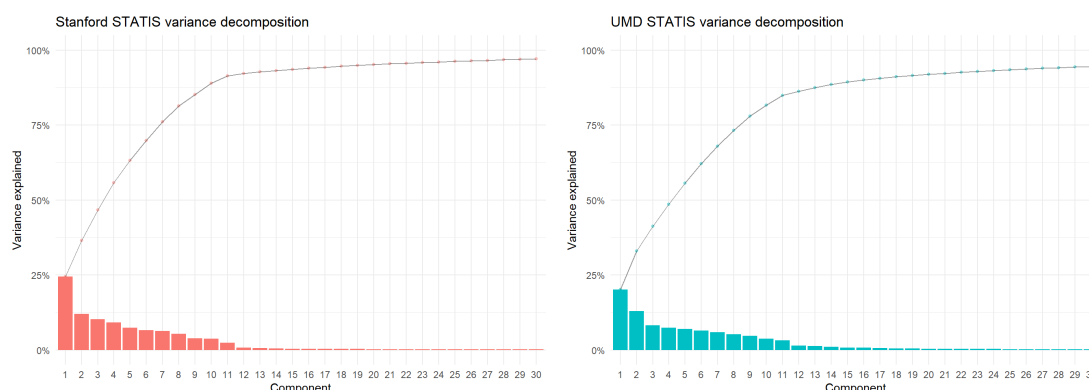


Figure 2.34: STATIS' variance decomposition barplot of the first 30 dimensions out of 141 for Stanford (left) and out of 174 for UMD (right)

the triangles don't have sharp edges and pointy corners is because the STATIS found a compromise between the samples' position through the metabolite dataset (probably a cloud of points) and the samples' position through the composition dataset (*i.e.*, a simplex).

Finally and most importantly, one can look at the feature's coordinates (the loadings) in figure 2.36. A focus on a list of metabolites has been made using colors: lactate, hippurate, cysteine, spermidine, putrescine, and tyramine. The essential elements to take away from these graphs are the following:
Stanford:

- Metabolites: metabolites are all negatively correlated to the first compromise axis, just like they were in the descriptive analysis' PCA. This phenomenon happens because the size effect hasn't been removed. Phosphate is the only first dimensions positively correlated metabolite. Tyramine is positively correlated to the fourth principal component and adenosine is oppositely negatively correlated to that same component. Hippurate is negatively correlated to the third PC.
- Subcommunity: In the first two dimensions, *Lactobacillus*-dominated subcommunities I, II, V, and VI form a group of vectors opposite of non-*Lactobacillus*-dominated subcommunities IV-B.b and IV-O.a. This opposition explains the gradient of samples in the scores' plot of figure 2.35. Furthermore, subcommunity III is opposite of IV-O.b in all shown dimensions. Because of the right angle between the two vectors, subcommunity I and IV-O.a look independent of both III and IV-O.b. In dimensions 3 and 4, the contribution of subcommunities to the axis is dominated by subcommunities I, II, and III.
- Both: Lactate, as expected, is negatively correlated with the first axis, along

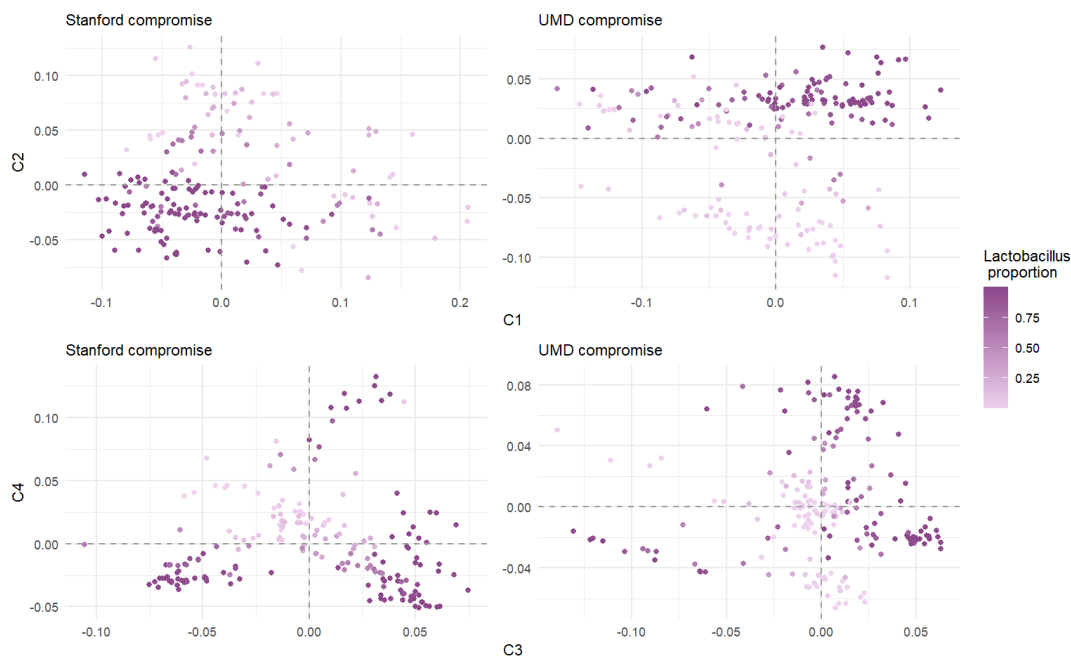


Figure 2.35: STATIS' row coordinates on the compromise' first 4 dimensions for Stanford (left) and UMD (right), colored by the proportion of *Lactobacillus* spp.

with all *Lactobacillus*-dominated subcommunities. Hippurate is positively associated with subcommunities I, II, III, V, VI. Tyramine and putrescine are correlated to the same dimensions as subcommunities IV-B.b and IV-B.a. Tyramine could be associated with subcommunity II. Adenosine is associated with I and III. Phosphate is associated with subcommunity IV-O.b in dimension 1 and 2. Phosphate is associated with IV-O.b.

UMD:

- Metabolites: Phosphate and maleate are negatively correlated to the first principal component. Putrescine, tyramine, xanthine, and spermidine are negatively correlated to the second PC. Tyramine and putrescine are correlated to one another in all dimensions. Hippurate, asparagine, lactate, and cysteine are positively correlated to the second PC. Hippurate is also strongly correlated to orotate in dimension 3 and 4, and they oppose 1-carboxyethylphenylalanine.
- Subcommunity: the lactobacilli subcommunities are grouped together in dimensions 1 and 2: two subgroups composed of subcommunities I, II, and VI, and III and V. In dimensions 1 and 2, IV-O.a and IV-B.b are correlated to each other and might be independent of IV-O.b. The third and fourth

dimensions do not have as distinctive groups of variables as the first two dimensions did. Instead, they hold the key to understanding the triangular shape of the scores plot of figure 2.35: the subcommunities I, II, and III are the corners of this triangle, and all non-lactobacilli subcommunities are "contained" in that triangle.

- Both: Lactate is associated with all lactobacilli subcommunities. 1-carboxyethyl-phenylalaline is associated with subcommunity III. Hippurate is probably correlated to I, along with orotate. Putrescine is associated with IV-B.a, while tyramine is associated with IV-B.b. Phosphate is associated with IV-O.b.

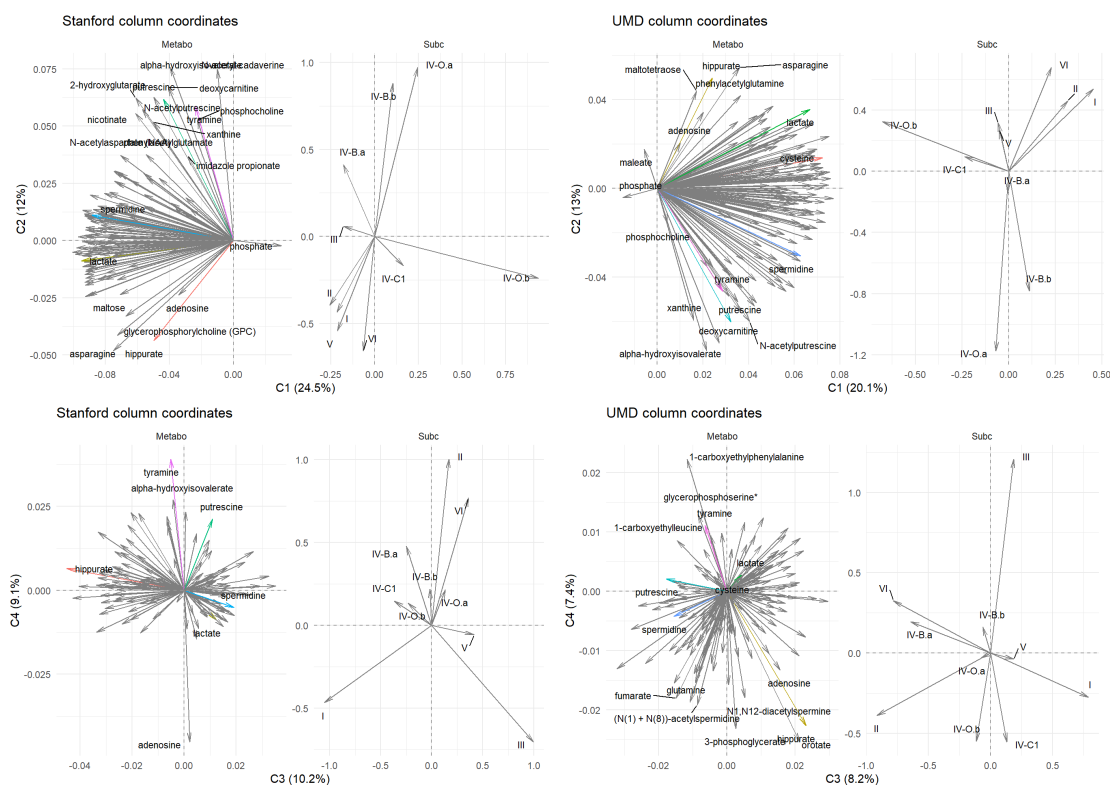


Figure 2.36: STATIS' column coordinates (loadings) on the compromise' first 4 dimensions for Stanford (left) and UMD (right)

SCA results

The graphs to explore in Simultaneous Component analysis are the same as STATIS (*i.e.*, variance decomposition barplot, scores plot, loadings plot), but the results might differ due to both the weights and the PCA rotation. Indeed, the SCA

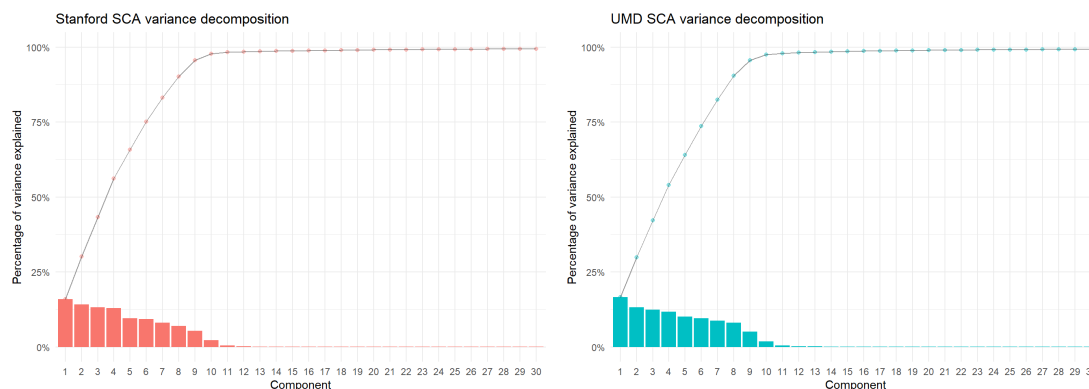


Figure 2.37: SCA' variance decomposition barplot of the first 10 dimensions for Stanford (left) and UMD (right)

procedure is the same as STATIS' in terms of computation and input tables except the block weights are all equal to 1. Moreover, a rotation or coordinate flip might occur when computing the PCA.

