Characterization of non-antibiotic medication associated microbiome module in the human gut microbiome

A Project Report

submitted by

VINDHYA REGONDA

MT22208

in partial fulfilment of the requirements

for the award of the degree of

MASTER OF TECHNOLOGY

Computational Biology

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

18th May, 2024

THESIS CERTIFICATE

This is to certify that the thesis titled **Characterization of non-antibiotic medication associated microbiome module in the human gut microbiome**, submitted by **Vindhya Regonda,** to the Indraprastha Institute of Information Technology, Delhi, for the award of the degree of **Master's of Technology**, is a bona fide record of the research work done by him under my supervision. This thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Advisor's Name: Dr. Tarini Ghosh
Thesis Supervisor
Assistant Professor
Dept. of Computational Biology
IIIT Delhi, 110020

Place: New Delhi

Date: 18th May 2024

ACKNOWLEDGEMENTS

I extend my heartfelt appreciation to Dr. Tarini Shankar Ghosh for their invaluable guidance and mentorship, which played a pivotal role in shaping this research endeavor. Additionally, I am grateful to the dedicated team at the Microbiome Informatics lab for their technical support and collaboration, without which this project would not have been possible. I also want to acknowledge my family and friends for their unwavering encouragement and support throughout this journey. Their belief in me has been a constant source of motivation and inspiration.

Abstract

Pharmacomicrobiomics, an emerging field at the intersection of microbiology and pharmacology, explores the intricate relationship between drug response and the gut microbiome, offering new avenues for precision medicine. Lifestyle factors, including diet, exercise, smoking, and alcohol consumption, significantly influence drug metabolism and clearance, adding to the complexity of individual variability in drug response. Recent studies have investigated the impact of medications used for various diseases on the gut microbiome; however, categorized medications by class, a comprehensive understanding of microbiome-drug interactions with drug-specific structural characteristics is still unclear. Moreover, existing literature often focuses on specific therapeutic areas such as non-steroidal anti-inflammatory drugs (NSAIDs), anti-diabetic agents, and cancer therapeutics, overlooking the diverse chemical structures of drugs within each category. This study aims to identify the complex relationships between medication structure, gut microbiota composition, and host factors. We performed meta-analysis across more than 10 global cohorts encompassing > 6000 subjects. Additionally, we identify non-antibiotic medication associated modules of gut microbial taxa that show distinct associations with different drug classes. These drug-responsive modules also show differential enrichment of markers of health and disease, indicating that while certain drugs have a minimal or beneficial impact on the gut microbiomes, others can potentially drive the gut microbiome to a detrimental state. Furthermore, our findings highlight the pivotal role of cohort lifestyle, particularly industrialized living, in shaping bacterial species' response to drugs, underscoring the need for holistic approaches in understanding drug-microbiome interactions. Findings from this study can inform public health guidelines regarding the prescription and usage of different non-antibiotic drugs across global populations.

Keywords: gut microbiota, drug-microbiome interactions, association, microbial species, age, lifestyle, personalized healthcare, therapeutic outcomes.

Contents

List of Figures

Chapter 1

Introduction

The emerging field of pharmacomicrobiomics, describes the complex interplay between drug response and gut microbiome, represents a frontier in precision medicine. This interdisciplinary domain holds immense potential for enhancing therapeutic outcomes and minimizing adverse reactions by elucidating the impact of microbial composition on drug metabolism and efficacy. While the potential of individual variability in drug response (IVDR) is considerable, it presents a substantial challenge in clinical settings. Therefore, there is a pressing need for a thorough comprehension of the numerous determinants that impact drug efficacy and safety. [1], [2].

IVDR is a multifaceted phenomenon shaped by a myriad of factors, ranging from genetic predispositions to environmental influences. Genetic variations in drug metabolizing enzymes, transporters, and targets contribute to interindividual differences in drug metabolism and response. Age-related physiological changes, namely alterations in body composition, kidney and liver function, and hormonal levels, also influence drug pharmacokinetics and pharmacodynamics. Similarly, gender differences and hormone levels can lead to differential drug responses, highlighting the importance of considering dissimilarities between women and men in personalized medicine approaches [3].

Moreover, lifestyle factors, including diet, exercise, smoking, and alcohol consumption, exert profound effects on drug metabolism and clearance, further complicating IVDR. Disease states can alter organ function, drug metabolism, and receptor expression, thereby influencing drug response variability. Additionally, drug-drug interactions and environmental toxins and pollutants exposure can modify medicine pharmacokinetics and pharmacodynamics, resulting into IVDR. Amidst this complex landscape, emerging evidence suggests a pivotal role for gut microbiota in modulating drug responses, underscoring the need for deeper exploration into microbiome-drug interactions [2].

In addition to elucidating the factors influencing individual variability in drug response, it is imperative to consider the structural diversity inherent in pharmaceutical agents. Various drugs within categorized groups often possess distinct chemical structures, which can profoundly influence their interactions with gut microbiota. While prior research, such as the study by Forslund et al. [4], has explored the impact of cardiovascular disease (CVD) on the microbiome, grouping drugs by class, a comprehensive understanding of microbiome-drug interactions necessitates a nuanced examination of drug-specific structural characteristics. Furthermore, existing literature predominantly focuses on specific therapeutic areas, including non-steroidal anti-inflammatory drugs (NSAIDs), anti-diabetic agents, cancer therapeutics [5], cardiovascular medications, and central nervous system (CNS) disorder treatments [6], among others. However, these studies often overlook the diverse chemical structures of drugs within each therapeutic category, thereby limiting insights into the different effects of individual medicine on microbiome abundance and drug response variability. By considering the structural diversity of pharmaceutical agents and their implications for microbiome-drug interactions, our study aims to bridge this gap in knowledge and enhance the understanding of the complex pattern between gut microbiota, drug structure, and host factors in shaping drug responses.

In our study, we seek to unravel the complex relationships between medication structure, gut microbiota composition, and host factors. By leveraging comprehensive datasets from diverse studies, we aim to delineate how variations in drug structure influence bacterial species abundance and how host factors such as gender, age, lifestyle, and disease status contribute to alterations in microbiome composition. Through our interdisciplinary approach, we aspire to shed light on the mechanisms underpinning IVDR and microbiome-drug interactions, ultimately paving the way for personalized therapeutic interventions tailored to individual patient profiles. Our research holds the promise of revolutionizing precision medicine by optimizing drug therapy, mitigating the risk of adverse reactions, and improving patient outcomes in the era of personalized healthcare.

Chapter 2

METHODS

2.1 Dataset collection and pre-processing

For our study on the association between drug structure and the gut microbiome, we gathered relevant datasets from four distinct sources: MetaCardis [7], NuAge [8], ElderMet [9], and Japanese4D [10]. These datasets were selected based on the availability of comprehensive medication and bacterial species data. These datasets were shortlisted after reviewing that each dataset provided sufficient information to analyze the relationship between drug exposure and microbial composition in the gut. By incorporating these diverse datasets, we aimed to capture a broad spectrum of drug-microbiome associations across different populations and contexts. The comprehensive profiling of 759 drugs in Japanese 4D, along with the metadata collection, including anthropometrics, diets, lifestyles, diseases and physical activities, offers a rich dataset for identifying the effects of drugs on the gut microbiota. Additionally, fecal samples were collected before and after drug intake and which longitudinal insights into the dynamic nature of drug-microbiome interactions. However, for our analysis, we took the data from a particular time point. Unlike the previous paper, ElderMet didn't identify the association between medication and gut microbiome association; nevertheless, the study collected comprehensive metadata pertaining to medication usage alongside other demographic and clinical variables within a geriatric population cohort. MetaCardis collated 1,241 samples of middle-aged European cohort, encompassing healthy people, those with dysmetabolic conditions such as type 2 diabetes and obesity, and individuals diagnosed with IHD. In the case of NuAge, 612 individuals from five European countries (UK, France, Netherlands, Italy, and Poland) were taken for study and collected cohort metadata like body mass index, age, disease, gender, medication usage and pathophysiologies.

Additionally, we included a dataset used in a recently published research paper from

the Microbiome Informatics Lab at IIIT Delhi comprising 138 studies.

In the data preprocessing phase, the raw datasets obtained from ElderMet, NuAge, and MetaCardis were transformed into a standardized binary matrix format, with samples represented along the rows and medications along the columns. To ensure consistency, drug names that were initially recorded as brand names were systematically annotated into their corresponding generic names. Annotation guidelines were made to ensure uniformity across different datasets. This annotation process adhered to a predefined guideline, referencing authoritative sources such as the ATC/DDD Index 2024 WHO website and "The Essentials of Medical Pharmacology" book.

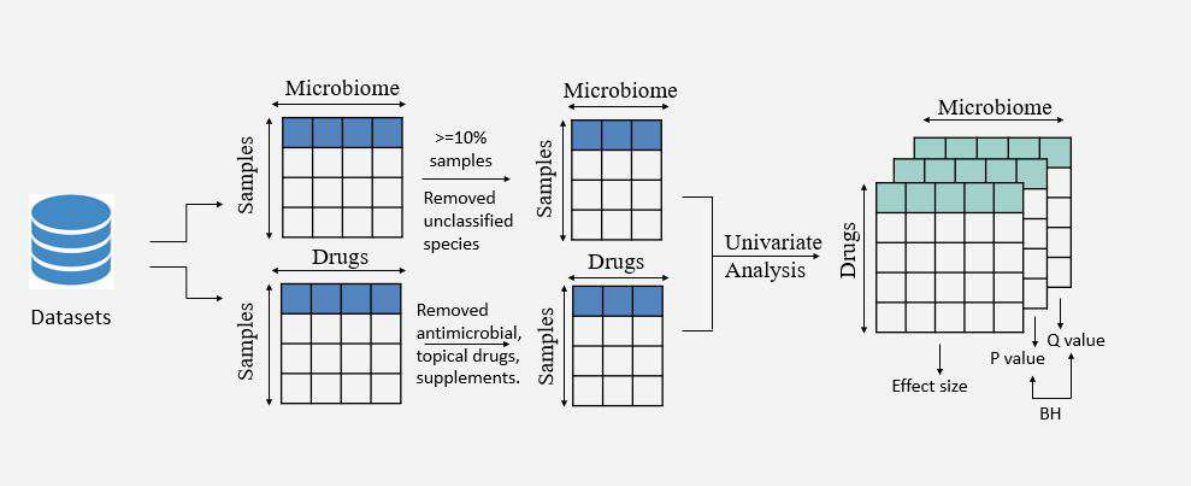

Figure 2.1: Schematic of Dataset processing

Subsequently, a comprehensive filtering step was implemented to exclude certain classes of medications from further analysis. Specifically, antimicrobial drugs, antiviral, antifungal, antibiotics, health supplements, herbal drugs, inhalers, vitamins, biologics, topical medications, and ophthalmic drugs were removed from consideration as they bypass the GIT, thus having less influence on the gut microbiome. This refinement ensured that the subsequent analyses focused specifically on therapeutically relevant drugs with potential implications for the gut microbiome. Additionally, a total of 1,972 viral species, comprising unclassified species and those identified only at the family level, were excluded from the ElderMet; 12 species were excluded due to their classification at the family taxonomic level in MetaCardis' dataset and 66 unclassified species were excluded from NuAge dataset. Further computed univariate analysis to identify the association between microbiome and drugs by computing effect size, p and q values.