First, looking at the barplot of figure 2.37, we see the decomposition of the variance through the principal components has managed to gather 15.9% of the variance in Stanford and 16.6% in UMD. Moreover, it requires 6 components to reach 75% of variance explained in Stanford and 7 components in UMD. Additionally, the first 10 components explain a great deal of the variance before an elbow occurs in the cumulative variance.

The scores plot of the individuals in dimension 1 and 2 of figure 2.38 display a homogeneous cloud of points for Stanford and a vague triangular shape for UMD. If we look at the coloring, we can see subjects for which the data points are closer together, suggesting a subject effect or within subject correlation.

Similarly to the STATIS results, gradients of samples exist when considering the *Lactobacillus* spp. proportion of this genus. The scores plot using lactobacilli proportion as a covariate can be found in appendix 4.7.

The main results of the loadings' plot in figure 2.39 are summarized below: Stanford:

- Metabolites: In contrast to STATIS results, the size effect is not present. Hippurate really stands out and is opposed to tyramine and putrescine's vectors in the first 2 dimensions. N-acetyl-cadaverine also stands out in the same plot. There is a group of correlated vectors around spermidine and lactate in both 1-2 and 3-4 dimensions. Phosphate stands out in dimension 3 and 4 because it is one of the only two metabolites, along with adenosine, positively correlated to the third PC.

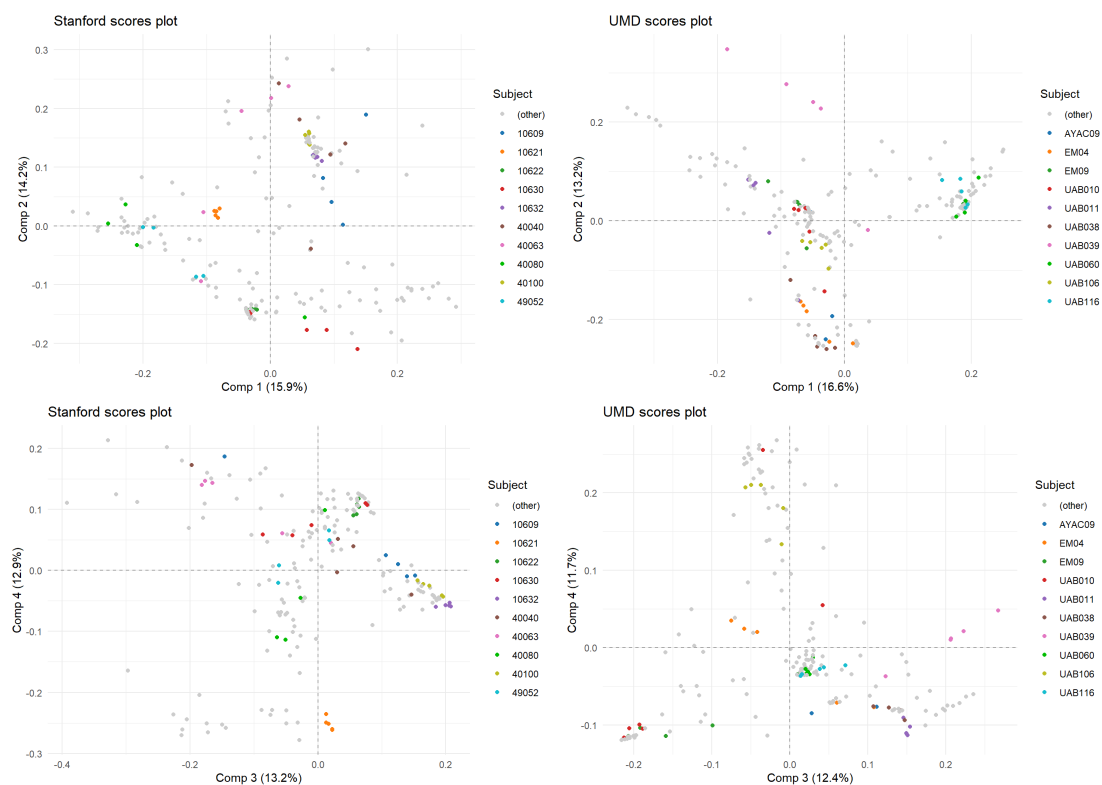


Figure 2.38: SCA' scores plot using the first 4 principal components for Stanford (left) and UMD (right); 10 random subjects have been identified using different colors

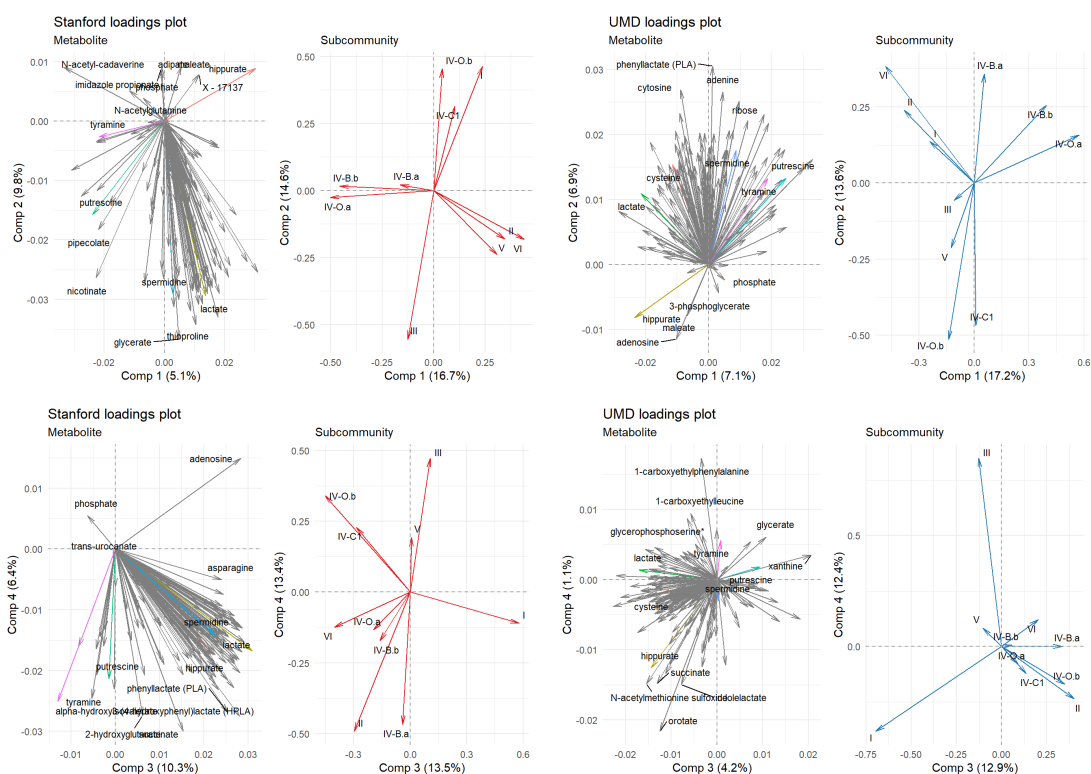


Figure 2.39: SCA' loadings plot using the first 4 principal components for Stanford (left) and UMD (right)

- Subcommunities: In the first and second dimension loadings plot, 4 groups of variables are shown: a first group with II, V, and VI; a second group with I, IV-O.b, IV-C1; a third group composed only of subcommunity III; and a last group with IV-O.a, IV-B.a, and IV-B.b. In dimensions 3 and 4, these groups of variables are not retained. Instead, subcommunity I seems independent of II and III. II is correlated to IV-B.a. But, the correlation between IV-C1 and IV-O.b remains strong.
- Both: Based on first two dimensions, hippurate is associated with subcommunity I. N-acetyl-cadaverine, tyramine and putrescine are associated with subcommunity IV-O.a and IV-B.b. Spermidine, lactate and glycerate are associated with all *Lactobacillus* subcommunity except I. In dimensions 3 and 4, adenosine is associated with III and I. Tyramine and putrescine are associated with II and IV-B.a. Phosphate is associated with IV-O.b and IV-C1.

UMD:

- Metabolites: Hippurate, adenosine, and maleate along with phosphate oppose the majority of the metabolites positively correlated to the second dimension. The most outstanding metabolite in this crowd of vectors is phenyllactate. 1-carboxyethyl-phenylalanine is highly correlated to the fourth dimension. Orotate is the most negatively correlated metabolite to the fourth dimension.
- Subcommunities: Three groups of variables are present in dimension 1 and 2: a first group with subcommunities I, II, and IV; a second group with IV-B.a, IV-B.b, and IV-O.a; and a last group with IV-C1, IV-O.b, III, and V. In dimensions 3 and 4, subcommunities I and III make two corners of a triangle, and subcommunities II, IV-O.b, and IV-C1 make the third corner. All other variables are contained in that triangle.
- Both: Hippurate and adenosine might be associated with subcommunities IV-C1, IV-O.b, and V. Hippurate, along with orotate, is also associated with subcommunity I. Tyramine, putrescine, and spermidine are associated with IV-B.a, IV-B.b, and IV-O.a. 1-carboxyethyl-phenylalanine is associated with subcommunity III. Lactate and cysteine are probably associated with I, II, and VI.

CCA results

In canonical correlation analysis, one directly uses the variables of each data block, without involving a compromised matrix, to form linear combinations that maximize a correlation criteria. It is important to recall that the compositional

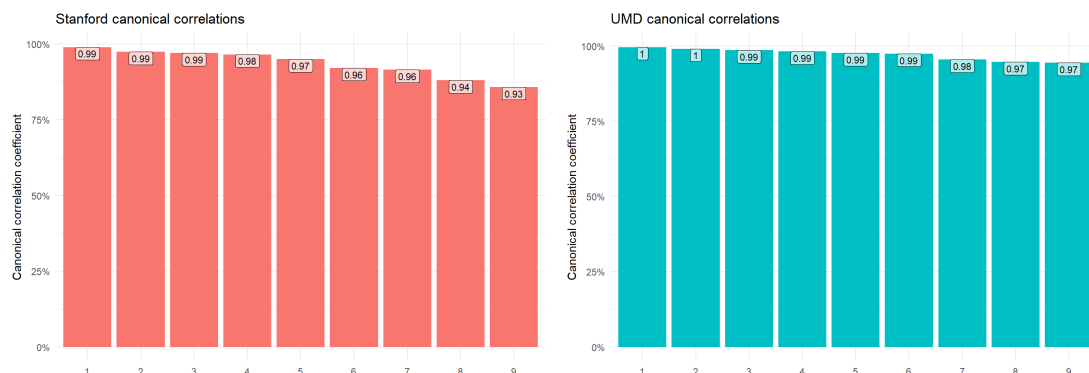


Figure 2.40: Barplot of canonical correlations rounded to the nearest hundredth for Stanford (left) and UMD (right)

data, which originally had 10 variables, was modified: subcommunity IV-O.b was discarded because the compositional data was not a full rank matrix. Indeed, any one variable could be explained by the 9 remaining variables. Therefore, only 9 columns were selected in the compositional datasets of Stanford and UMD for the CCA.

The barplots of figure 2.40 show the value of the correlations, called *canonical correlations*. As we can see, all canonical correlations ρ are very high (greater than 90%) which means the algorithm was able to find good linear combinations of the variables.

We then turn our attention towards the contribution of the block variables to the variates shown in figure 2.41. It is important to realize that the names of the axis can be misleading. For example, *variate 1* of metabolites is not the same as *variate 1* of subcommunities. These two *variate 1* are just highly correlated with one another. In our case, the correlations are extremely high (0.99 between variate 1 of metabolite and subcommunity).