Schematic of above-explained process is given in [Figure 2.1]

2.2 Methods

2.2.1 Computation of association between gut microbiome and medication

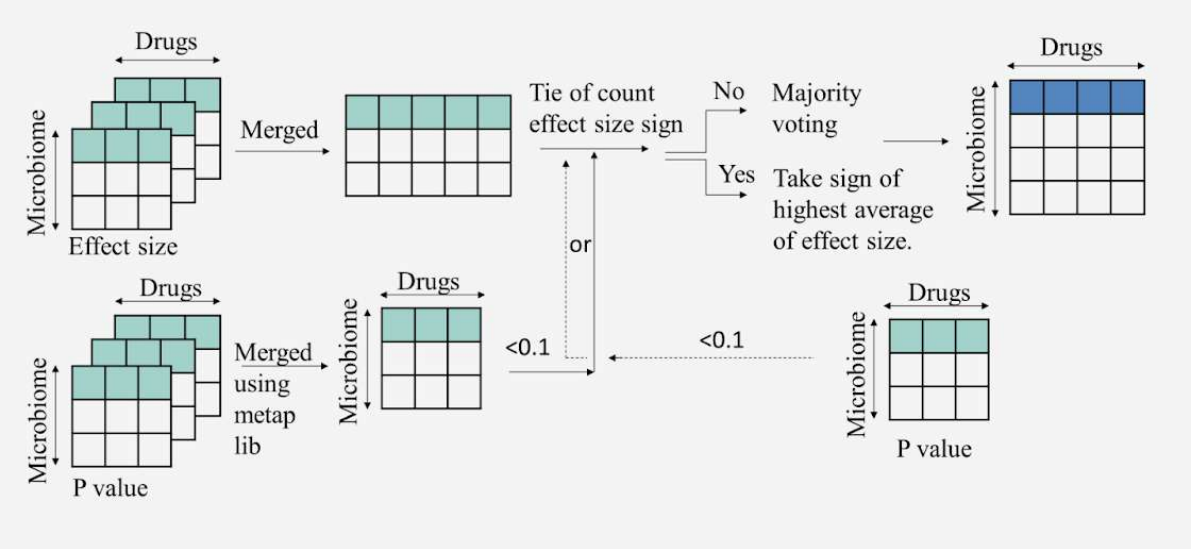

Figure 2.2: Schematic of Univariate analysis of microbiome and medication

To delineate the association between drug structure and gut microbiome, an analytical approach was undertaken. Initially, univariate analysis was conducted across three distinct datasets: MetaCardis, ElderMet, and NuAge. The glm2 function, a component of the R statistical environment (version 3.2.0), was employed for this purpose, as outlined in the documentation provided by the Comprehensive R Archive Network (CRAN).

Each dataset underwent simple linear regression, treating the species column as the response variable and the drugs column as the predictor variable. This method facilitated the computation of effect size, p-values, and q-values, crucial metrics for assessing the significance and magnitude of associations. Given the utilization of three distinct datasets, three matrices encompassing effect sizes, p-values, and q-values were derived. Furthermore, a similar analytical framework was adopted in the context of the Japanese 4D dataset, and these results were seamlessly integrated into subsequent phases of analysis.

In continuation of the preceding analysis, it is imperative to expound upon fundamental statistical concepts utilized therein, namely effect size and q value, alongside elucidating the utilization of the sumlog function.

Effect size is a pivotal metric in statistical analysis, denoting the magnitude of the relationship between variables under consideration. Typically, effect size is computed utilizing various measures depending on the nature of the analysis, with common metrics including Cohen's d which is used for comparing mean values, odds ratios used for categorical data or Pearson's r used for correlation,.

Furthermore, the q value also referred to as the false discovery rate (FDR), represents an adjusted p-value that accounts for multiple hypothesis testing. The q value is computed via methods such as the Benjamini-Hochberg procedure, which controls the FDR at a predetermined threshold. The formula for calculating the q value entails ordering the p values obtained from multiple tests and applying a specified correction method to ascertain the adjusted significance level.

$$\text{q value} = p \times \frac{R}{N} \tag{2.1}$$

Where:

- p represents the p value obtained from individual tests.
- N denotes the total number of tests conducted.
- R signifies the rank of the p value within the ordered list of all p values.

During merging the results across datasets, the effect size matrices are concatenated column-wise, facilitating a comprehensive overview of the effect sizes observed across diverse datasets. Additionally, the consolidation of p values is accomplished by employing the sumlog function from the metap library in R, as documented in the Comprehensive R Archive Network (CRAN). This function aggregates p values utilizing a logarithmic transformation, followed by summation, to yield a unified metric representative of the combined statistical significance across multiple analyses.

"It is pertinent to emphasize that the p value denotes the probability of obtaining specific results under the assumption of a given hypothesis, rather than indicating the likelihood of the hypothesis being true based solely on the observed results" [11]. Consequently, the q value, as an adjusted p-value, serves to mitigate the risk of false positives in the context of multiple hypothesis testing, thereby enhancing the robustness of statistical inference.

A custom function was devised to compute the aggregate effect size based on the provided p-value and q-value for each drug-species pair. The function adheres to a

systematic approach, first evaluating the significance of the association by assessing whether the p-value corresponding to the drug-species pair falls below a predetermined threshold, typically set at 0.1.

If the p-value surpasses this threshold, indicating insufficient evidence to establish a meaningful relationship, the effect size is defaulted to zero. Conversely, if the p-value satisfies the criterion, the function proceeds to ascertain the direction and magnitude of the effect size across multiple datasets.

Given the consideration of four distinct datasets, potential scenarios arise where an equal number of datasets exhibit positive and negative effect sizes, thereby necessitating a tie-breaking mechanism. To resolve such instances, a majority voting system is implemented, where a consensus is reached based on the prevailing sign of the effect size across datasets. Specifically, if three datasets concur on either a positive or negative effect size, a respective value of +3 or -3 is assigned to the drug-species pair, indicating a robust majority consensus.

In cases of parity, where two datasets manifest positive effects while the remaining two demonstrate negative effects, a more nuanced approach is adopted. The average of the positive effect sizes is compared against the absolute average of the effect sizes that are negative. If the average of the effect sizes that are positive surpasses the absolute average of the negative counterparts, a value of +2 is assigned, signifying a discernible trend toward positivity. Conversely, if the absolute average of the negative effect sizes predominates, a value of -2 is allocated, indicative of a prevailing negative trend.

Thus, the devised function systematically integrates statistical evidence from multiple datasets, leveraging a judicious combination of majority voting and comparative analysis to derive a comprehensive and nuanced assessment of the aggregate effect size for each drug-species pair. A Schematic of the above-described procedure is given in the [Figure 2.2]

2.2.2 Computation of unifrac distance for of species that are positively and negatively associated with medications.

Following the initial analysis of the combined dataframe (combined_df), further investigation is conducted to elucidate microbiome modules predicated on drug structure. This involves the creation of two distinct dataframes derived from "combined_df", each delineating associations of differing polarities.

The first dataframe encapsulates positive associations by retaining solely positive values from "combined_df", while setting negative values to zero. Conversely, the second dataframe isolates negative associations by preserving negative values and nullifying positive values.

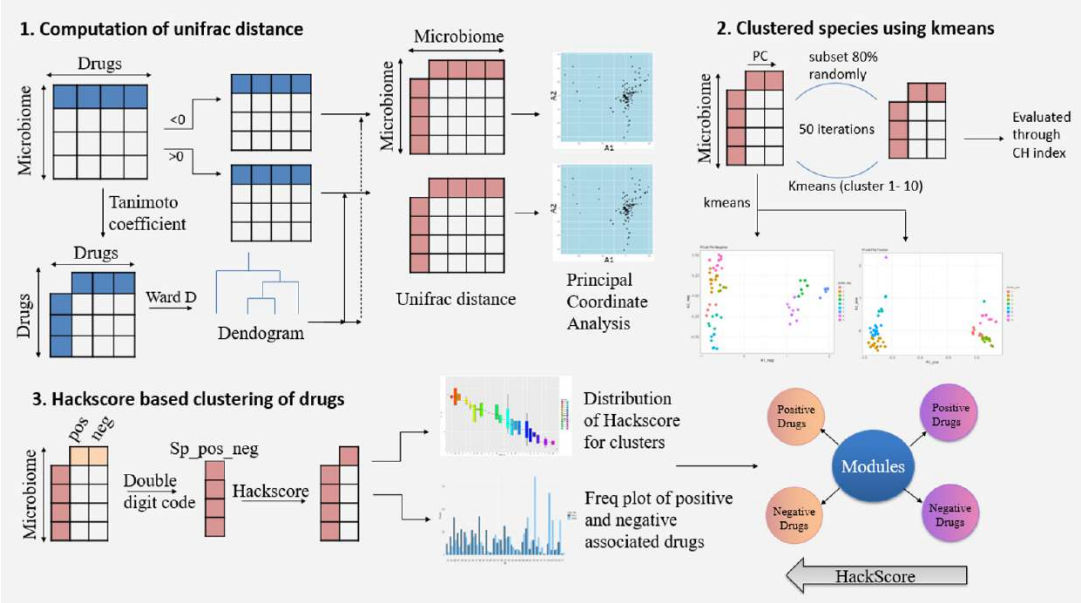

Figure 2.3: Schematic of Association between drug structure and gut microbiome

Subsequently, the drug names extracted from "combined_df" serve as pivotal entities for additional analysis. Leveraging web scraping techniques, SMILES (Simplified Molecular Input Line Entry System) representations are procured from the PubChem database. SMILES notations provide a compact textual representation of molecular structures, facilitating computational analyses.

To assess structural similarities between drugs, the RDKit library in Python is harnessed. Within this framework, the Tanimoto coefficient was computed using TanimotoSimilarity function that emerges as a key metric for quantifying molecular similarity. This coefficient measures the degree of resemblance of two molecules structurally based on their molecular fingerprints.

The Tanimoto coefficient (T) is computed using the formula:

$$T = \frac{A + B - C}{C} \tag{2.2}$$

Where:

- A represents the number of bits or features set in the fingerprint of molecule A.

- B signifies the number of bits or features set in the fingerprint of molecule B.
- C denotes the number of common bits or features set in both fingerprints of B and A molecules.

From this tanimoto coefficient a dendogram was built using hclust function for both dataframes. The resulting coefficient ranges from 0 to 1, with higher values denoting increased structural resemblance among the molecules. Tanimoto coefficient of 1 implies identical molecular structures, while a coefficient of 0 signifies complete dissimilarity. Utilizing this metric facilitates the identification of drug pairs exhibiting significant structural homology, thereby elucidating potential associations with microbiome modules.

The "ward.D" method, utilized in hierarchical clustering (hclust) in R, employs the Ward's minimum variance criterion to merge clusters iteratively. This approach seeks to reduce the overall variance within clusters while merging them, effectively optimizing the homogeneity of clusters while maximizing the separation between them.

In detail, the "ward.D" method calculates the increase in total within-cluster variance that results from merging two clusters. It chooses the merge that minimizes this increase, thus favoring compact and spherical clusters. The algorithm iteratively merges clusters until a specified stopping criterion is met, resulting in a hierarchical clustering dendrogram that visually represents the relationship between data points.

The Unifrac distance, as computed by the GUniFrac, library in R, is a metric used to quantify dissimilarities between microbial communities. It is particularly relevant in microbiome studies where understanding the diversity and structure of microbial communities is essential.

The Unifrac distance measures the fraction of branch length that is unique to one of the two microbial communities being compared within a phylogenetic tree. It considers the presence and absence of taxa in the communities, as well as the evolutionary relationships between taxa. The formula for computing Unifrac distance is as follows:

The UniFrac distance between communities A and B is computed using the formula:

$$\text{UniFrac}(A, B) = \frac{\text{Sum of branch lengths unique to A + Sum of branch lengths unique to B}}{\text{Total branch lengths in phylogenetic tree}} \tag{2.3}$$

Where:

- A and B represent two microbial communities being compared.
- The numerator sums the lengths of branches in the phylogenetic tree that are unique to either community A or community B.

- The denominator represents the total branch lengths in the phylogenetic tree.

Applying Unifrac distance to the dendrogram derived from drug association data captures the dissimilarities in microbiome composition associated with different drug structures. The Unifrac distance metric quantifies the extent of divergence in microbial community composition induced by positive associations between drugs and microbiome modules. A lower Unifrac distance suggests greater similarity in microbiome composition, indicating potential shared ecological niches or functional roles among microbial communities influenced by structurally similar drugs. Conversely, a higher Unifrac distance signifies greater dissimilarity, indicative of distinct microbial community compositions driven by structurally dissimilar drugs. This information aids in deciphering the impact of drug structure on the gut microbiome's abundance and its diversity, offering insights into potential therapeutic or adverse effects on host-microbiome interactions.

2.2.3 Clustering of Gut Microbiome into Modules Based on the Structure of Drugs with which they are Associated: Implications for Understanding Microbiome-Drug Interactions

Principal Coordinate Analysis (PCoA) was executed on the two Unifrac distance matrices separately, employing the dudi.pco function from the ade4 package in R. PCoA is a multivariate statistical technique utilized to visualize and explore dissimilarities in data points, particularly useful in microbiome studies to discern patterns of microbial community structure and composition.

To identify the most suitable cluster count within the PCoA-transformed dataset, we performed k-means clustering. This unsupervised machine learning algorithm divides data points into k separate clusters according to their similarities. Through iterative steps, the algorithm assigns each data point to the nearest cluster centroid and updates these centroids until convergence is reached.

The Calinski-Harabasz (CH) index functions as a measure to assess the quality of clustering. It evaluates the separation between clusters, measuring the distances between clusters relative to the compactness of each cluster. By considering both inter-cluster separation and cluster compactness, the CH index aids in determining the optimal number of clusters for a given dataset. The CH index is computed using the following formula:

The CH index is computed using the formula:

$$\text{CH index} = \frac{\text{Within-cluster sum of squares (WCSS)}}{\text{Between-cluster sum of squares (BCSS)}} \times \frac{k-1}{N-k} \tag{2.4}$$

Where:

- N denotes the total number of datapoints.
- k represents the number of clusters.
- BCSS corresponds to the sum of squared distances between cluster centroids and the overall centroid.
- WCSS signifies the sum of squared distances between data points and their respective cluster centroids.

A higher CH index value indicates better clustering, with improved separation between clusters and greater compactness within clusters.

For the implementation of k-means clustering, the dataset comprising PCoA components was randomly subsampled, encompassing 80% of the original dataset. Subsequently, k-means clustering was performed iteratively for cluster numbers ranging from 2 to 12, with a maximum of 100 iterations per clustering attempt. This random sampling procedure was repeated 50 times to ensure robustness and reliability of results. Finally, the CH index was computed for each clustering iteration using ClusterStability package in R, facilitating the construction of a boxplot to visualize the distribution of CH index values across different cluster numbers. The suitable cluster number was determined based on the highest Calinski Harabasz (CH) index values, with 8 clusters identified as suitable for positive associations and 9 clusters deemed appropriate for negative associations.

In the subsequent phase of analysis, each species within the dataset was assigned a unique tag denoting its respective cluster affiliation for positive and negative associations. This tag convention employed an underscore to delineate between the cluster numbers representing positive and negative associations, thus facilitating clear identification and differentiation. For instance, a species denoted as "SpeciesName_PosCluster _NegCluster" signifies its membership in both positive and negative clusters.

Subsequently, the dataset was examined to ascertain the prevalence and distribution of species sharing identical cluster tags (PosCluster_NegCluster). A comprehensive tally revealed the existence of 33 distinct groups, each comprising species with similar cluster affiliations.

To elucidate the relationship between drug associations and cluster groups, a grouped barplot was constructed. This graphical representation delineated the distribution of drugs exhibiting positive and negative associations within each group. Such visualization enables a nuanced understanding of the prevalence and distribution of drug effects across different microbial cluster affiliations.

2.2.4 Identification of Drug Modules Based on the Health-Associated Core Keystone (Hack) Score associated with the putatively good or bad gut microbiome

Moreover, leveraging recent advancements in microbiome research, emphasis was placed on evaluating the health implications of drug associations with specific microbial species. The Hack score, a metric developed and recently published by our laboratory, serves as a proxy for assessing the health-associated core keystone status of bacterial species. A greater Hack score signifies a favorable microbial composition, while a lower score suggests possible adverse impacts on the health of the host.

In alignment with this framework, drugs associated with species exhibiting the highest Hack scores are inferred to be detrimental, as they impede the growth and proliferation of beneficial bacterial species. To elucidate these associations further, boxplots depicting Hack scores for drugs associated with each grouped cluster were generated. This process entailed plotting a total of 33 boxplots, each representing the distribution of Hack scores for drugs within a specific cluster group.

In summary, this multifaceted analysis integrates cluster-based associations, drug effects, and microbiome health assessments, thereby offering comprehensive insights into the interplay between drug exposure and microbial community dynamics. Such insights are invaluable for discerning the potential health implications of pharmacological interventions on host-microbiome interactions. Schematic of this process is given in [Figure 2.3]

2.2.5 Identification of association of species with drugs influenced by Cohort lifestyle

In the second part of the thesis, the focus shifts to identifying associations between microbial species and drugs influenced by cohort lifestyle. To achieve this, datasets from 72 studies were curated, each representing distinct cohort populations and their respective microbial compositions. These studies encompass a broad spectrum of research endeavors, providing a diverse and comprehensive dataset for subsequent analysis.

Subsequently, the mean abundance of microbial species was computed for each study, resulting in a structured dataframe delineating study names against microbial species. Principal Coordinate Analysis (PCoA) was then employed to elucidate the underlying patterns and variability within the microbial community composition across different cohorts. Through visualization in scatter plots, the principal components capturing the maximum variance of specific cohort lifestyles were identified.

Notably, the cohort lifestyles under consideration were classified into three distinct

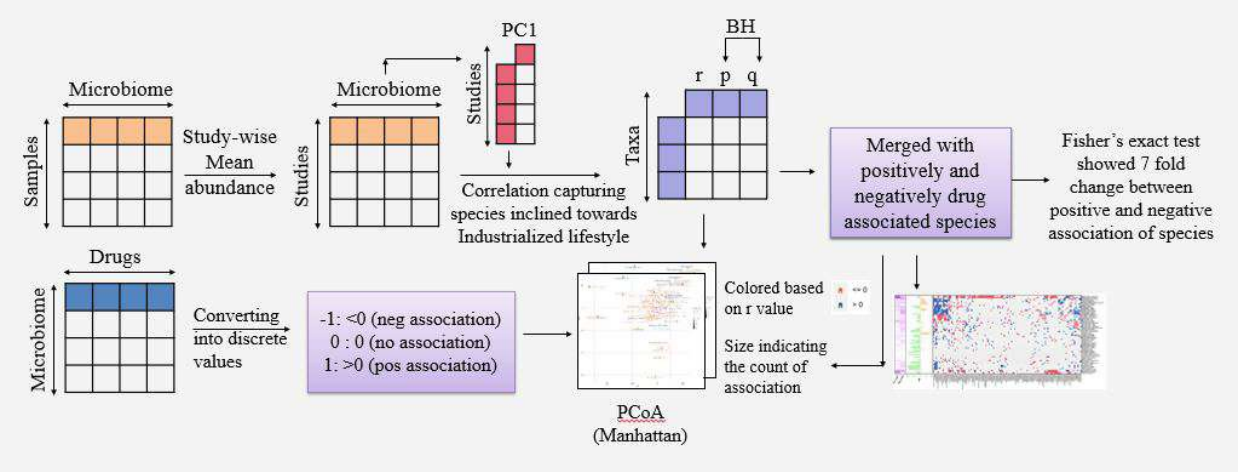

Figure 2.4: Schematic of Association of species with drugs influenced by Cohort lifestyle

categories: Industrialized, Rural Tribal, and Rural Urban Mixed. Upon analysis, it was observed that PC1 predominantly captured non-industrialized lifestyles, encompassing both Rural Tribal and Rural Urban Mixed cohorts.

Further investigation entailed correlating PC1 values with the mean abundance of microbial species. This correlation analysis aimed to discern potential relationships between the principal component representing cohort lifestyle and the microbial composition. The correlation coefficients (r), along with their associated p-values and q-values, were computed utilizing the corr.test function from the psych package in R.

The utilization of correlation analysis facilitated the identification of statistically significant associations between cohort lifestyle represented by PC1 and the abundance of microbial species. By delineating these associations, insights into the impact of cohort lifestyle on microbial composition and potential implications for drug-microbiome interactions can be gleaned. Such insights are invaluable for elucidating the intricate dynamics underlying host-microbiome relationships and may inform personalized therapeutic interventions tailored to specific cohort demographics.

In continuation of the aforementioned analysis, the results dataframe, comprising species as rows and correlation coefficients (r), p-values, and q-values as columns, was

augmented with additional information pertaining to the count of drugs positively and negatively associated with each species. This augmentation aimed to enrich the dataset with contextual information regarding the drug-microbiome associations under investigation.

Subsequently, to ascertain the significance of the observed disparity in the counts of positively and negatively associated drugs, Fisher's exact test was employed using fisher.test function in R. Fisher's exact test is a statistical technique employed to assess the significance of the relationship between two categorical variables, especially in scenarios with small sample sizes or when the chi-squared test assumptions are not fulfilled.

The test is particularly relevant in scenarios where binary outcomes are compared across different groups or categories. It calculates the probability of obtaining the observed distribution of outcomes, or more extreme distributions, under the assumption of independence between the variables. This probability, known as the p-value, quantifies the strength of evidence against the null hypothesis of independence.

For example, consider a hypothetical scenario where the contingency table might look like this:

	Group 1 (Positive)	Group 2 (Negative)
Category 1	a	b
Category 2	c	d

The formula for Fisher's exact test is:

$$p = \frac{\binom{a+b}{a} \cdot \binom{c+d}{c}}{\binom{n}{a}\binom{n}{c}} \tag{2.5}$$

Where:

- a is the count of observations in Category 1 (e.g., positively associated drugs).
- b is the count of observations in Group 1 but not in Category 1 (e.g., negatively associated drugs).
- c is the count of observations in Category 2 (e.g., negatively associated drugs).
- d is the count of observations in Group 2 but not in Category 2 (e.g., positively associated drugs).
- n is the total number of observations (i.e., $a + b + c + d$).

- $\binom{n}{k}$ denotes the binomial coefficient, also known as "n choose k", representing the number of combinations of k elements chosen from a set of n elements.

The resulting p-value quantifies the probability of observing the observed distribution of outcomes, or a more extreme distribution, assuming independence between the two categorical variables. A small p-value suggests that the observed association is unlikely to have occurred by chance alone, indicating statistical significance.

In continuation of the analysis, the dataframe `combined_df` underwent discretization to facilitate a categorical representation of the data. This discretization involved assigning a value of -1 to entries less than zero, a value of +1 to entries greater than zero, and a value of zero to entries equal to zero. This transformation allowed for the representation of the data in a format conducive to subsequent analyses.

Subsequently, Manhattan distance was computed for this discretized dataframe. The Manhattan distance, alternatively known as city block distance or L1 distance, is a metric used to measure the distance between two points in a multidimensional space. Unlike the Euclidean distance, which computes the straight-line distance between points, the Manhattan distance is determined by summing the absolute differences between the corresponding coordinates of the points.

For 2 points (x_1, y_1) and (x_2, y_2) in a two-dimensional space, the formula for Manhattan distance is:

$$\text{Manhattan distance} = |x_2 - x_1| + |y_2 - y_1| \tag{2.6}$$

Manhattan distance is particularly suitable for our case due to its robustness to outliers and its ability to handle data with discrete or categorical features effectively. In the context of microbial species association analysis with drugs, the discretized representation of the data aligns well with the nature of microbiome composition data, where the presence or absence of species abundance is of primary interest.

Subsequent to the computation of Manhattan distance, Principal Coordinate Analysis (PCoA) was performed on this distance matrix using the dudi.pco function in R, implemented under the ade4 package. PCoA is a multivariate statistical technique utilized to visualize and explore dissimilarities in data points, with the resulting principal coordinates capturing the underlying patterns and variability within the dataset.

In the PCoA plot, coloration was applied based on the sign of the previously computed correlation coefficients (r values), distinguishing between positive and negative associations. Additionally, the size of each data point in the PCoA plot was determined by the number of drugs positively associated with the corresponding microbial species. A separate PCoA plot was generated using the same methodology, with dot size now indicating the number of drugs negatively associated with each microbial species.

Principal Coordinate Analysis (PCoA), also known as classical multidimensional scaling (MDS), is a dimensionality reduction technique commonly employed in exploratory data analysis to visualize and interpret patterns of dissimilarity or similarity within a dataset. PCoA operates on distance or dissimilarity matrices, transforming high-dimensional data into a lower-dimensional space while preserving the pairwise distances between data points as much as possible.

The fundamental principle underlying PCoA is to represent the original data in a new coordinate system, such that the first few principal coordinates capture the maximum variance in the dataset. This transformation enables the visualization of complex relationships among data points in a lower-dimensional space, facilitating the identification of underlying structures or clusters.