A naive and simple interpretation of the previous plots could mention that:

- For Stanford: hippurate, lactate, spermidine, and creatinine are associated with subcommunity I. Methionine is associated with II and III. Succinate is associated with II. Adenosine and aconitate is associated with III. Putrescine and tyramine are associated with IV-B.b and IV-O.a. Spermidine is also associated with IV-B.a. Phosphate might be associated with V.
- For UMD: Lactate is associated with all lactobacilli subcommunities. Orotate and cystein are associated with I. Uridine is associated with II. 1-carboxyethyl-phenylalanine and tyramine are associated with III. Putrescine and tyramine

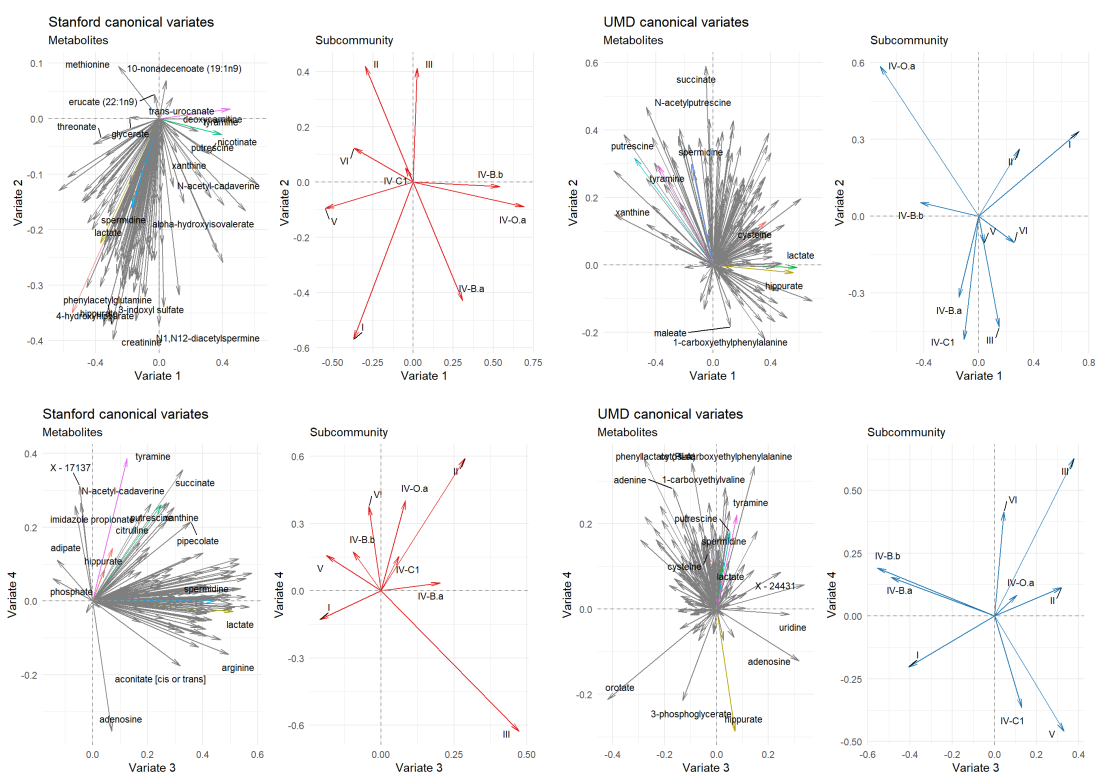


Figure 2.41: CCA's variable contribution to their respective block-specific variates 1, 2, 3, and 4 for Stanford (left) and UMD (right)

could be associated with IV-O.a. Hippurate is associated with subcommunity VI.

One can further these interpretation by focusing on the 8 most contributing variables, in absolute value, of variate 1 for instance. The results are presented in table 2.12 and 2.13.

Block	V1 coefficient	Stanford	
		Abs(V1 coeff)	Feature
Metabolites	-0.6322	0.6322	asparagine
Metabolites	-0.5752	0.5752	glycerophosphorylcholine (GPC)
Metabolites	-0.5485	0.5485	hippurate
Metabolites	-0.4859	0.4859	tryptophan
Metabolites	-0.4746	0.4746	maltose
Metabolites	-0.4741	0.4741	acetylcarnitine (C2)
Metabolites	0.5438	0.5438	alpha-hydroxyisovalerate
Metabolites	0.6193	0.6193	N-acetyl-cadaverine
Sub-community	-0.5487	0.5487	V
Sub-community	-0.3711	0.3711	I
Sub-community	-0.3699	0.3699	VI
Sub-community	-0.2995	0.2995	II
Sub-community	-0.04955	0.04955	IV-C1
Sub-community	0.3083	0.3083	IV-B.a
Sub-community	0.5386	0.5386	IV-B.b
Sub-community	0.6905	0.6905	IV-O.a

Table 2.12: Stanford's 8 most contributing metabolites to CCA's first variate

Block	V1 coefficient	UMD	
		Abs(V1 coeff)	Feature
Metabolites	-0.6948	0.6948	alpha-hydroxyisovalerate
Metabolites	-0.6907	0.6907	xanthine
Metabolites	-0.6242	0.6242	deoxycarnitine
Metabolites	-0.5504	0.5504	putrescine
Metabolites	0.5602	0.5602	hippurate
Metabolites	0.5855	0.5855	lactate
Metabolites	0.6144	0.6144	glycerophosphoglycerol
Metabolites	0.6935	0.6935	asparagine
Sub-community	-0.7168	0.7168	IV-O.a
Sub-community	-0.4265	0.4265	IV-B.b
Sub-community	-0.1462	0.1462	IV-B.a
Sub-community	-0.1075	0.1075	IV-C1
Sub-community	0.1485	0.1485	III
Sub-community	0.2577	0.2577	VI
Sub-community	0.2976	0.2976	II
Sub-community	0.7285	0.7285	I

Table 2.13: UMD’s 8 most contributing metabolites to CCA’s first variate

In Stanford’s results, asparagine is the most contributing metabolite: when asparagine average levels increase, one expects the sub-community V to become more predominant, and the sub-community IV-O.a to decrease in population. Hippurate is also amongst the 8 most contributing metabolites: when the average levels of hippurate increase, one expects the proportion of sub-community I to increase.

In UMD’s results, when the average levels of putrescine increase, one expects the proportion of IV-O.a to increase as well. And when the average levels of hippurate and lactate increase, one expects the general population of textitLactobacillus to increase.

One could repeat such a procedure with the other variates.

2.2.5 Comparative study

The aim of this section is to find matching results in Stanford and UMD cohorts from the previous section. We first quickly delve into the differences of the preprocessing stage. In a second part, we make comparisons of the data integration section.

The preprocessing was slightly adjusted to the cohort for some steps. For the ASV metagenomic data, the preparation was identical in UMD and Stanford cohort. Then, for the metabolites, the data filtering and transformation was also identical.

However, Stanford metabolite dataset required a batch correction, but UMD did not. In addition to this, a size effect, correlated to the first principal component of a PCA analysis, was identified in both cohort’s metabolite data. Correcting this size effect could easily be done by discarding the first PC. However, because removing that size effect led to lower correlation between *Lactobacillus* spp. proportions and lactate abundances in both cohorts, the size effect was believed to be of biological rather than technical origin. Therefore, the integration analyses were performed on the data from which the size effect was not removed.

This data integration section revealed potentially important and re-occurring associations between metabolites and bacterial subcommunities. The most important visible associations are compiled in figure 2.42 showing a colormap of associations between a list of metabolites and every subcommunities for each approach (STATIS, SCA, and CCA).

Let us detail the consistent discoveries that the colormap portrays. In every approach, the cells at the junction of the in the lactate rows and the lactobacilli subcommunities column are colored, which means a first consistent association is the one between lactate and all *Lactobacillus*-dominated subcommunities. Though it was expected, it is satisfying to see it. Secondly, subcommunity I (dominated by *Lactobacillus crispatus*) is frequently associated with hippurate, orotate, and adenosine. This association with orotate is only present in UMD results, while adenosine association with I is only present in Stanford results. This is an example of difference between Stanford and UMD. However, we cannot say this difference is the product of a recruitment site effect or a pregnancy effect because they are confounded.

Moving on, there is hardly any consistency in the associations found with *Lactobacillus gasseri* dominated subcommunity II (except for lactate). Conversely, subcommunity III (dominated by *Lactobacillus iners*) has been associated with lactate, spermidine, adenosine, and 1-carboxyethyl-phenylalanine more than once. This previous discovery contrasts with the fact that spermidine, along with tyramine and putrescine, has been recursively identified in association with the non-*Lactobacillus*-dominated subcommunity IV-B.a. Therefore, there might be a link between III and IV-B.a through spermidine, regardless of cohort.

A very obvious and re-occurring association is the one between the correlated metabolite couple tyramine-putrescine and the subcommunity IV-B.b and IV-O.a: all approaches allowed us to identify this link, which is particularly relevant in UMD and Stanford for the SCA results. Besides, even though subcommunity IV-O.b was not explored in CCA (due to rank issues), SCA and STATIS suggest its association with phosphate. Finally, subcommunities V and VI (respectively dominated by *Lactobacillus jensenii* and other *Lactobacillus* spp.) were associated with hippurate and lactate. Furthermore, in Stanford, both subcommunities were

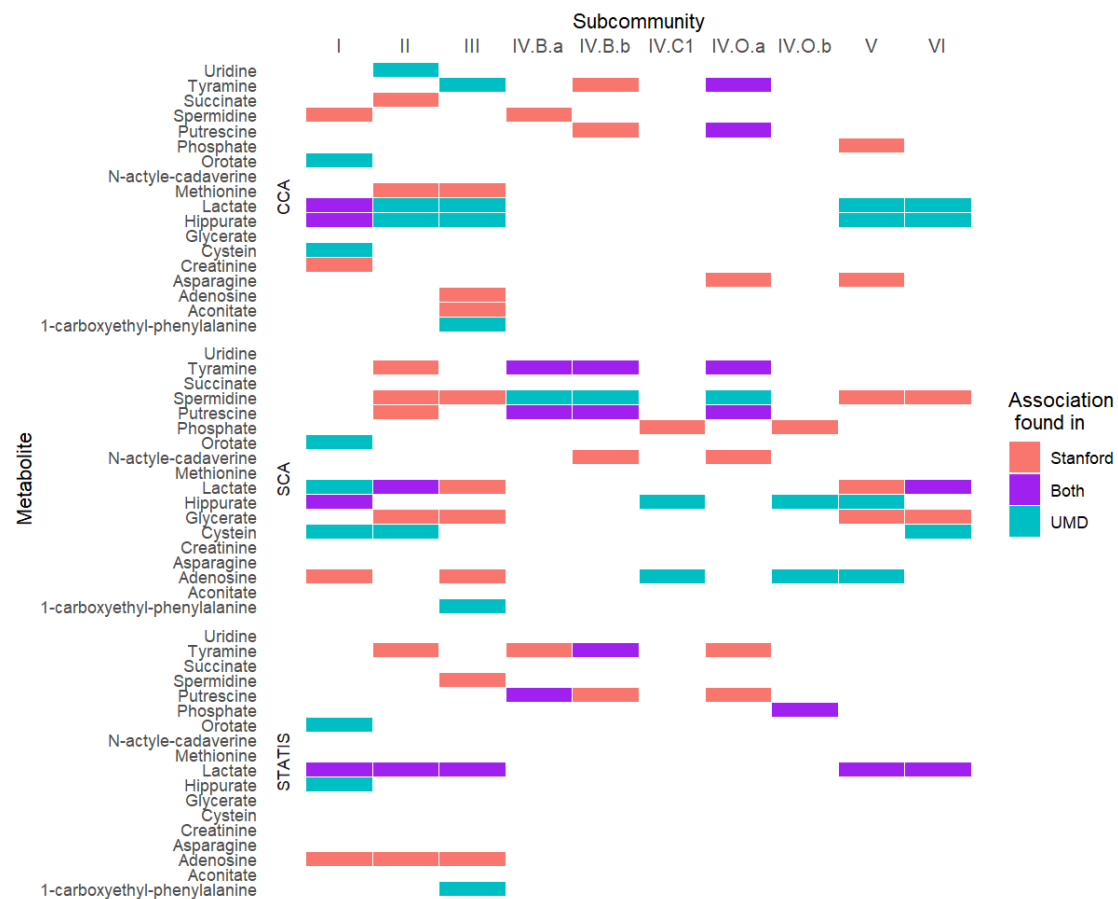


Figure 2.42: Colormap of associations identified between metabolites and microbial sub-communities per approach (CCA, SCA, STATIS); associations are colored according to whether they were identified in **Stanford**, **UMD** or **both**

associated with spermidine and glycerate.