PCoA accomplishes this transformation through eigenvalue decomposition of the distance matrix, yielding a set of eigenvectors (principal coordinates) and corresponding eigenvalues. The principal coordinates represent the axes of the new coordinate system, while the eigenvalues quantify the amount of variance explained by each principal coordinate.

The PCoA plot visualizes the data points in the new coordinate system, typically with two or three principal coordinates representing the axes. Data points that are closer together in the PCoA plot are more similar to each other in terms of their original multidimensional representation, while those that are farther apart are more dissimilar.

PCoA is particularly useful in microbiome studies, where high-dimensional microbial community data are often analyzed to elucidate patterns of diversity, composition, and ecological relationships. By visualizing the relationships among microbial communities or species based on their dissimilarities, PCoA aids in identifying factors driving microbial community structure, such as environmental gradients, disease states, or treatment effects.

In summary, PCoA is a powerful tool for visualizing and interpreting complex patterns of dissimilarity or similarity within multidimensional datasets. Its application in microbiome research enables researchers to gain insights into the structure and dynamics of microbial communities, with potential implications for understanding host-microbiome interactions and informing clinical interventions.

Through this comprehensive analysis, insights into the spatial arrangement of microbial species associations with drugs, as well as their differential impacts based on drug positivity and negativity, can be gleaned. Such visualizations provide valuable information for understanding the complex interplay between microbial community composition and drug influences, with potential implications for therapeutic interventions and personalized medicine approaches. Schematic of this process is given in [Figure 2.4]

2.2.6 Identification of association of species with drugs influenced by Age

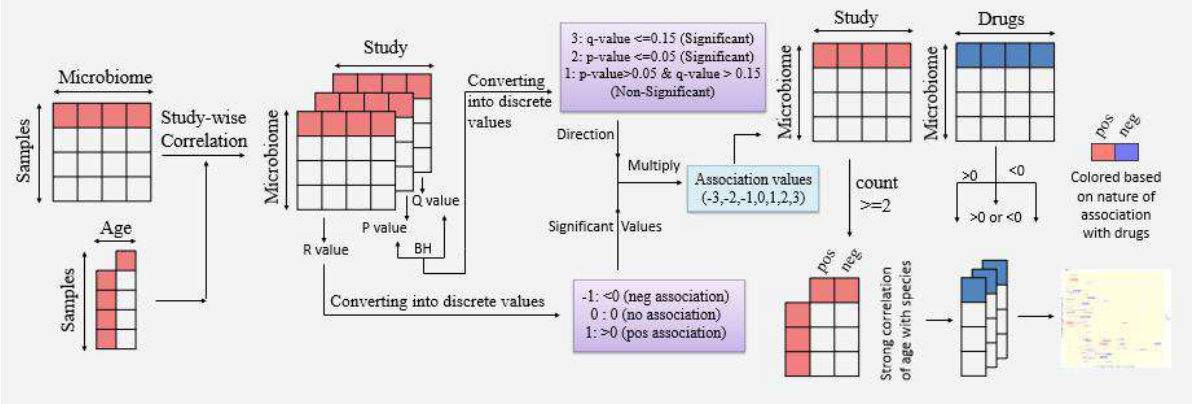

Figure 2.5: Schematic of association of species with drugs influenced by age

The primary objective of this section is to discern potential associations between microbial species and drugs, taking into account the influence of age as a crucial factor. The analysis conducted herein leverages data extracted from 35 distinct studies, each characterized by subjects falling within the age bracket of 18 to 60 years.

At the outset, a comprehensive examination was undertaken, correlating the abundance of microbial species observed in each study with the age of the corresponding patients. This correlation analysis, employing Pearson correlation coefficients, aimed to elucidate any potential relationships between microbial community composition and age. By computing correlation coefficients alongside associated p-values and q-values, derived through the rigorous Benjamin-Hochberg method to account for multiple testing, a nuanced understanding of age-associated microbial shifts was attained.

To further elucidate the significance of these correlations, a discretization approach was adopted. This involved transforming continuous p and q values into discrete categories, thereby facilitating a more interpretable assessment of statistical significance. Specifically, q-values below the threshold of 0.15 were categorized as highly significant (assigned a value of 3), while p-values below 0.05 were deemed statistically significant (assigned a value of 2); all other values were assigned a value of 1. Multiplying these discrete values by their corresponding correlation coefficients allowed for the determination of the directionality of the age-species associations.

The subsequent step entailed a meticulous examination of the identified directionalities to discern age-related associations with microbial species. Directionalities exhibiting a magnitude greater than or equal to 2 were considered indicative of robust associations. By tallying the total number of directionalities meeting this criterion, the prevalence of positive age and species associations (in cases where the r-value was negative) and negative age and species associations (where the r-value was less than or equal to -2) was quantified for each microbial species under investigation.

Integrating these findings with information regarding drug associations, the counts of drugs positively and negatively associated with each microbial species were merged. This allowed for a comprehensive assessment of potential relationships between drug effects and age-related microbial shifts.

To provide a visual representation of the interplay between drug associations and age-related microbial shifts, a scatter plot was generated. In this plot, the x-axis depicted the number of positively associated drugs and positive age associations, while the y-axis represented the number of negatively associated drugs and negative age associations. By color-coding species based on their respective associations (blue for positive drug and positive age associations, and red for negative drug and negative age associations), patterns of co-occurrence and potential synergies or antagonisms between drug effects and age-related microbial shifts were discerned. The schematic of this process is given in [Figure 2.5]

Chapter 3

RESULTS

3.1 Results of the drug structure and gut microbiome association analysis

SMILES were obtained for all drugs included in the MetaCardis, ElderMet, NuAge, and Japanese 4D studies. From these SMILES, Tanimoto coefficients were calculated to assess structural similarity between drug molecules. The coefficients, ranging from 0 to 1, indicate the degree of structural resemblance, with higher values suggesting greater similarity. A Tanimoto coefficient of 1 denotes identical molecular structures, while 0 signifies complete dissimilarity. This metric aids in identifying drug pairs with significant structural homology, providing insights into potential associations with microbiome modules. Subsequently, dendrograms were constructed based on these coefficients, as depicted in Figure 3.1.

The application of Unifrac distance to the dendrogram derived from drug association data facilitates the identification of dissimilarities in microbiome composition linked to various drug structures. This distance metric quantifies the extent of divergence in microbial community composition resulting from positive associations between drugs and microbiome modules. A lower Unifrac distance suggests greater similarity in microbiome composition, indicative of potential shared ecological niches or functional roles among microbial communities influenced by structurally similar drugs. Conversely, a higher Unifrac distance signifies greater dissimilarity, reflecting distinct microbial community compositions driven by structurally dissimilar drugs. This information enhances our understanding of the impact of drug structure on gut microbiota composition and diversity, shedding light on potential therapeutic or adverse effects on host-microbiome interactions.

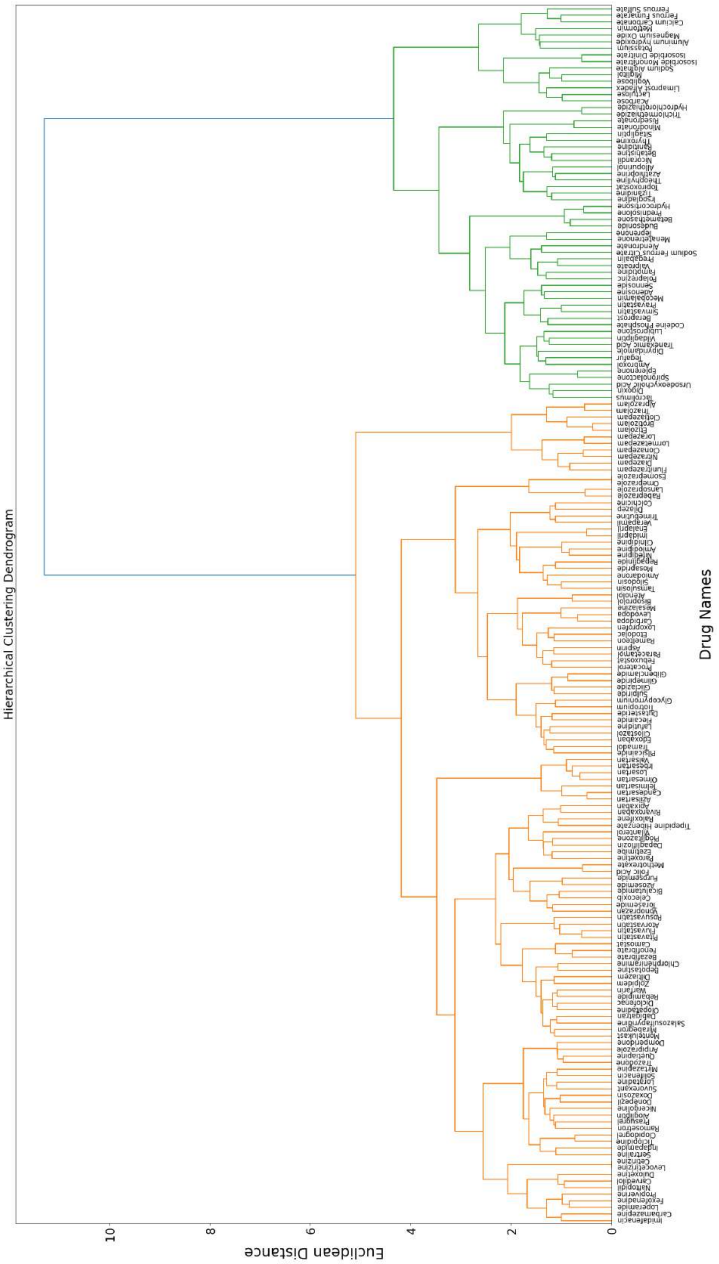

Figure 3.1: Dendrogram of Tanimoto coefficient.

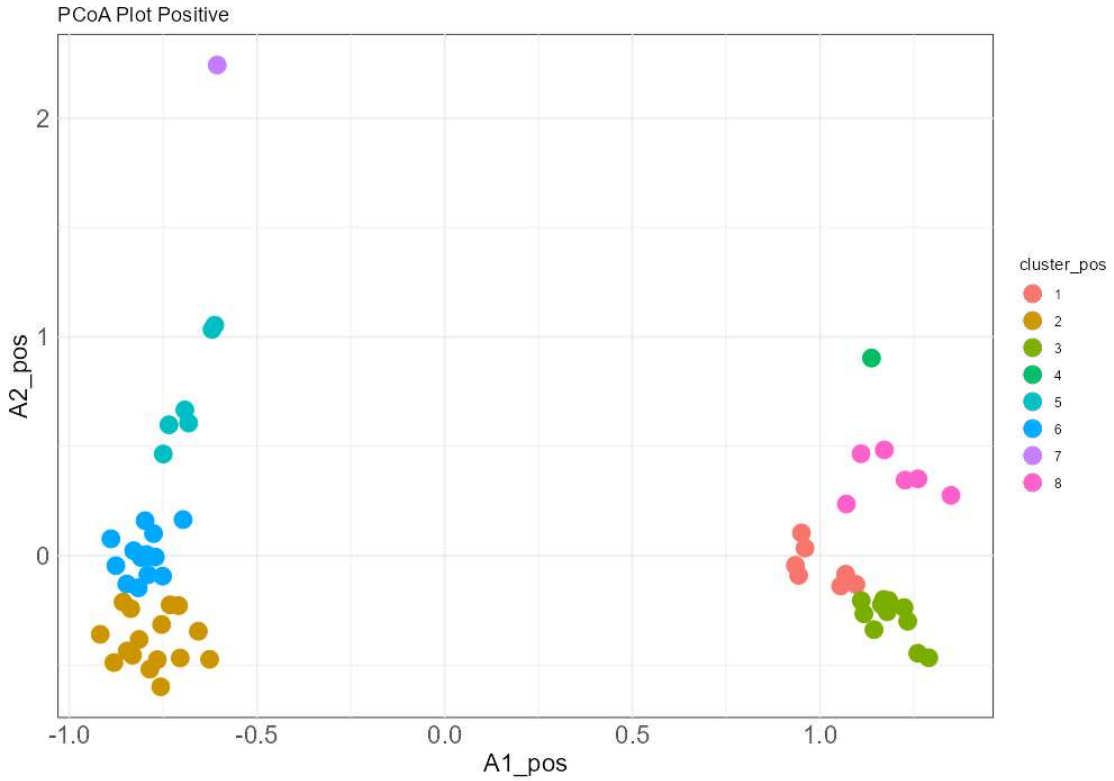

Figure 3.2: Principal Coordinate of Analysis (PCoA) of unifrac distance of gut microbiome that is positively associated with drugs.