In conclusion to this comparative study, we have reasons to believe that some of the outcomes of the methods used in this project are present in the larger population of women from the United States and beyond pregnancy status. This is the case for lactate and hippurate which have been frequently associated with lactobacilli dominated subcommunities in both Stanford and UMD. Despite this, putrescine and tyramine are consistently associated with non-lactobacilli subcommunities except IV-C1 in both cohorts. But the occurrence of some results still cannot be trusted to be a product of the pregnancy effect or the recruitment site effect. For example, spermidine has been associated with non-lactobacilli-dominated subcommunity IV-B.a, IV-B.b, and IV-O.a in UMD, but spermidine has also been linked to lactobacilli-dominated subcommunities I, III and V in Stanford. It is impossible to reconcile this discovery without investigating this association further.

2.3 Discussion

The overarching goal of this work was to integrate multi-omic datasets to understand and identify metabolites associated with microbiota composition. To do so, we used two sources: a set of 200 samples coming from pregnant women enrolled in *Stanford University* cohort and a set of 200 samples from non-pregnant women enrolled in *University of Maryland* (UMD). It was important to notice that the two cohorts differ in two major points: on one hand, they differed by pregnancy, and on the other hand, they differed in location (recruitment site). Thus, we pointed out the confounding effects of the pregnancy and the recruitment site. Since we could not isolate either effect, we opted for a parallel analysis of Stanford and UMD cohort respectively, with the intention of making a cautious comparative study of downstream results.

Both cohorts were associated with a metabolomic and metagenomic dataset. The metabolomic dataset presented the abundance of 853 metabolites for each sample quantified by mass-spectrometry, and the metagenomic dataset was a contingency table of 3590 amplified microbial 16S rRNA gene counts. In their raw form, neither of these sets could be efficiently used to infer associations. They both had to be preprocessed before the integration analysis.

This preprocessing had to be carefully planned to adjust for the non-essential effects and to discharge statistical noise while retaining the data variability of interest. It was not only data-specific but also cohort-specific to some extent.

For the metagenomics data, we used relative bacterial abundance rather than the absolute abundance since microbiota data are considered compositional (the proportions of each taxa are typically more relevant than the absolute number of taxa to describe ecological samples and the total number of reads usually results from

technical artifacts than from the true biological depth of the sample). Furthermore, we reduced the dimensionality of the dataset by applying topics model: soft-classification model assuming samples are made up of proportions of each different class instead of belonging to one unique class. This modelling application was performed on the combined Stanford–UMD compositional data because we believe that any difference in samples’s biological or technical aspects between Stanford and UMD is negligible compared to the biological synergies existing between bacteria. The result of said model was able to summarize the sample’s bacterial composition into ten different variables called topics or subcommunities. These ten topics could be grouped into two categories: *Lactobacillus* and non-*Lactobacillus* dominated microbiota. This distinction was actually important because the vaginal microbiota is typically considered healthy when its flora is dominated by lactobacilli strains, and considered unhealthy otherwise.

The preprocessing of the metabolomics data was as complicated as it was necessary to the downstream analyses. The first challenge was to address the very high proportion of missing values. Exploration of the missingness pattern led us to believe that some metabolite concentrations were below a certain limit of detection, which we didn’t know about, and therefore metabolites were not missing at random. Additionnaly, self-collected samples most certainly induced poor sample quality and thus meant some samples contained more missing values than others. We partially addressed these issues by first filtering out metabolites with more than 85% of missing values and then filtering out samples with more than 50% of missing values among the remaining metabolites. From the 853 original metabolites, only 132 remained in Stanford and 165 in UMD. The choice of filter thresholds was arbitrary and ulterior works could apply other percentages.

We also checked for the biochemical’s abundance’s distribution and found out that, out of a series of potential procedures, a log10 transformation was the most efficient transformation to stabilize the variance and decrease the mean-variance correlation. Both of these conditions were essential to making the sample wise information comparable.

The filtering and transformation was followed by a principal component analysis to further explore the data and remaining sources of variability. This PCA highlighted a subject effect (samples from the same subject were correlated), a batch effect (samples from the same batch were correlated) and a size effect (all samples were correlated to the first principal component).

The batch effect was only present in Stanford cohort as this cohort was the only one with samples emanating from two different batches. The batch effect was found to be confounded with metabolites mean abundance and the racial group. It was later corrected using `limma` package’s `removeBatchEffect()` function taking into account both metabolite abundance and subject’ self-proclaimed race. The

correction was verified through a PCA which showed no more batch distinction in the scores plot.

Apart from this, the size effect was present in both Stanford and UMD. This size effect is a direct consequence of the highly variable metabolite concentration in the collection vials. Because it was correlated to the first principal component, the size effect could be resolved by removing this first PC. However, it was shown that the size effect correction led to too much information loss in terms of metabolite-association with microbiota composition (inferred from the *Lactobacillus*-vs-lactate scatterplot). In that case, it was decided to use the filtered, transformed, and batch effect corrected metabolite dataset for Stanford and the filtered and transformed dataset for UMD, and to omit the size effect correction.

After the preprocessing steps, we underwent the integration analysis using three different approaches: STATIS, Simultaneous Component Analysis (SCA), and Canonical Correlation Analysis (CCA).

STATIS and SCA approaches have a similar framework: compute the similarity matrices of the component data blocks (*i.e.*, metabolite and compositional data in this project), take a weighted average of these similarity matrices, and perform a PCA on this average. They also have the same plots to investigate: variance decomposition barplot, scores plot, loadings plot. However, these methods differ by the weight they give to each block when computing the average. In STATIS, the weights are computed based on the RV-coefficient (a coefficient of correlation between two matrices). In SCA, all weights are equal to 1. This difference in weights explains why the STATIS and SCA results are different. Another source of difference is the PCA rotation.

In STATIS approach, recall that the weight of the metabolite data block were greater than the composition data weight for both Stanford and UMD. An interpretation to this weight allocation is because the metabolite data divides its total variance between many more variables (132 biochemicals for Stanford and 165 for UMD) than the compositional data does (10 subcommunities). Therefore, STATIS compensates this variance imbalance by giving more importance to the metabolite block.

In CCA, we also found orthogonal linear combinations of block variables, but the process was more direct and did not involve eigendecomposition. Furthermore, CCA builds its linear combinations, called canonical variates, by maximizing the correlations between these variates. It is worth mentioning that, in order to find linear combinations, CCA relies on the number of variables in each data block, and more specifically to the rank of the blocks. Because any 1 variable in the compositional data is a linear combination of the remaining 9 variables, the rank of this datatable was 9. Thus, CCA found a maximum of 9 pairs of canonical variates, all of which had a respective canonical correlation greater than 90%. The reason

why all canonical correlations are high is because the rank of the metabolite data was very large (132 for Stanford and 165 for UMD). The way CCA approach is built makes it that one can always find high canonical correlations if the rank of either data block is "virtually" infinite. Instead of using high dimensionality datasets, ulterior works could suggest a more appropriate method like Sparse Canonical Correlation Analysis (but this approach was outside the scope for this project).

In terms of result comparison, STATIS and SCA, in spite of having similar frameworks, yielded different outcomes. A first difference observed between their respective results is the decomposition of the variance. Even though STATIS had a first principal component with greater explained variance, SCA required less PC to reach a cumulative variance threshold of 75%. Furthermore, STATIS' first PC explained the data variability linked to the previously identified size effect and SCA's was not. It would seem like SCA method "disolved" the size effect into different principal components.

A second discovery made through STATIS and SCA is that the samples were still characterized by within subject correlations. In that sense, subjects with similar metabolic profiles or microbial composition could potentially give more or less weight to the associations found throughout the analysis. The subject effect also artificially inflated the correlation between the data blocks (measured by the RV coefficient): correlations between subjects in both the metabolite and subcommunity data implies that similar patterns are found between the blocks. Alternatively, we would expect a lower RV coefficient if all samples were independent and the within subject correlation was null.

Finally, the associations between metabolites and microbial subcommunities were revealed in the canonical variates plot of CCA approach and the loadings plot of STATIS and SCA approach. They were first identified independently in all method' subsection then compiled in a comparative study. We lacked time to make a thorough comparative study, and therefore it mainly relied on qualitative observations. Nevertheless, a use of coefficients, metrics, and more quantitative methods would be welcome in the future. Besides, it would be interesting to have the expert opinion of biologists and/or clinicians on the matter to have a better apprehension of the information.

In the meanwhile, associations were identified in all three approaches and in both Stanford and UMD cohorts. The most consistent and predictable one was the lactate's link to all *Lactobacillus*-characterized subcommunities (*i.e.*, subcommunities I, II, III, V, and VI). Hippurate is also reoccurringly associated with lactobacilli subcommunities I, V, and VI. Subcommunity III was associated with adenosine, and 1-carboxyethyl-phenylalanine. Another reoccurring association is the one between tyramine and putrescine and the subcommunities IV-B.b and IV-O.a.

However, some associations were not consistent between the two cohorts. For

example, spermidine was associated with *L. iners*' subcommunity III in UMD results. But spermidine was also associated with non-*Lactobacillus*-dominated subcommunity IV-B.a in Stanford. Another example on non-consistent association is the one between orotate and adenosine and subcommunity I. Orotate's association was only present in UMD results, while adenosine's association with I was only present in Stanford results. These intriguing discoveries cannot be simply explained by a subject's attribute differences. It is impossible to tell whether these associations are a consequence of the pregnancy effect or the site effect.

Perspectives

A posteriori, it is obvious the data collection method lacked organization. A more carefully planned *experimental design* to collect the samples and generate the data would have given many opportunities and allow for deeper analysis of the data. Indeed, a thoughtful design would have resolved the confounding effects issue and let us have a guess of the causality of results' differences through hypothesis testing. Furthermore, it would have helped the preprocessing phase of this work.

Regarding the preprocessing, ulterior works could consider different options. It was mentioned in section 2.1.7 that the filtering threshold of the metabolite's missing values was arbitrary. Hence, this threshold could be changed in the future and one could explore the impact of changing this threshold on the final results. One could also change the imputation method to deal with missing data. Or one could also evade the imputation altogether and plan on using missing-tolerant integration methods. Moreover, the choice of the transformation could be different: for example, the hyperbolic arc-sine transformation revealed to a good candidate yielding comparable variance-stabilized biochemicals. Besides, knowing there are many packages to adjust for batch effect, one could modify or replace the batch correction step. In the metagenomics data preprocessing, the value of the number of topics K to summarize the information can be increased to have more precise subcommunities.

Ultimately, a sensitivity analysis of these hyper-parameters could be useful to estimate the impact of any change on the outcomes of the integration analysis. The retaining of the size effect correction would certainly have major repercussion on the results and would be worth investigating in the future.

Finally, an additional omic layer could yield more information on the present situation. An addition of proteomic data, for instance, could help us make a more direct link between bacterial presence and metabolite profiles.

Chapter 3

Conclusions

In conclusion, this work project successfully integrated multi-omic data to characterize the associations between metabolite abundance and vaginal microbiota composition. The meticulous preprocessing of both metabolomic and metagenomic data was essential for ensuring the reliability and accuracy of our findings. Through comprehensive data integration techniques, we uncovered consistent associations between specific metabolites and microbial subcommunities, shedding light on the complex interplay within the vaginal environment. These insights enhance our understanding of vaginal health, although the analysis was constrained by the confounding effects of pregnancy status and site collection differences between Stanford and UMD cohorts. Future studies should address these confounding factors by planning an experimental design more carefully, thereby further refining our understanding of the intricate biological networks that underpin vaginal microbiota and metabolome interactions.