To further explore these associations, Principal Coordinate Analysis (PCoA) was conducted on the two Unifrac distance matrices separately using the dudi.pco function from the ade4 package in R. PCoA, a multivariate statistical technique, enables the visualization and exploration of dissimilarities in data points. This analysis, depicted in Figure 3.2 and Figure 3.3 for positive and negative drug associations respectively, aids in elucidating patterns of microbial community structure and composition influenced by drug associations with the gut microbiome.

To ascertain the most suitable number of clusters within the PCoA-transformed dataset, k-means clustering was employed. K-means clustering, an unsupervised learning technique in machine learning, categorizes data points into separate clusters according to their likeness. This iterative process involves assigning each data point to its closest cluster centroid and adjusting the centroids until convergence occurs.

The Calinski-Harabasz (CH) index was utilized as a metric to assess the quality of

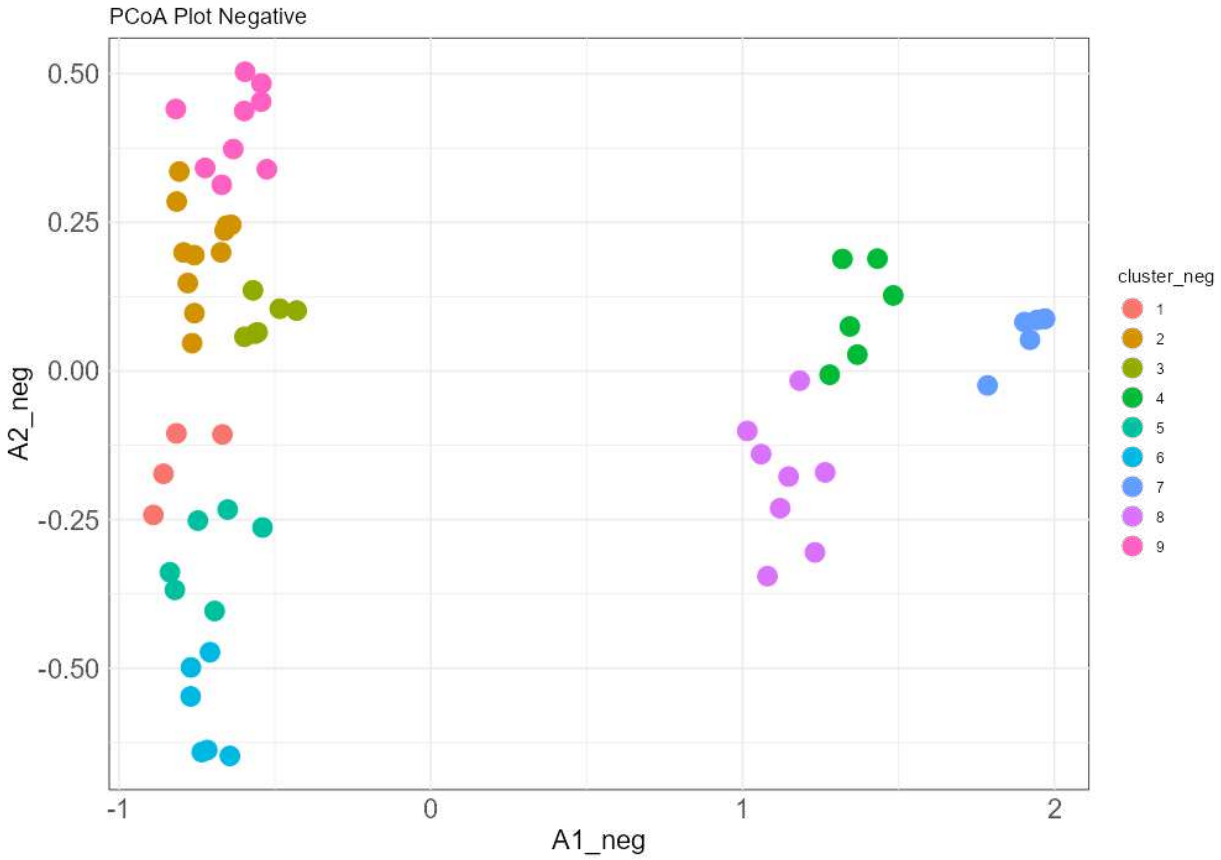

Figure 3.3: PCoA of unifrac distance of gut microbiome that is negatively associated with drugs.

clustering solutions. A higher CH index value indicates better clustering, with improved separation between clusters and greater compactness within clusters.

For the implementation of k-means clustering, the dataset containing PCoA components was randomly subsampled, covering 80% of the original dataset. Subsequently, k-means clustering was iteratively performed for cluster numbers ranging from 2 to 12, with a maximum of 100 iterations per clustering attempt. This random sampling procedure was repeated 50 times to ensure the robustness and reliability of the results.

Finally, the CH index was computed for each clustering iteration using the ClusterStability package in R. A boxplot was generated to visualize the distribution of CH index values across different cluster numbers. The optimal number of clusters was determined based on the highest CH index values, with 8 clusters identified as suitable

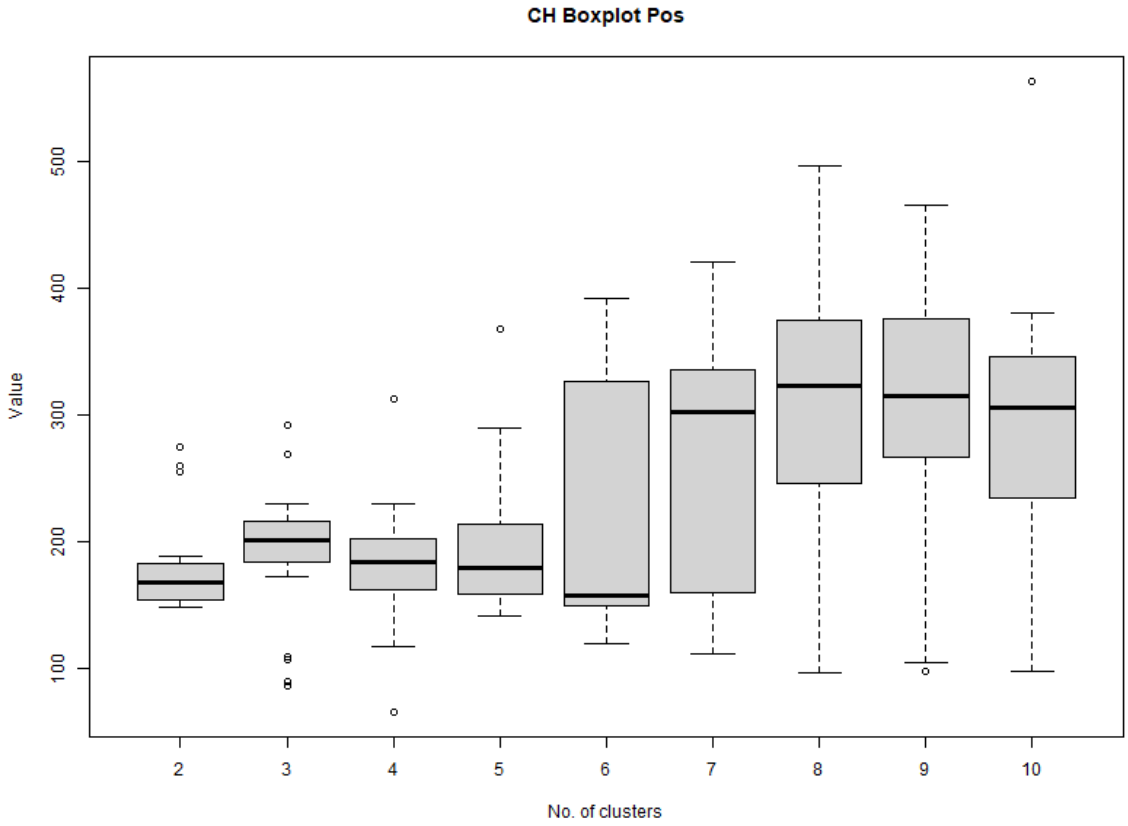

Figure 3.4: These boxplots display the distribution of Calinski-Harabasz (CH) index values across different cluster numbers for the positively associated gut microbiome with drug structure. Higher values indicate better separation between clusters and greater compactness within clusters. In this instance, the highest CH index value corresponds to a cluster number of 8, indicating that this number of clusters provides the most cohesive and distinct grouping of samples based on their gut microbiome composition and its association with drug structure.

for positive associations and 9 clusters deemed appropriate for negative associations. Refer figure 3.4 and 3.5. Final clusters is shown in figure 3.12 and 3.11

In the subsequent phase of analysis, each species within the dataset received a unique tag indicating its cluster affiliation for both positive and negative associations. This tagging convention utilized underscores to distinguish between cluster numbers representing positive and negative associations, facilitating clear identification and dif-

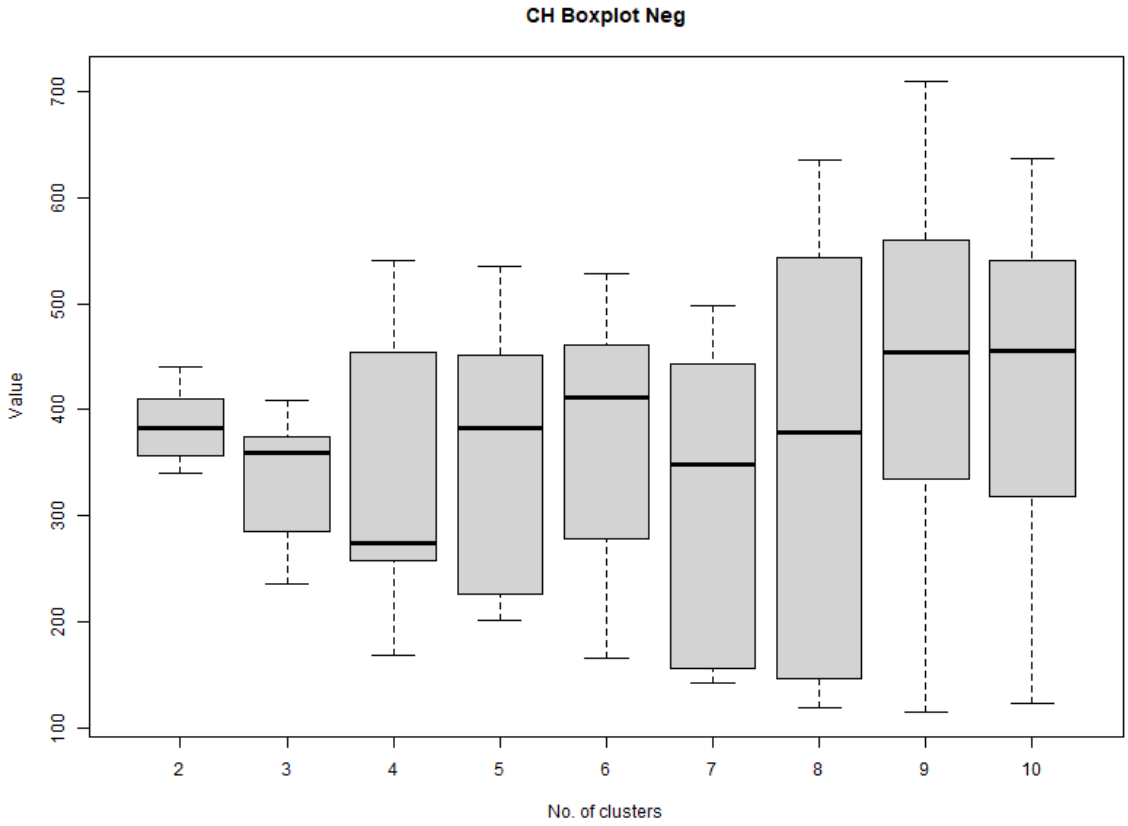

Figure 3.5: These boxplots display the distribution of Calinski-Harabasz (CH) index values across different cluster numbers for the negatively associated gut microbiome with drug structure. Higher values indicate better separation between clusters and greater compactness within clusters. In this instance, the highest CH index value corresponds to a cluster number of 9, indicating that this number of clusters provides the most cohesive and distinct grouping of samples based on their gut microbiome composition and its association with drug structure.

ferentiation. For example, a species labeled as "SpeciesName_PosCluster_NegCluster" indicates membership in both positive and negative clusters.