Chapter 4

Appendixes

4.1 Appendix 1

Microbiota proportion barplots of *Lactobacillus iners* and *Gardnerella vaginalis*

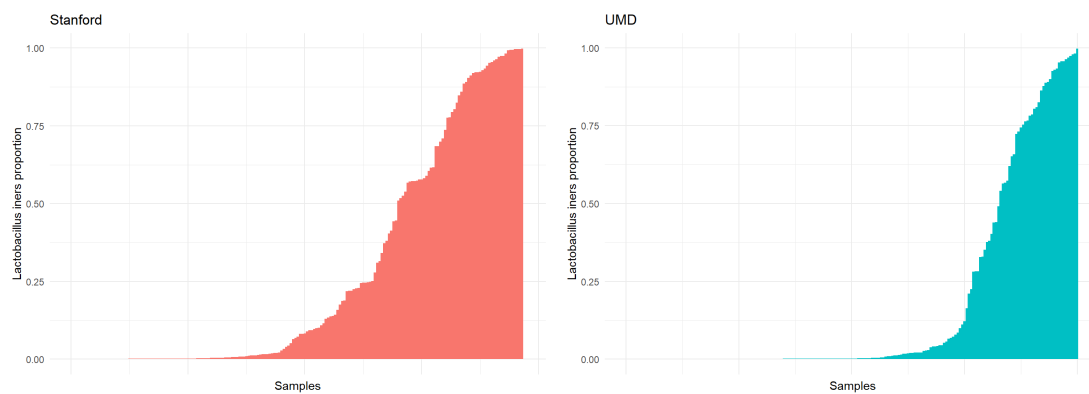


Figure 4.1: Ordered proportion of *Lactobacillus iners* in Stanford (left) and UMD (right) samples

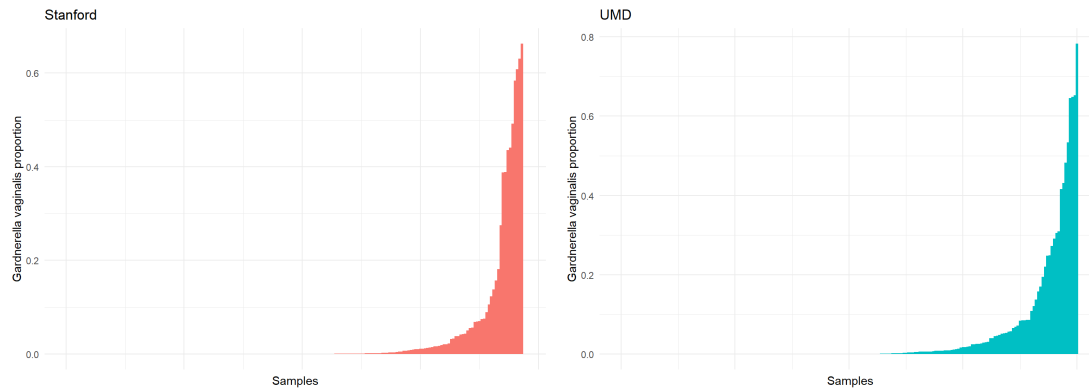


Figure 4.2: Ordered proportion of *Gardnerella vaginalis* in Stanford (left) and UMD (right) samples

4.2 Appendix 2

Correlation plots between mean metabolite abundance and sample attributes

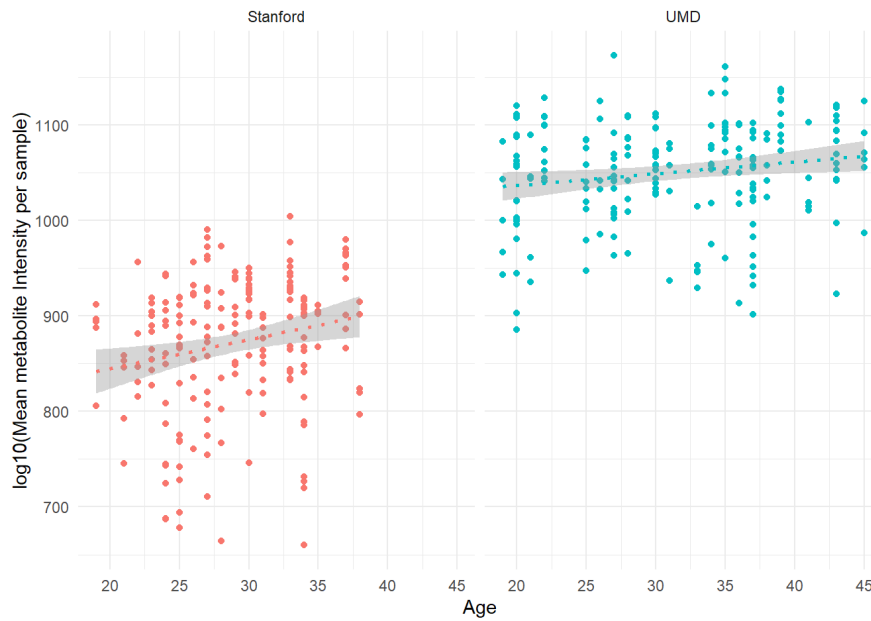


Figure 4.3: Scatterplot of the log10 mean metabolite peak intensities per sample as a function of the women's age (in years) for Stanford (left) and UMD (right)

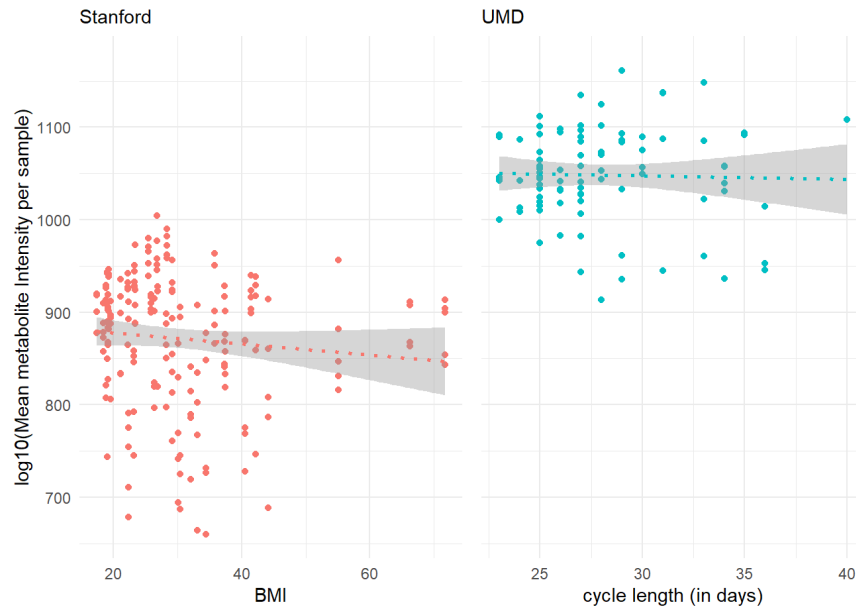


Figure 4.4: Scatterplot of the log10 mean metabolite peak intensities per sample as a function of Stanford's women's BMI (left) or UMD's women menstrual cycle length in days (right)

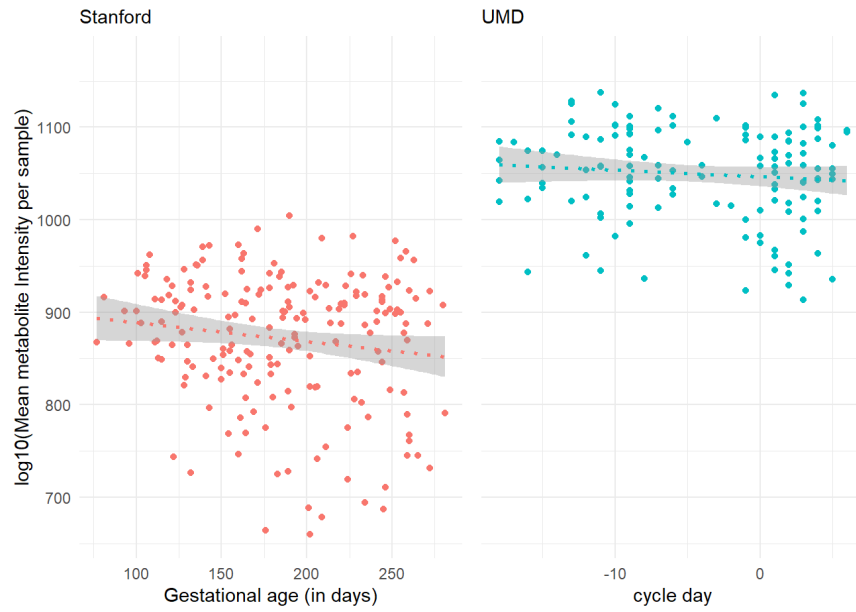


Figure 4.5: Scatterplot of the log10 mean metabolite peak intensities per sample as a function of Stanford's women's gestational age in days (left) or UMD's women's cycle day (right)

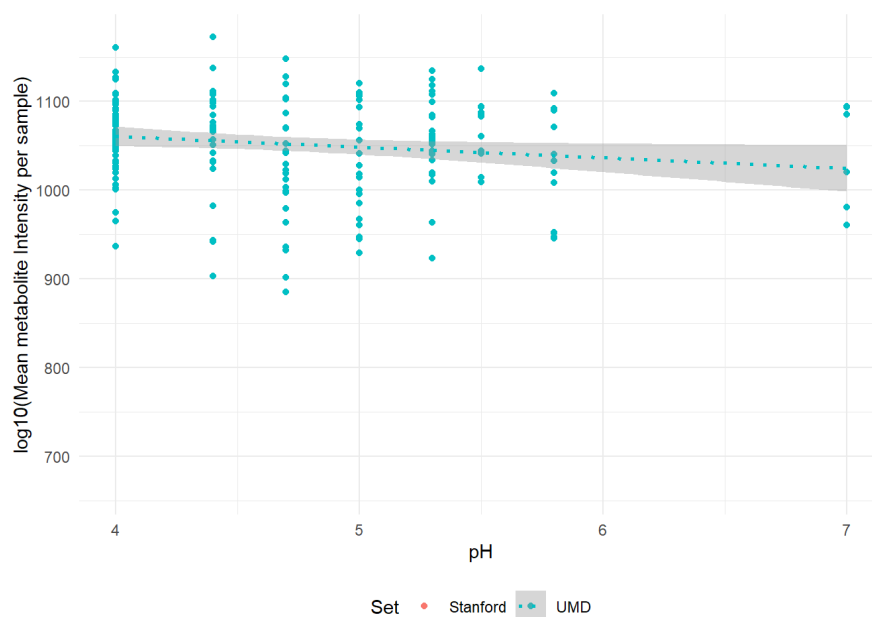


Figure 4.6: Scatterplot of the log10 mean metabolite peak intensities per sample as a function of UMD's women's vaginal pH

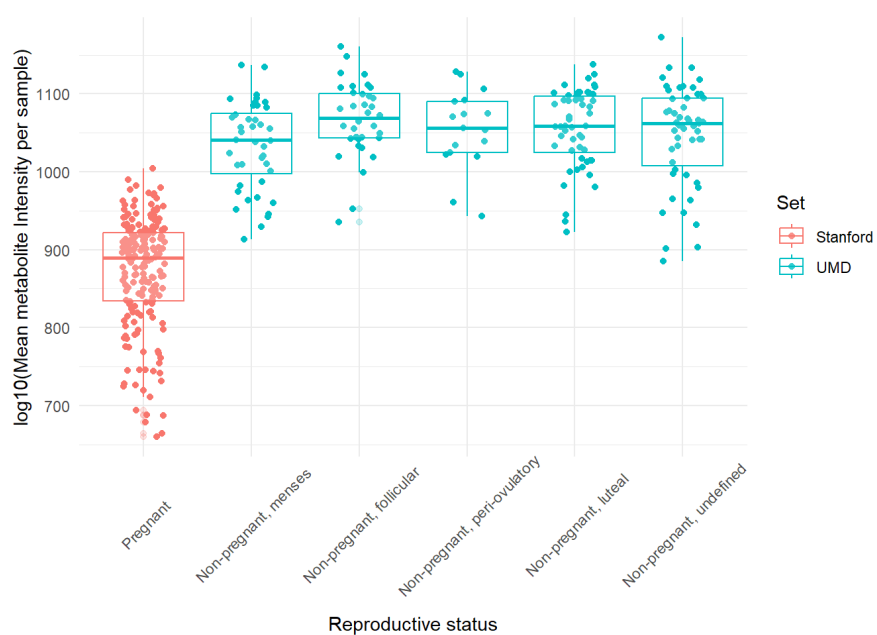


Figure 4.7: Scatterplot of the log10 mean metabolite peak intensities per sample as a function of all women's reproductive status