Subsequently, the dataset was scrutinized to determine the prevalence and distribution of species sharing identical cluster tags (PosCluster_NegCluster). A thorough tally revealed the presence of 33 distinct groups, each consisting of species with similar cluster affiliations.

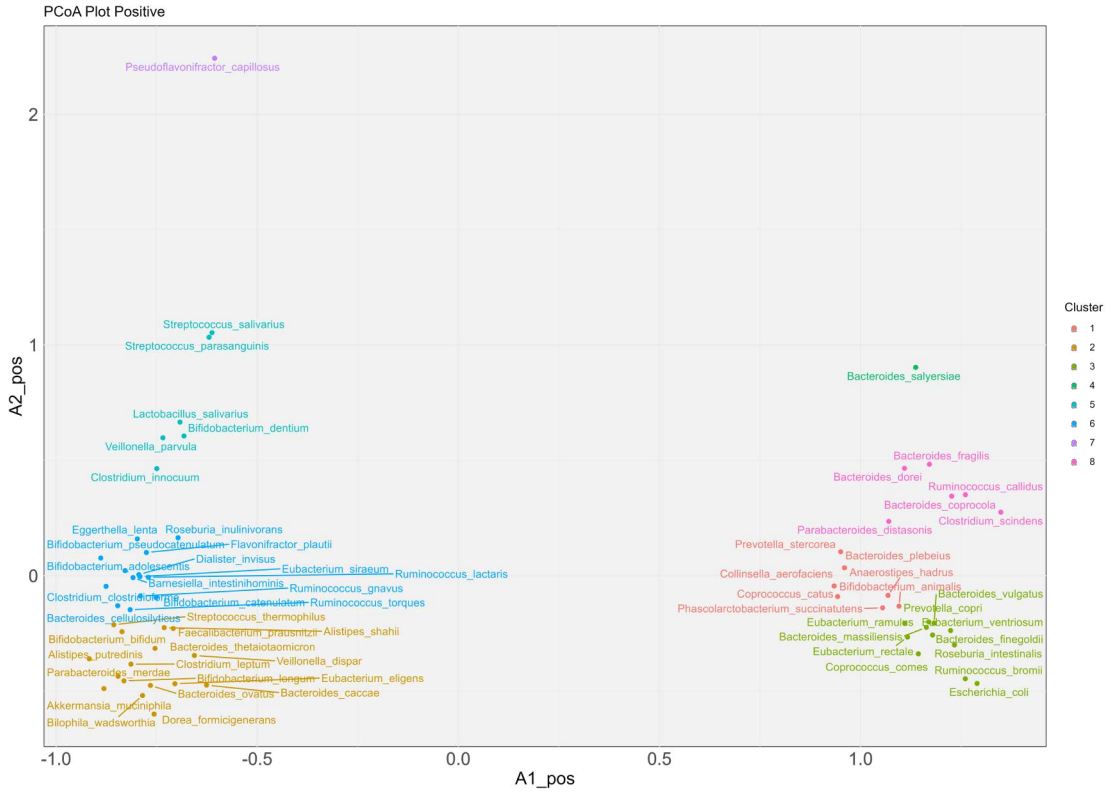

Figure 3.6: This Principal Coordinate Analysis (PCoA) plot illustrates the positive association of gut microbiome composition with drug structure. Each point in the plot represents a bacterial species and is coloured according to the cluster number identified through k-means clustering. The plot reveals distinct clusters of samples, indicating similarities in gut microbiome composition among individuals exposed to structurally similar drugs. This visualization provides insights into the potential impact of drug structure on gut microbiome composition and highlights patterns of microbial community structure influenced by drug associations.

To explore the relationship between drug associations and cluster groups, a grouped barplot was constructed. This visual representation delineated the distribution of drugs exhibiting positive and negative associations within each group. Such visualization provides insights into the prevalence and distribution of drug effects across different microbial cluster affiliations.

Furthermore, leveraging recent advancements in microbiome research, emphasis was placed on evaluating the health implications of drug associations with specific microbial

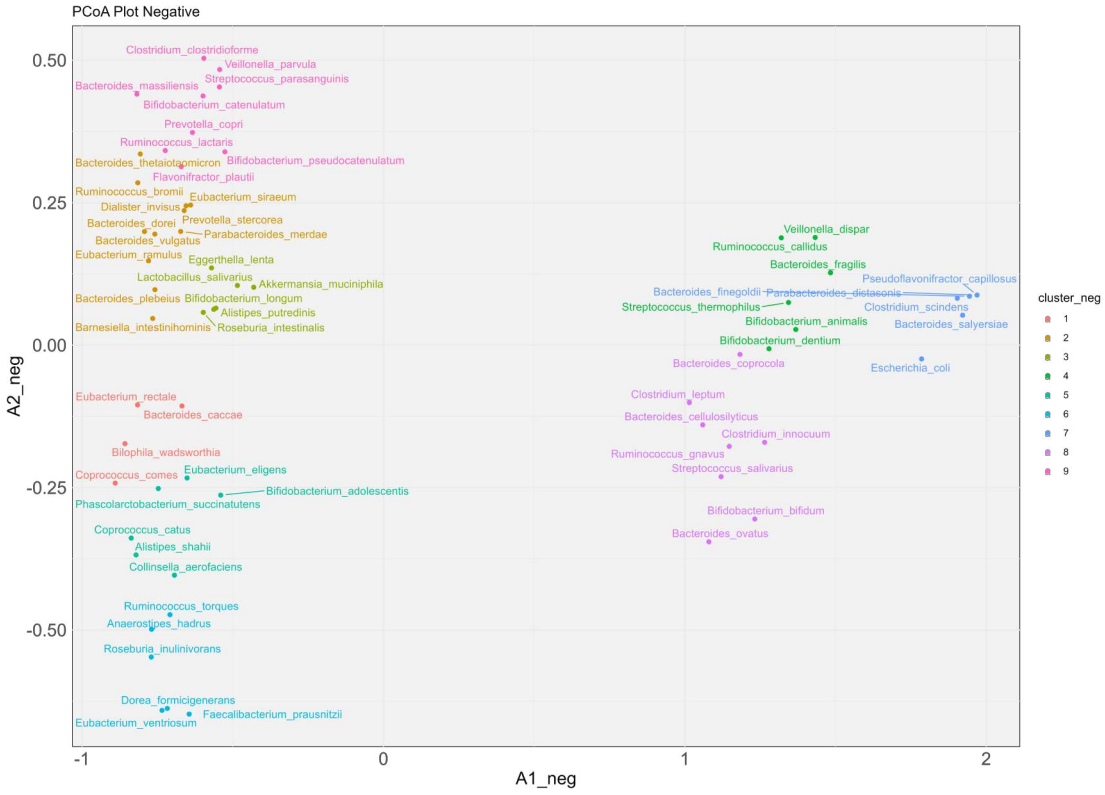

Figure 3.7: This Principal Coordinate Analysis (PCoA) plot illustrates the negative association of gut microbiome composition with drug structure. Each point in the plot represents a bacterial species and is coloured according to the cluster number identified through k-means clustering. The plot reveals distinct clusters of samples, indicating similarities in gut microbiome composition among individuals exposed to structurally similar drugs. This visualization provides insights into the potential impact of drug structure on gut microbiome composition and highlights patterns of microbial community structure influenced by drug associations.

species. The Hack score, a metric developed and recently published by our laboratory, serves as a proxy for assessing the health-associated core keystone status of bacterial species. A higher Hack score indicates a beneficial microbial profile, while a lower score suggests potential detrimental effects on host health.

Consistent with this framework, drugs associated with species exhibiting the highest Hack scores are inferred to be detrimental, as they inhibit the growth and proliferation of beneficial bacterial species. To further elucidate these associations, boxplots

depicting Hack scores for drugs associated with each grouped cluster were generated. This involved plotting a total of 33 boxplots, each representing the distribution of Hack scores for drugs within a specific cluster group. Refer figure 3.8 and 3.9

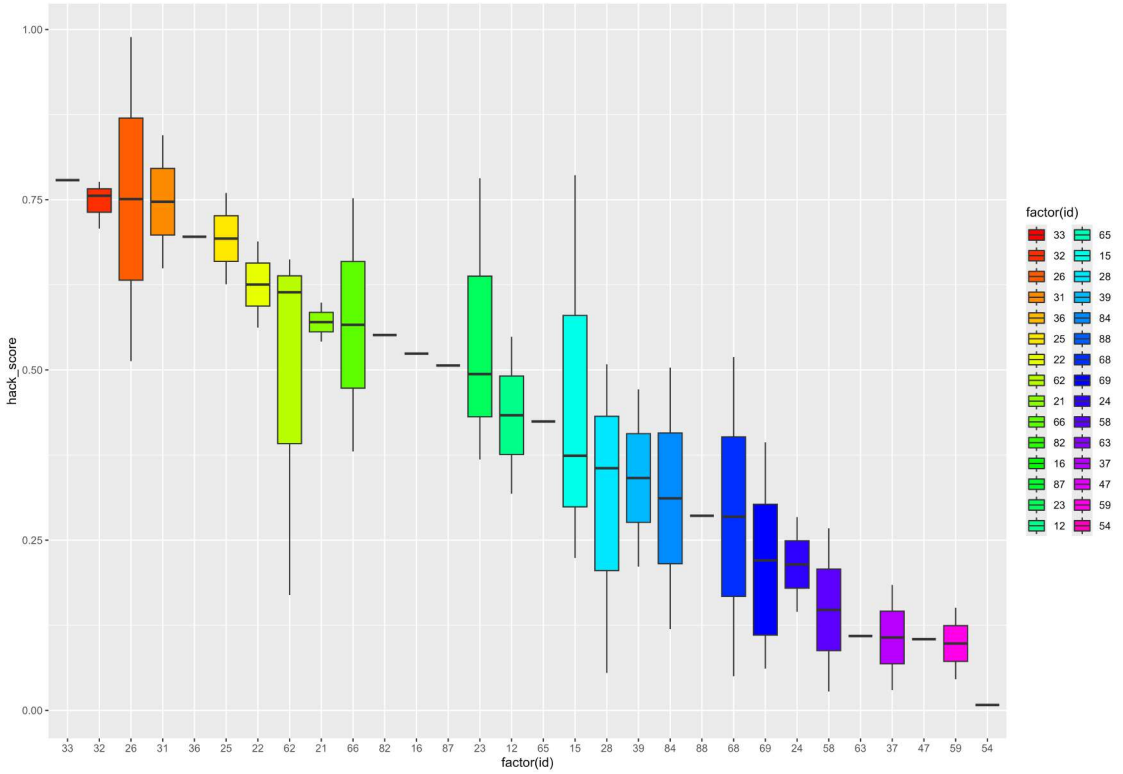

Figure 3.8: This boxplot illustrates the distribution of Hack scores for drugs associated with each of the 33 grouped clusters. The Hack score serves as a metric for assessing the health-associated core keystone status of bacterial species. Higher Hack scores indicate a beneficial microbial profile, while lower scores suggest potential detrimental effects on host health. The boxplot provides a visual depiction of the variability in Hack scores across different cluster groups, shedding light on the potential health implications of drug associations with specific microbial species.