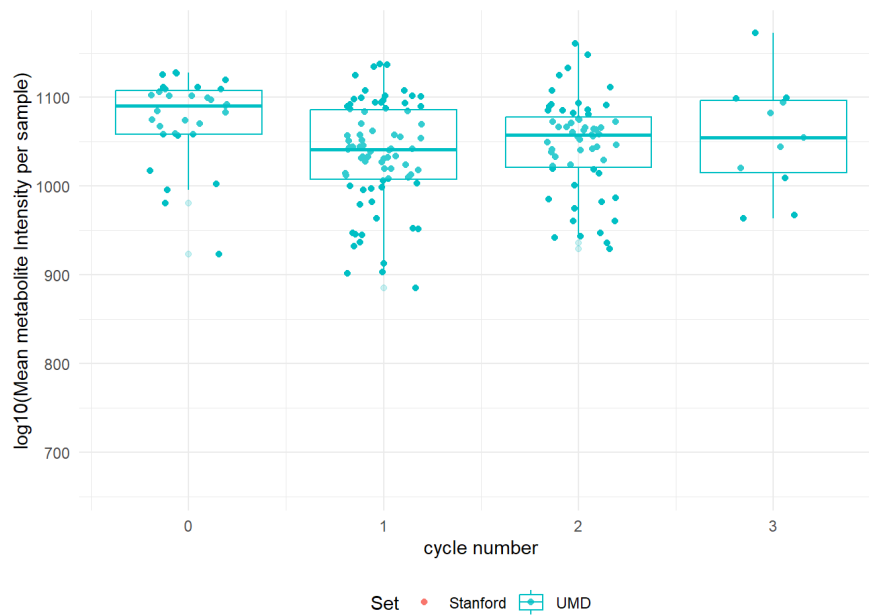


Figure 4.8: Scatterplot of the log10 mean metabolite peak intensities per sample as a function of UMD's women's cycle number

4.3 Appendix 3

Supplementary graphs of the descriptive analysis' principal component analysis

	eigenvalue	variance.percent	cumulative.variance.percent		eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	76.56	58	58	Dim.1	68.86	41.73	41.73
Dim.2	11.99	9.085	67.09	Dim.2	18.28	11.08	52.81
Dim.3	9.535	7.223	74.31	Dim.3	13.68	8.293	61.1
Dim.4	5.834	4.419	78.73	Dim.4	6.624	4.014	65.12
Dim.5	2.318	1.756	80.49	Dim.5	5.21	3.157	68.27
Dim.6	1.919	1.454	81.94	Dim.6	4.423	2.68	70.95
Dim.7	1.827	1.384	83.33	Dim.7	3.305	2.003	72.96
Dim.8	1.449	1.098	84.42	Dim.8	2.767	1.677	74.63
Dim.9	1.263	0.9571	85.38	Dim.9	2.595	1.572	76.21
Dim.10	1.119	0.8476	86.23	Dim.10	1.944	1.178	77.38

Figure 4.9: Eigenvalue table of Stanford (left) and UMD (right) principal component analysis on filtered, imputed, transformed metabolomic dataset

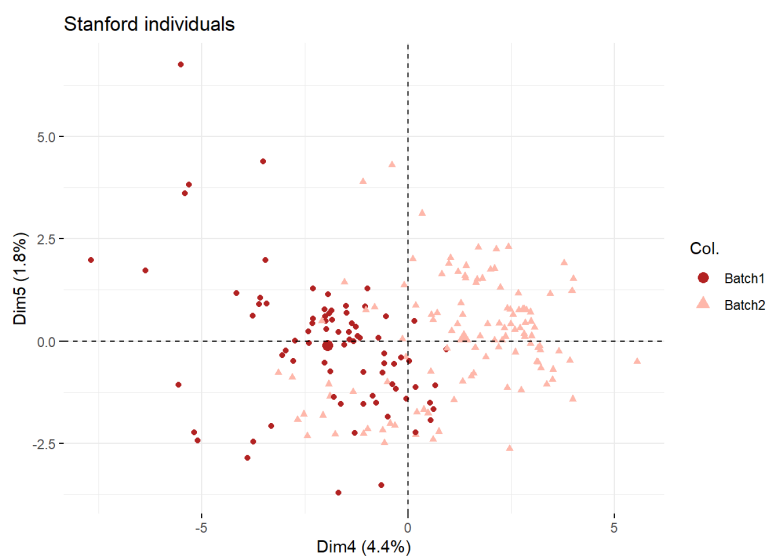


Figure 4.10: Scores plot of Stanford (left) colored by batch; PCA on filtered, imputed, transformed metabolomic dataset

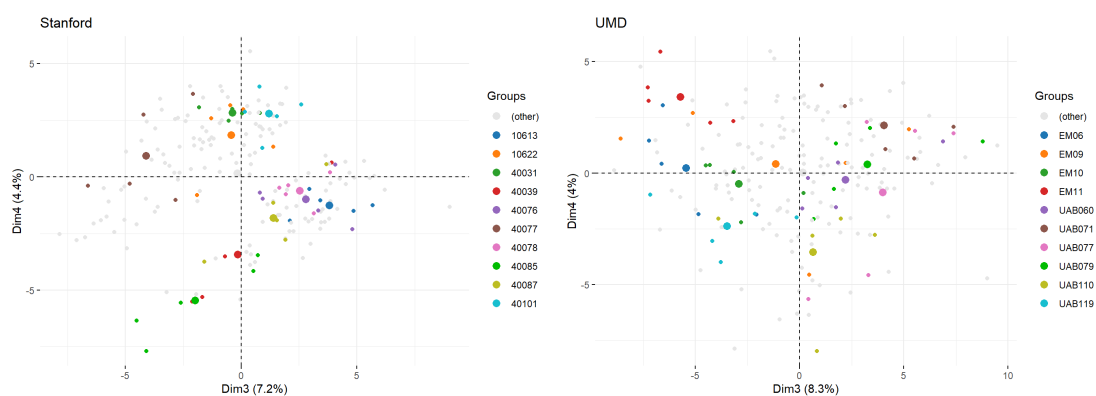


Figure 4.11: Scores plot of Stanford (left) and UMD (right) colored according to 10 random subjects; PCA on filtered, imputed, transformed metabolomic dataset

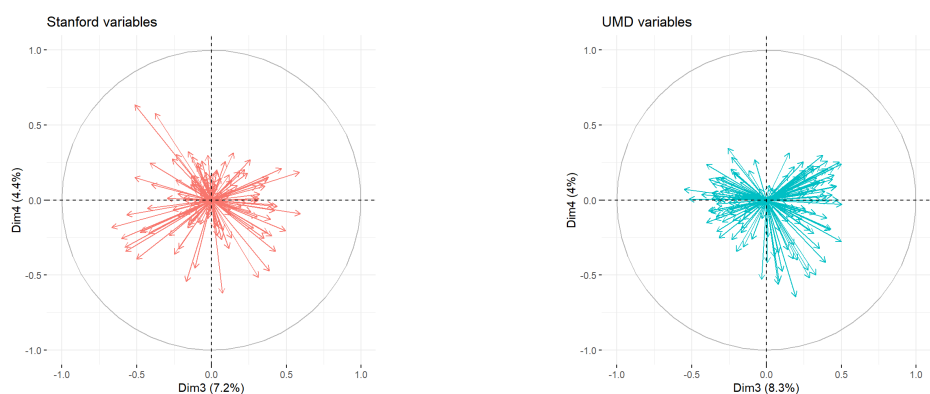


Figure 4.12: Loadings plot of Stanford (left) and UMD (right); PCA on filtered, imputed, transformed metabolomic dataset

4.4 Appendix 4

Supplementary PCA results after batch correction and size effect correction

	eigenvalue	variance.percent	cumulative.variance.percent		eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	14.6	7.055	7.055	Dim.1	12.73	6.064	6.064
Dim.2	8.831	4.266	11.32	Dim.2	12.48	5.94	12
Dim.3	5.959	2.879	14.2	Dim.3	11.67	5.555	17.56
Dim.4	5.051	2.44	16.64	Dim.4	10.64	5.066	22.63
Dim.5	4.418	2.134	18.77	Dim.5	10.24	4.878	27.5
Dim.6	4.221	2.039	20.81	Dim.6	9.455	4.502	32.01
Dim.7	3.863	1.866	22.68	Dim.7	9.092	4.329	36.34
Dim.8	3.357	1.622	24.3	Dim.8	8.964	4.269	40.6
Dim.9	3.253	1.572	25.87	Dim.9	8.926	4.251	44.86
Dim.10	2.879	1.391	27.26	Dim.10	8.296	3.95	48.81

Figure 4.13: Eigenvalue and variance decomposition tables of Stanford (left) and UMD (right) PCA results for size effect corrected metabolomic datasets

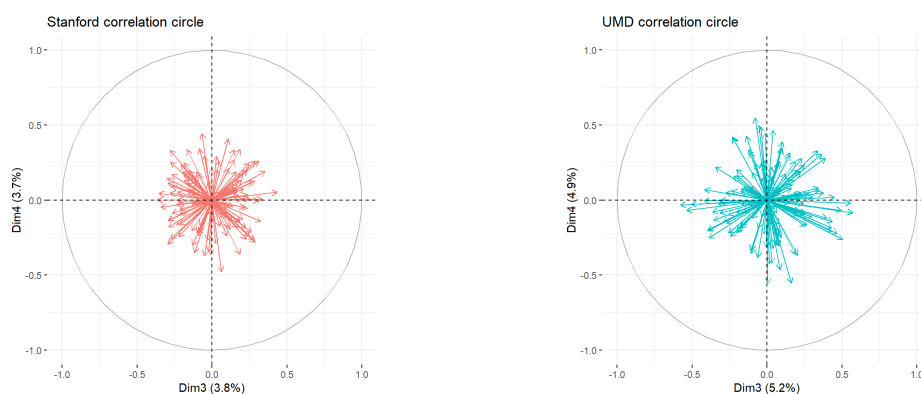


Figure 4.14: Loading's plot (dimension 3 and 4) of Stanford (left) and UMD (right) PCA results for size effect corrected metabolomic datasets

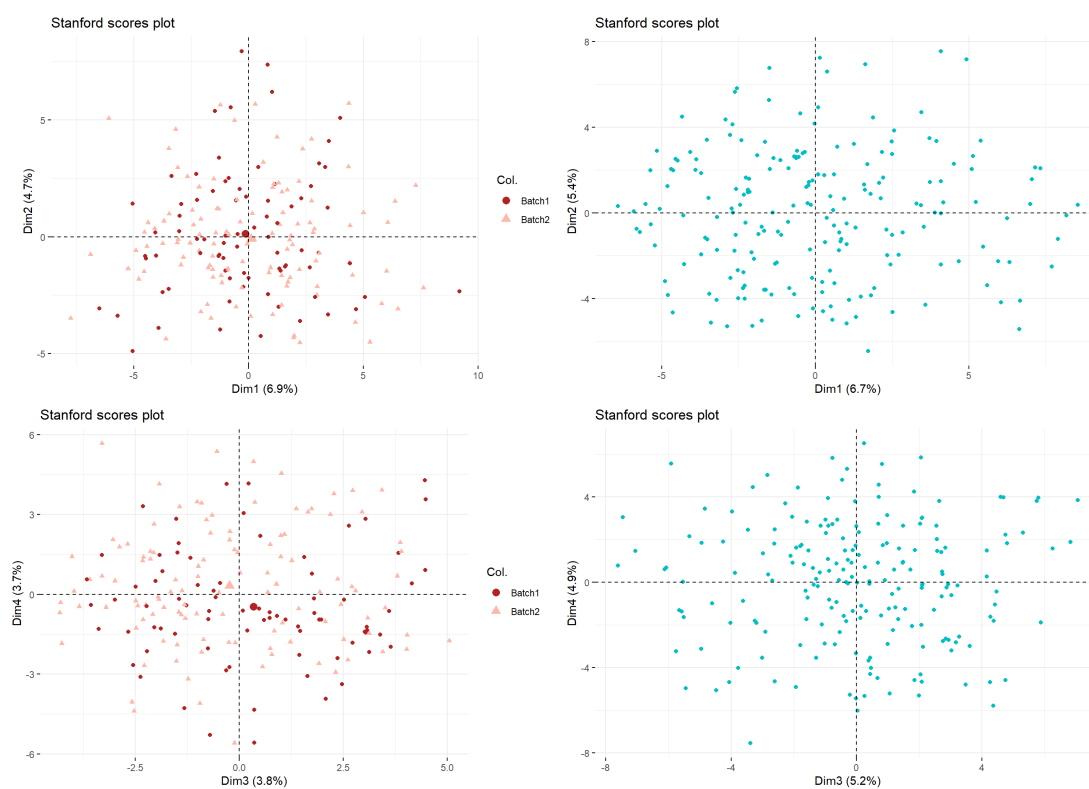


Figure 4.15: Scores's plot of Stanford (left) and UMD (right) PCA results for size effect corrected metabolomic datasets

4.5 Appendix 5

Supplementary notes on package `limma` in the context of batch correction

Many studies have investigated the *correction of batch effect* (Yu et al., 2023) (Liu et al., 2020). Most of them rely on homogenizing the location and dispersion of data points of all batches. It is rather similar to a standardizing or normalization procedure.

In R, different packages exist for batch correction: `sva`, `limma`, `MetabolomicsQC`, `BatchCorr`, `xcms`, etc. Out them all, `limma` might be the most adequate one as it can control for potential confounding variables and take incomplete datasets (i.e. with missing values) as input.

The `limma` package from Bioconductor offers multiple tools for exploring and modeling omics data. It was initially designed to work with gene expression count data, but was later adapted to a great number of different -omics data, including metabolomics. The main function of this package is to fit a linear model for every gene (or metabolite) of a dataset and to compile the information to infer differential expression. It not only generalizes the analysis, but it can also take into account the between gene variance and the sample correlation structures (for example, repeated measures or batch). However, `limma` is also adapted for pre-processing steps of a standard data analysis procedure. (Ritchie et al., 2015)

4.6 Appendix 6

Additional context and information regarding STATIS method and plot interpretation

The interpretation of the interstructure plot

The cosine of the angle between two vectors $O\vec{Q}_k$ and $O\vec{Q}_l$ in the interstructure subspace is an approximation of the RV coefficient between the block matrices k and l (Lavit, 1988). Furthermore, the interstructure plot tells us two things:

1. If the vectors are close to one another, it means the S_k objects are close together. And if the S_k are close together, it means there is a structure amongst individuals between tables.
2. Vectors' magnitude reflect the contribution of the tables to the compromise: only the vectors (derived from the matrix S_k of the k^{th} datatable) with a great norm on the first axis (horizontal) contribute to the construction of the compromise.

Additionnal notes on the compromise sub-space, the trajectories, and the column coordinates on the compromise.

The compromise subspace can be considered as the scalar product between the individuals. We can use the scalar products to portray the *position* of the individuals taking into account all tables. When the distances between the S_k objects (translated by the interstructure) are small, we can conclude of a significant connection between individuals across datatables. Lavit (1988) also mentions that the trajectories reveal individuals who are responsible for the dissimilarities between tables (the dissimilarities are of course a synonym of the distances computed in the interstructure).

We can compute the correlation between the variables of each table with the first 2 axis of the compromise c_1 and c_2 . Moreover, we can compute the correlation between the tables (through S_k) with the compromise axis; and the correlation between the individuals and the compromise axis.

PCA framework

All PCA in the STATIS analysis are performed using eigendecomposition (or "factorisation canonique" as L'Hermier des Plantes and Thiébaud (1977) used to call it) and follow the framework below:

- Assume the PCA is performed on the centered and scaled $(n \times p)$ matrix X
- Compute the correlation matrix $\rho = X'X$
- Apply eigendecomposition of ρ and obtain the $[u_\alpha]_{\alpha=1,\dots,p}$ eigenvectors associated with the $[\lambda_\alpha]_{\alpha=1,\dots,p}$ eigenvalues. The eigenvectors are orthonormal such that $u'_\alpha u_\beta = 1$ if $\alpha = \beta$ and $u'_\alpha u_\beta = 0$ if $\alpha \neq \beta$.
- The $(p \times p)$ loadings matrix is $P = U = [u_1 \ u_2 \ \dots u_p]$
- The $(n \times p)$ scores matrix is $T = [t_1 \ t_2 \ \dots t_p]$ where $t_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} X \cdot u_\alpha$
- The PCA can be written such that $T = XP$

4.7 Appendix 7

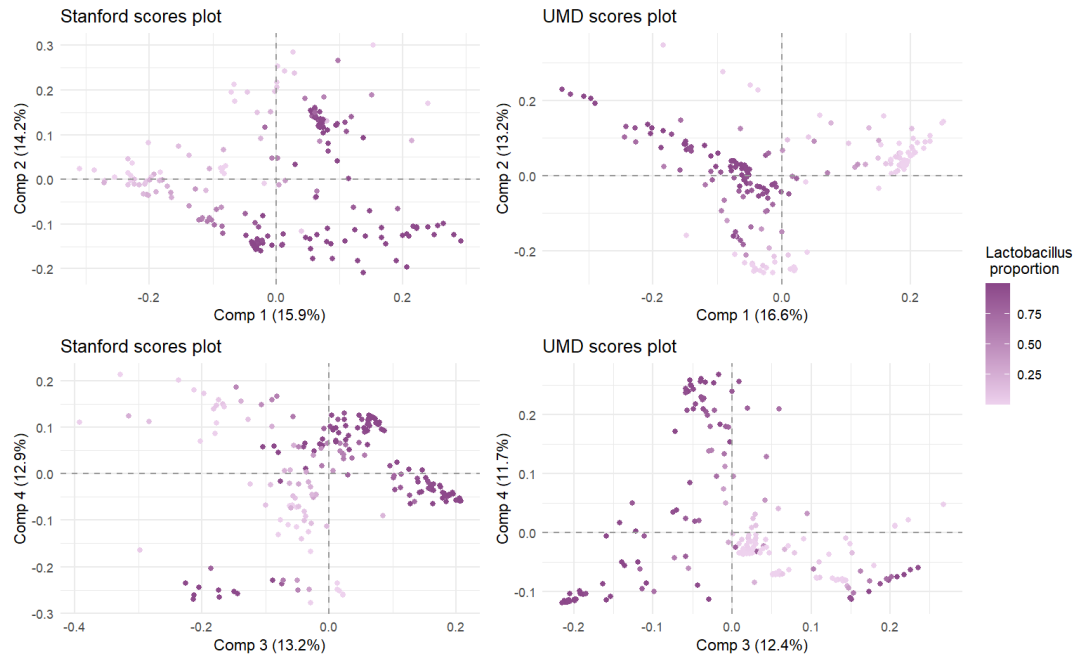


Figure 4.16: SCA' scores plot using the first 4 principal components for Stanford (left) and UMD (right)

Bibliography

- George Armstrong, Gibraan Rahman, Cameron Martino, Daniel McDonald, Antonio Gonzalez, Gal Mishne, and Rob Knight. Applications and comparison of dimensionality reduction methods for microbiome data. *Frontiers in Bioinformatics*, 2, 2022. ISSN 2673-7647. doi: 10.3389/fbinf.2022.821861. URL <https://www.frontiersin.org/articles/10.3389/fbinf.2022.821861>.
- J-LC. Borgogna, MD. Shardell, SG. Grace, EK. Santori, B. Americus, Z. Li, A. Ulanov, L. Forney, TM. Nelson, RM. Brotman, J. Ravel, and CJ. Yeoman. Biogenic amines increase the odds of bacterial vaginosis and affect the growth of and lactic acid production by vaginal lactobacillus spp. *Appl Environ Microbiol*, 87:e03068–20, 2021. doi: doi.org/10.1128/AEM.03068-20.
- Stéphanie Bougeard and Stéphane Dray. Supervised multiblock analysis in r with the ade4 package. *Journal of Statistical Software*, 86(1):1–17, 2018. doi: 10.18637/jss.v086.i01. URL <https://www.jstatsoft.org/index.php/jss/article/view/v086i01>.
- J. Bruin. newtest: command to compute new test @ONLINE, February 2011. URL <https://stats.oarc.ucla.edu/stata/ado/analysis/>.
- B. Callahan, P. McMurdie, and S. Holmes. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J*, 11:2639–43, 2017a. doi: doi.org/10.1038/ismej.2017.119.
- BJ. Callahan, DB. DiGiulio, DSA. Goltsman, CL. Sun, EK. Costello, P. Jeganathan, JR. Biggio, RJ. Wong, ML. Druzin, GM. Shaw, DK. Stevenson, SP. Holmes, and DA. Relman. Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of us women. *Proc Natl Acad Sci USA*, 114(37):9966–71, Sep 2017b. doi: doi.org/10.1073/pnas.1705899114.
- Wikipedia contributors. 16s ribosomal rna — Wikipedia, the free encyclopedia, 2024a. URL https://en.wikipedia.org/w/index.php?title=16S_ribosomal_RNA&oldid=1212133039. Online; accessed 2024 5 April.

- Wikipedia contributors. Human microbiome, Apr 2024b. URL https://en.wikipedia.org/wiki/Human_microbiome. Online; accessed April 04 2024.
- Wikipedia contributors. Omics, Feb 2024c. URL <https://en.wikipedia.org/wiki/Omics>. Online; accessed April 05 2024.
- K. De Roover, E. Ceulemans, M. E. Timmerman, K. Vansteelandt, J. Stouten, and P. Onghena. Clusterwise simultaneous component analysis for analyzing structural differences in multivariate multiblock data. *Psychological Methods*, 17(1):100–19, 2012a. doi: <https://doi.org/10.1037/a0025385>.
- Kim De Roover, Eva Ceulemans, and Marieke E. Timmerman. How to perform multiblock component analysis in practice. *Behavior Research Methods*, 44(1): 41–56, 01 2012b. doi: 10.3758/s13428-011-0129-1. URL <https://doi.org/10.3758/s13428-011-0129-1>.
- F. De Seta, G. Campisciano, N. Zanotta, G. Ricci, and M. Comar. The vaginal community state types microbiome-immune network as key factor for bacterial vaginosis and aerobic vaginitis. *Frontiers in microbiology*, 10(2451), 2019. doi: doi.org/10.3389/fmicb.2019.02451.
- K. Dettmer, P. A. Aronov, and B. D. Hammock. Mass spectrometry-based metabolomics. *Mass spectrometry reviews*, 26(1):51–78, 2007. doi: doi.org/10.1002/mas.20108.
- Llobell Fabien, Cariou Véronique, Vigneau Evelyne, Labenne Amaury, and Qannari El Mostafa. Analysis and clustering of multiblock datasets by means of the stasis and clustatis methods. application to sensometrics. *Food Quality and Preference*, 79:103520, 2020. doi: <https://doi.org/10.1016/j.foodqual.2018.05.013>. URL <https://www.sciencedirect.com/science/article/pii/S0950329318302520>.
- Julia Fukuyama, Kris Sankaran, and Laura Symul. Multiscale analysis of count data through topic alignment. *Biostatistics*, 24(4):1045–1065, 06 2022. ISSN 1465-4644. doi: 10.1093/biostatistics/kxac018. URL <https://doi.org/10.1093/biostatistics/kxac018>.
- Ignacio González, Sébastien Dejean, Pascal G.P. Martin, and Alain Baccini. CCA: An R package to extend canonical correlation analysis. *Journal of Statistical Software*, 23(12):1–14, 2008. doi: 10.18637/jss.v023.i12. URL <https://hal.inrae.fr/hal-02665561>.
- C. Gosmann, M. N. Anahtar, S. A. Handley, M. Farcasanu, G. Abu-Ali, B. A. Bowman, N. Padavattan, C. Desai, L. Droit, A. Moodley, M. Dong, Y. Chen,