Four Drug modules were formed on the basis of hack score. Drug module 1- Lowest hackscore and are positively associated with gut microbiome - Zolpidem, Sertraline, Pregabalin, Alprazolam, Etizolam, Flunitrazepam, Nitrazepam, Brotizolam, Valproate, Lansoprazole, Rabeprazole, Esomeprazole, Omeprazole, Losartan, Valsartan, Azilsartan, Candesartan, Olmesartan, Carvedilol, Enalapril, Furosemide, Hydrochlorothiazide,

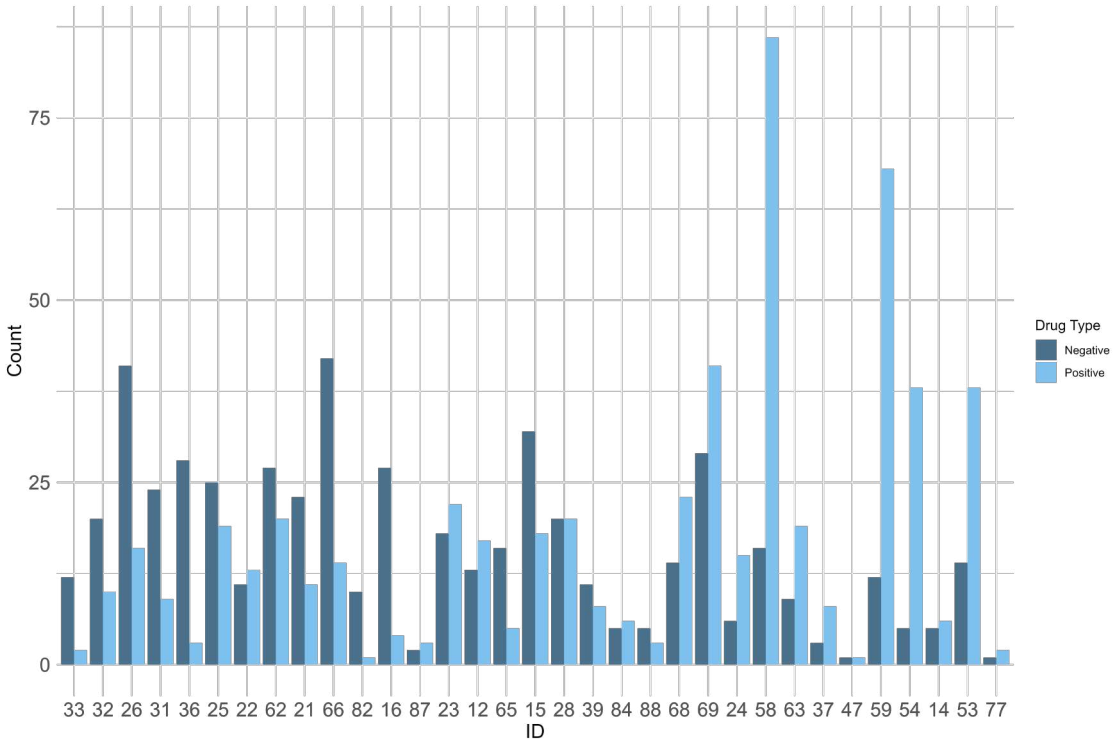

Figure 3.9: This grouped barplot illustrates the distribution of drugs exhibiting positive and negative associations within each of the 33 distinct cluster groups identified in the dataset. Each cluster group represents a collection of microbial species with similar cluster affiliations for positive and negative associations. The barplot provides a visual depiction of the prevalence and distribution of drug effects across different microbial cluster affiliations, offering insights into the relationship between drug associations and microbial community composition. This visualization aids in understanding how drugs impact specific microbial clusters and their potential implications for host-microbiome interactions and health outcomes.

Spironolactone, Torasemide, Nifedipine, Tamsulosin, Vildagliptin, Glimepiride, Gliclazide, Bicalutamide, Repaglinide, Bisoprolol, Doxazosin, Atorvastatin, Simvastatin, Fluvastatin, Rosuvastatin, Clopidogrel, Montelukast, Chlorpheniramine, Domperidone, Methotrexate, Mirabegron, Paracetamol, Aspirin, Ranitidine, Risedronate, Rivaroxaban, Silodosin, Solifenacin, Tramadol, Warfarin, Aripiprazole, Celecoxib, Nicorandil, Prednisolone, Vonoprazan, Polaprezinc, Sucralfate, Teprenone, Mosapride, Salazosul-

fapyridine, Azosemide, Isosorbide Dinitrate, Isosorbide Mononitrate, Camostat, Tegafur, Naftopidil, Loxoprofen, Minodronate, Procaterol, Ambroxol, Apixaban, Ursodeoxycholic Acid, Neurotropin, Alendronate, Duloxetine, Fenofibrate, Fexofenadine

Drug module 2- Lowest hackscore and are negatively associated with gut microbiome - Atorvastatin, Simvastatin, Pravastatin, Nicorandil, Aspirin, Fenofibrate, Flecainide, Atenolol, Amlodipine, Nifedipine, Amlodipine, Verapamil, Telmisartan, Furosemide, Acarbose, Metformin, Risedronate, Diltiazem, Lansoprazole

Drug module 3 - Highest hackscore and are positively associated with gut microbiome - Aspirin, Candesartan, Chlorpheniramine, Diltiazem, Fenofibrate, Gliclazide, Raloxifene, Aripiprazole, Carbidopa, Atenolol, Bisoprolol, Febuxostat, Metformin, Telmisartan, Olmesartan, Rosuvastatin, Pravastatin, Quetiapine, Rivaroxaban, Solifenacin, Trazodone, Valsartan, Acarbose, Budesonide, Dutasteride, Flecainide, Indapamide, Lansoprazole, Azosemide

Drug module 4 - Highest hack score and are positively associated with gut microbiome - Ramelteon, Zolpidem, Levodopa, Carbidopa, Alprazolam, Clonazepam, Pregabalin, Alendronate, Domperidone, Mesalazine, Methotrexate, Paracetamol, Thyroxine, Azathioprine, Prednisolone, Polaprezinc, Rebamipide, Sucralfate, Teprenone, Mosapride, Salazosulfapyridine, Ursodeoxycholic acid, Voglibose, Tranexamic Acid, Loxoprofen, Sulpiride, Amlodipine, Dipyridamole, Levocetirizine, Loperamide, Quetiapine, Rivaroxaban, Silodosin, Sitagliptin, Aripiprazole, Famotidine, Lafutidine, Sodium Alginate, Lubiprostone, Calcium Carbonate, Apixaban, Aspirin, Irbesartan, Lactulose, Losartan, Rabeprazole, Ranitidine, Spironolactone, Warfarin, Beraprost, Rosuvastatin, Pravastatin, Tizanidine, Pioglitazone, Acarbose, Miglitol, Metformin, Isosorbide Dinitrate.

3.2 Results of association of species with drugs influenced by Cohort lifestyle

The second part of the thesis focused on identifying associations between microbial species and drugs influenced by cohort lifestyle. Initially, datasets from 72 studies were compiled, each representing distinct cohort populations and their microbial compositions. Principal Coordinate Analysis (PCoA) was then employed to reveal underlying patterns and variability within the microbial community composition across different cohorts. Notably, PC1 predominantly captured non-industrialized lifestyles, such as Rural Tribal and Rural Urban Mixed cohorts.

Subsequent correlation analysis aimed to discern potential relationships between PC1 values representing cohort lifestyle and the abundance of microbial species. This

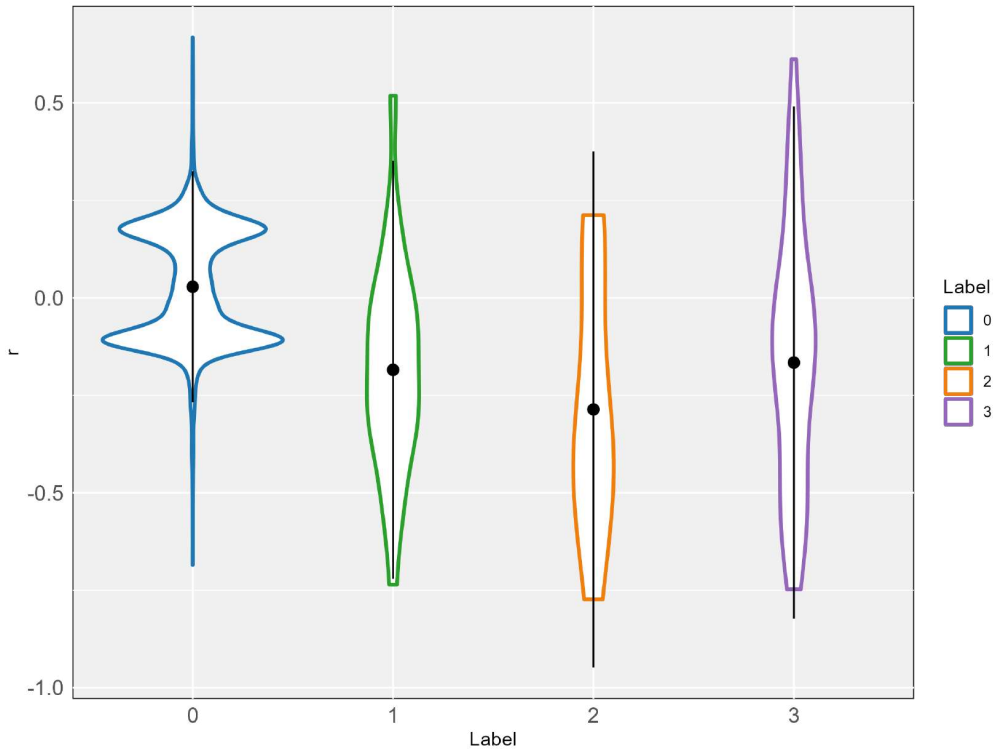

Figure 3.10: This violin plot illustrates the distribution of correlation coefficients (r) between cohort lifestyle represented by PC1 and the abundance of microbial species, categorized based on the presence or absence of positive and negative associations with drugs. The labels on the x-axis denote four distinct groups: Group 0 represents microbial species with no associations with either positive or negative drugs (pos_drug = 0 and neg_drug = 0); Group 1 consists of species positively associated with drugs but not negatively associated (pos_drug > 0 and neg_drug = 0); Group 2 comprises species negatively associated with drugs but not positively associated (pos_drug = 0 and neg_drug > 0); and Group 3 includes species with both positive and negative associations with drugs (pos_drug > 0 and neg_drug > 0). The violin plot provides a visual representation of the distribution and variability of correlation coefficients across these groups, offering insights into the relationship between cohort lifestyle, microbial composition, and drug associations.

analysis identified statistically significant associations between cohort lifestyle and microbial composition, providing insights into the impact of cohort lifestyle on the microbiome and potential implications for drug-microbiome interactions.

To enrich the dataset, additional information regarding the count of drugs positively and negatively associated with each microbial species was incorporated. This augmented dataset was then utilized to generate a violin plot categorizing microbial species based on their associations with positive and negative drugs. The violin plot visually depicts the distribution and variability of correlation coefficients across different groups of microbial species, offering insights into the relationship between cohort lifestyle, microbial composition, and drug associations. Refer figure 3.10.

In continuation of the analysis, the combined dataframe underwent discretization to enable a categorical representation of the data. This transformation involved assigning values of -1 to entries less than zero, +1 to entries greater than zero, and 0 to entries equal to zero, facilitating subsequent analyses.

Following the computation of Manhattan distance for the discretized dataframe, Principal Coordinate Analysis (PCoA) was conducted using the dudi.pco function in R. PCoA is a multivariate statistical technique utilized to visualize dissimilarities in data points, with resulting principal coordinates capturing underlying patterns and variability within the dataset.

In the PCoA plot, coloration was applied based on the sign of previously computed correlation coefficients (r values), distinguishing between positive and negative associations. Additionally, the size of each data point in the PCoA plot was determined by the number of drugs positively associated with the corresponding microbial species. A separate PCoA plot was generated using the same methodology, with dot size now indicating the number of drugs negatively associated with each microbial species. This analysis provides valuable insights into the relationship between microbial community composition, drug associations, and the underlying patterns captured by PCoA. Refer Figure 3.12 and 3.11

A heatmap was constructed to represent the discretized dataframe, where rows correspond to microbial species and columns to drug names. Values of -1, 0, and 1 indicate the negative, absence, or positive associations, respectively, between species and drugs. This heatmap provides a visual overview of the relationships between microbial species and drugs, aiding in the identification of potential associations.