- N. Ismail, T. Ndung'u, M. S. Ghebremichael, D. R. Wesemann, C. Mitchell, K. L. Dong, C. Huttenhower, B. D. Walker, H.W. Virgin, and D. S. Kwon. Lactobacillus-deficient cervicovaginal bacterial communities are associated with increased hiv acquisition in young south african women. *Immunity*, 46(1):29–37, 2017. doi: doi.org/10.1016/j.immuni.2016.12.013.
- Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3-4): 321–377, 12 1936. ISSN 0006-3444. doi: 10.1093/biomet/28.3-4.321. URL <https://doi.org/10.1093/biomet/28.3-4.321>.
- Farid Huma. Bacterial vaginosis, Nov 2023. URL <https://www.buoyhealth.com/learn/bacterial-vaginosis>. Online; accessed 2024 06 April.
- E. Kaambo, C. Africa, R. Chambuso, and J.S. Passmore. Vaginal microbiomes associated with aerobic vaginitis and bacterial vaginosis. *Frontiers in public health*, 6(78):pp–pp, 2018. doi: doi.org/10.3389/fpubh.2018.00078.
- Yankuba Kassama, Yun Xu, Warwick B. Dunn, Nick Geukens, Jozef Anné, and Royston Goodacre. Assessment of adaptive focused acoustics versus manual vortex/freeze-thaw for intracellular metabolite extraction from streptomyces lividans producing recombinant proteins using gc-ms and multi-block principal component analysis. *Analyst*, 135(5):934–42, 2010. doi: 10.1039/B918163F.
- Hendrik Albert Lambertus Kiers. *Three-way methods for the analysis of qualitative and quantitative two-way data*. PhD thesis, 1989. Relation: http://www.rug.nl/date_submitted : 2008Rights : University of Groningen.
- S. Kovachev. Defence factors of vaginal lactobacilli. *Critical reviews in microbiology*, 44(1):31–9, 2018. doi: doi.org/10.1080/1040841X.2017.1306688.
- C. Lavit. *Analyse conjointe de tableaux quantitatifs*. Masson, Paris, 1988.
- Joseph Levin. Simultaneous factor analysis of several gramian matrices. *Psychometrika*, 31(3):413–419, 1966. doi: 10.1007/BF02289472. URL <https://doi.org/10.1007/BF02289472>.
- Henri L'Hermier des Plantes, 06 Structuration des tableaux a trois indices de la statistique. URL https://three-mode.leidenuniv.nl/bibliogr/lhermierdesplantesh_thesis/lhermierdesplantesh1976.pdf. Online; accessed may 28 2024.
- Henri L'Hermier des Plantes and Bernard Thiébaut. Étude de la pluviosité au moyen de la méthode s.t.a.t.i.s. *Revue de Statistique Appliquée*, (2):57–81, 1977. URL http://www.numdam.org/item/RSA_1977__25_2_57_0/.

- Kristian Hovde Liland. *multiblock: Multiblock Data Fusion in Statistics and Machine Learning*, 2023. <https://khliland.github.io/multiblock/>, <https://github.com/khliland/multiblock/>.
- Qin Liu, Douglas Walker, Karan Uppal, Zihe Liu, Chunyu Ma, ViLinh Tran, Shuzhao Li, Dean P. Jones, and Tianwei Yu. Addressing the batch effect issue for lc/ms metabolomics data in data preprocessing. *Sci Rep*, 10(13856), August 2020. doi: doi.org/10.1038/s41598-020-70850-0.
- C. Mack, Z. Su, and D. Westreich. Types of missing data, Feb 2018. URL <https://www.ncbi.nlm.nih.gov/books/NBK493614/>. Online; accessed April 10 2024.
- Clinic Mayo. Bacterial vaginosis - symptoms&causes, June 2023. URL <https://www.mayoclinic.org/diseases-conditions/bacterial-vaginosis/symptoms-causes/syc-20352279>. Online; accessed April 04 2024.
- Metabolon. Metabolon’s client data table (cdt), a. URL <https://www.metabolon.com/resources/videos/client-data-table/>. Online; accessed October 27 2023.
- Metabolon. Quantitation, b. URL <https://www.metabolon.com/why-metabolon/quantitation/>. Online; accessed October 27 2023.
- Metabolon. Chapter 2 - metabolomics study workflow, c. URL <https://www.metabolon.com/study-design/chapter-2-metabolomics-study-workflow/>. Online; accessed April 10 2024.
- Metabolon. Chapter 9—designing a metabolomics study, d. URL <https://www.metabolon.com/why-metabolomics/your-guide-chapter-9-designing-metabolomics-study/>. Online; accessed October 27 2023.
- M. Morgan, V. Obenchain, J. Hester, and H. Pagès. Summarizedexperiment: Summarizedexperiment container. r package version 1.34.0, 2024. URL <https://bioconductor.org/packages/SummarizedExperiment>. Online; accessed May 22 2024.
- Elena Parkhomenko, David Tritchler, and Joseph Beyene. Genome-wide sparse canonical correlation of gene expression with genotypes. In *BMC proceedings*, volume 1, page S119. Springer, 2007.
- K. Peebles, J. Velloza, J. E. Balkus, R. S. McClelland, and R. V. Barnabas. High global burden and costs of bacterial vaginosis: A systematic review and meta-analysis. *Sexually transmitted diseases*, 46(5):304–311, 2019. doi: doi.org/10.1097/OLQ.0000000000000972.

- Mårdh Per-Anders. The vaginal ecosystem. *American Journal of Obstetrics and Gynecology*, 165(4, Part 2):1163–1168, 1991. ISSN 0002-9378. doi: [https://doi.org/10.1016/S0002-9378\(12\)90721-8](https://doi.org/10.1016/S0002-9378(12)90721-8). URL <https://www.sciencedirect.com/science/article/pii/S0002937812907218>.
- Marcel Ramos, Lucas Schiffer, Angela Re, Rimsha Azhar, Azfar Basunia, Carmen Rodriguez, Tiffany Chan, Phil Chapman, Sean R. Davis, David Gomez-Cabrero, Aedin C. Culhane, Benjamin Haibe-Kains, Kasper D. Hansen, Hanish Kodali, Marie S. Louis, Arvind S. Mer, Markus Riester, Martin Morgan, Vince Carey, and Levi Waldron. Software for the Integration of Multiomics Experiments in Bioconductor. *Cancer Research*, 77(21):e39–e42, 10 2017. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-17-0344. URL <https://doi.org/10.1158/0008-5472.CAN-17-0344>.
- J. Ravel, P. Gajer, Z. Abdo, G.M. Schneider, S.S. Koenig, S.L. McCulle, S. Karlebach, R. Gorle, J. Russell, C.O. Tacket, R.M. Brotman, C.C. Davis, K. Ault, L. Peralta, and L.J. Forney. Vaginal microbiome of reproductive-age women. *Proceedings of the National Academy of Sciences of the United States of America*, 108(1):4680–87, 2011. doi: doi.org/10.1073/pnas.1002611107.
- Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 01 2015. doi: 10.1093/nar/gkv007. URL <https://doi.org/10.1093/nar/gkv007>.
- P. Robert and Y. Escoufier. A unifying tool for linear multivariate statistical methods: The rv- coefficient. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 25(3):257–265, 1976. ISSN 00359254, 14679876. URL <http://www.jstor.org/stable/2347233>.
- Kris Sankaran, Laura Symul, and Julia Fukuyama. *alto: Aligns Topics across LDA Models*, 2024. R package version 0.1.0.
- Age K. Smilde, Jeroen J. Jansen, Huub C. J. Hoefsloot, Robert-Jan A. N. Lamers, Jan van der Greef, and Marieke E. Timmerman. Anova-simultaneous component analysis (asca): a new tool for analyzing designed metabolomics data. *Bioinformatics*, 21(13):3043–8, 05 2005. doi: 10.1093/bioinformatics/bti476. URL <https://doi.org/10.1093/bioinformatics/bti476>.
- M Steinert, Ute Hentschel, and J Hacker. Symbiosis and pathogenesis: Evolution of the microbe-host interaction. *Die Naturwissenschaften*, 87:1–11, 02 2000. doi: 10.1007/s001140050001.

- Laura Symul and Fanny Hoogstoel. Traitement statistique de données -omics - introduction, Feb 2024. URL <https://moodle.uclouvain.be/course/view.php?id=2964>. Catholic University of Louvain course.
- Laura Symul, Pratheepa Jeganathan, Elizabeth K. Costello, Michael France, Seth M. Bloom, Douglas S. Kwon, Jacques Ravel, David A. Relman, and Susan Holmes. Sub-communities of the vaginal microbiota in pregnant and non-pregnant women. *Proceedings of the Royal Society B: Biological Sciences*, 290(2011):pp–pp, Nov 2023. doi: doi.org/10.1098/rspb.2023.1461.
- J. Thioulouse, Patrice Cadet, and Alain Albrecht. The use of permutation tests in co-inertia analysis : application to the study of nematode-soil relationships. In *Fourth meeting of the east, central and southern African network of the international biometric society : proceedings*, pages 144–152, 1995. URL <https://www.documentation.ird.fr/hor/fdi:010014295>.
- Robert A. van den Berg, Huub C.J. Hoefsloot, Johan A. Westerhuis, Age K. Smilde, and Mariët J. van der Werf. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 7(142), 08 2006. doi: [10.1186/1471-2164-7-142](https://doi.org/10.1186/1471-2164-7-142). URL <https://doi.org/10.1186/1471-2164-7-142>.
- Katrijn Van Deun, Age.K. Smilde, Mariët J. van der Werf, Henk AL. Kiers, and Iven Van Mechelen. A structured overview of simultaneous component based data integration. *BMC Bioinformatics*, 10(1):1471–2105, 11 2009. doi: [10.1186/1471-2105-10-246](https://doi.org/10.1186/1471-2105-10-246). URL <https://doi.org/10.1186/1471-2105-10-246>.
- H. Verstraelen and R. Verhelst. Bacterial vaginosis: an update on diagnosis and treatment. *Expert review of anti-infective therapy*, 7(9):1109–24, 2009. doi: doi.org/10.1586/eri.09.87.
- WHO. Bacterial vaginosis, Aug 2023. URL <https://www.who.int/news-room/fact-sheets/detail/bacterial-vaginosis>. Online; accessed April 04 2024.
- Sang Jun Yoon, Chae Bin Lee, Soon Uk Chae, Seong Jun Jo, and Soo Kyung Bae. The comprehensive “omics” approach from metabolomics to advanced omics for development of immune checkpoint inhibitors: Potential strategies for next generation of cancer immunotherapy. *International Journal of Molecular Sciences*, 22(13), 2021. doi: [10.3390/ijms22136932](https://doi.org/10.3390/ijms22136932). URL <https://www.mdpi.com/1422-0067/22/13/6932>.
- Ying Yu, Naixin Zhang, Yuanbang Mai, Luyao Ren, Qiaochu Chen, Zehui Cao, Qingwang Chen, Yaqing Liu, Wanwan Hou, Jingcheng Yang, Huixiao Hong, Joshua

- Xu, Weida Tong, Lianhua Dong, Leming Shi, Xiang Fang, and Yuanting Zheng. Correcting batch effects in large-scale multiomics studies using a reference-material-based ratio method. *Genome Biol*, 24(201), Sept 2023. doi: doi.org/10.1186/s13059-023-03047-z.
- B. Zhou, JF. Xiao, L. Tuli, and HW. Ransom. Lc-ms-based metabolomics. *Mol Biosyst*, 8(2):470–81, Feb 2012. doi: doi.org/10.1039/c1mb05350g.