Adjacent to the discretized dataframe heatmap, a horizontal barplot was generated to illustrate the correlation coefficients (r values) between microbial species and drugs. Each bar represents a microbial species, with bars extending to the left or right based on negative or positive correlation coefficients, respectively. This barplot offers insights into the strength and directionality of associations between microbial species and drugs, facilitating the identification of significant relationships.

Furthermore, an additional heatmap was constructed to visualize the distribution of positive and negative associations between microbial species and drugs. In this

heatmap, the intensity of color represents the number of drugs associated with each microbial species, with darker shades indicating a higher number of associations. This visualization enables the identification of microbial species that are heavily or lightly associated with drugs, providing valuable insights into potential drug-microbiome interactions.

Overall, these visualizations offer a comprehensive understanding of the relationships between microbial species and drugs, elucidating the patterns and dynamics of drug-microbiome associations within the dataset. Refer figure 3.13

3.3 Results of association of species with drugs influenced by age

The correlation plot provided insights into the associations between microbial species abundance and age, revealing both positive and negative correlations. Warm-colored cells indicated positive correlations, suggesting that certain microbial species were associated with older age groups, while cool-colored cells indicated negative correlations, indicating associations with younger age groups. This information highlighted the complex interplay between microbial composition and age.

Similarly, the directionality heatmap elucidated the direction of associations between microbial species and age. Shades of blue indicated positive associations, where higher microbial abundance corresponded to older age groups, while shades of red indicated negative associations, suggesting associations with younger age groups. This visualization allowed for the identification of age-related shifts in microbial composition and provided context for further analysis.

Finally, the scatter plot integrated information on both age-related microbial shifts and drug associations, offering a comprehensive view of their interplay. By color-coding species based on their associations with age and drugs, patterns of co-occurrence and potential synergies or antagonisms between drug effects and age-related microbial shifts were discerned. This visualization facilitated a deeper understanding of the complex relationships between age, drug exposure, and microbial composition within the dataset.

From above results it is clearly evident that age and drug associated species are not correlated. Refer 3.14, 3.15, 3.16, 3.17, 3.18, 3.19 and 3.20

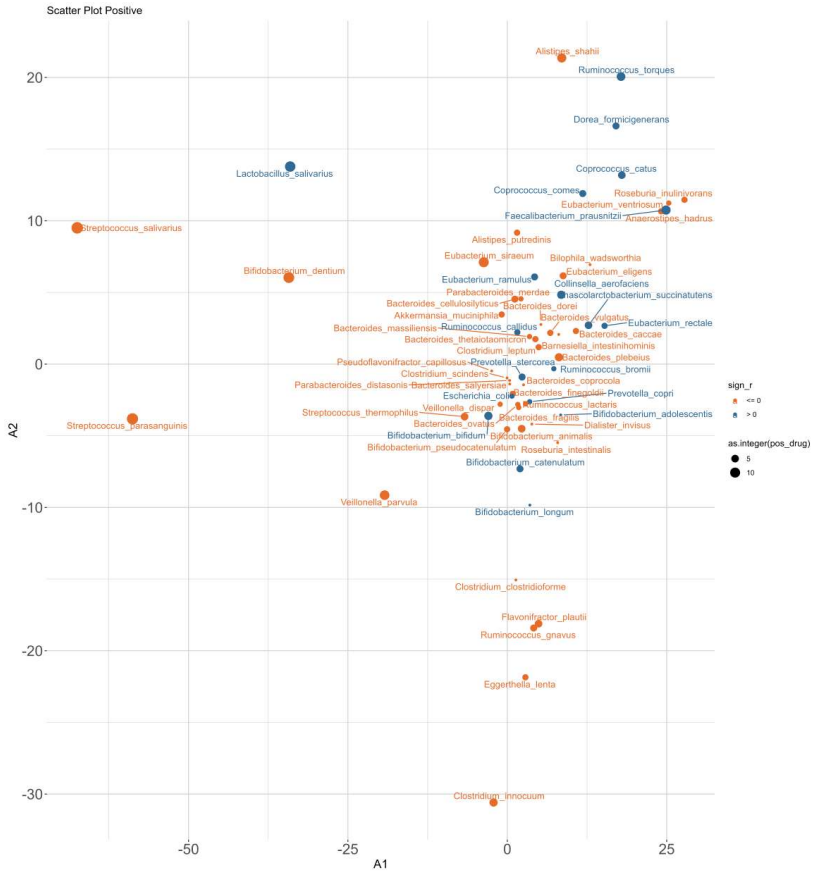

Figure 3.11: Above PCoA plot visualizes the dissimilarities in microbial community composition based on Manhattan distance of gut microbiome. The data points representing species are colored on the basis of the sign of r values (correlation coefficient), distinguishing between positive and negative cohort life style associations. Additionally, the size of each point corresponds to the number of drugs that species is positively associated with.

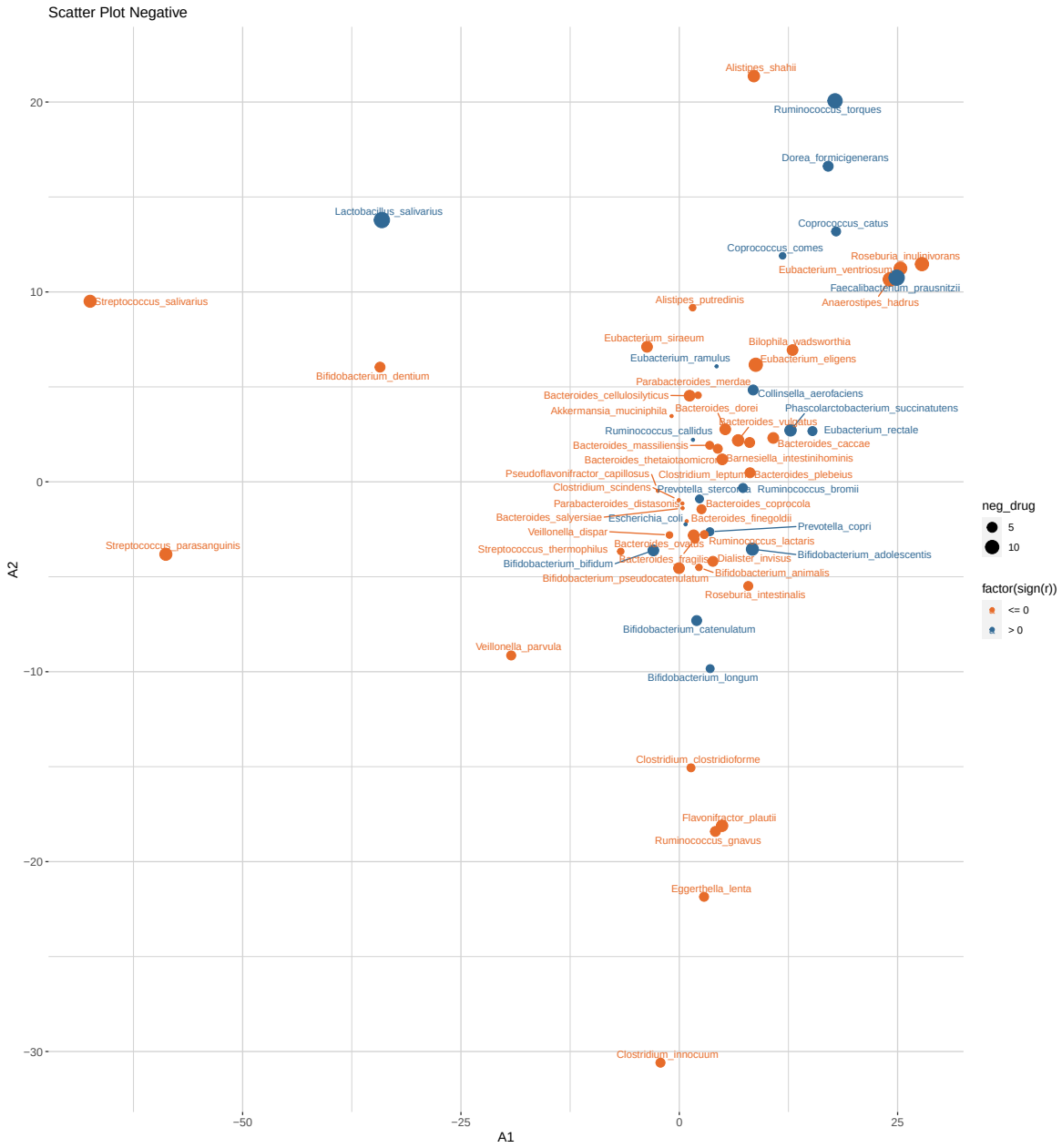

Figure 3.12: Above PCoA plot visualizes the dissimilarities in microbial community composition based on Manhattan distance of gut microbiome. The data points representing species are colored on the basis of the sign of r values (correlation coefficient), distinguishing between positive and negative cohort life style associations. Additionally, the size of each point corresponds to the number of drugs that species is negatively associated with

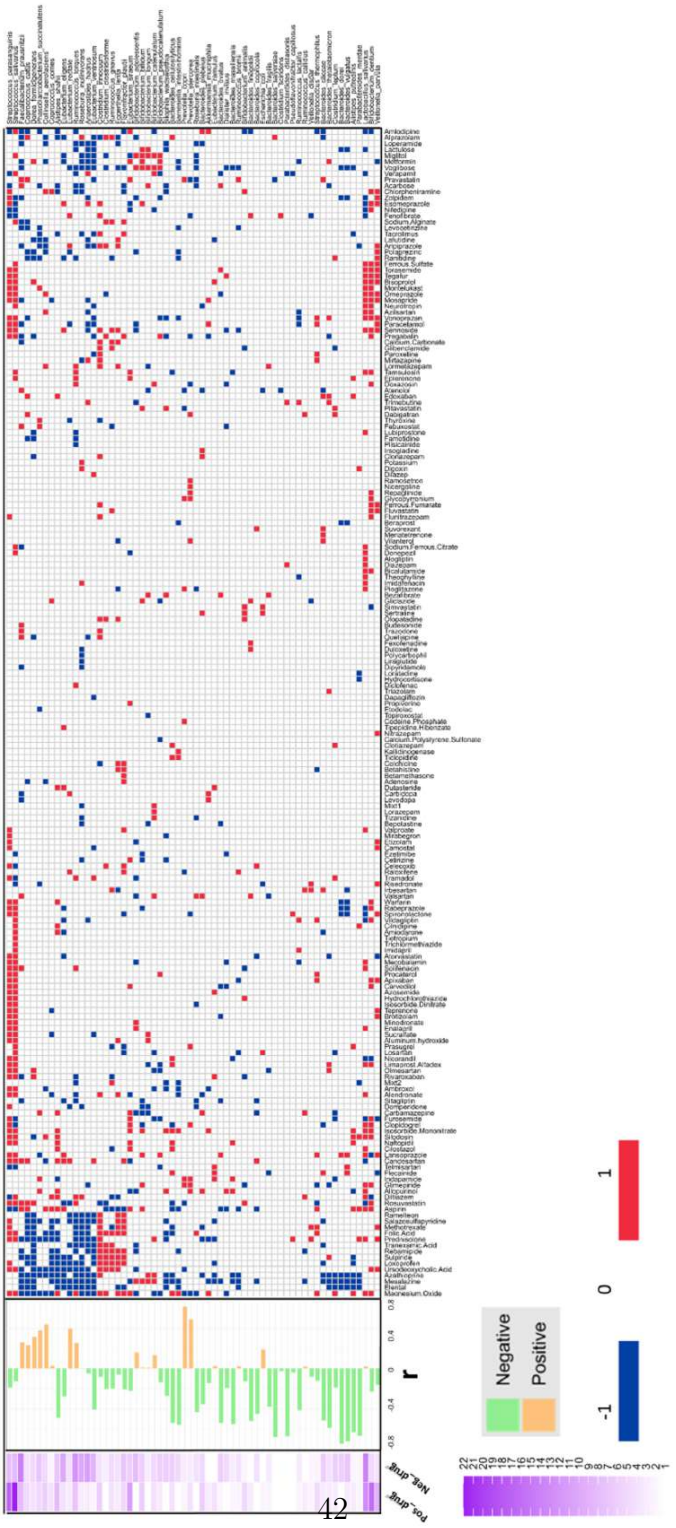

Figure 3.13: Heatmap

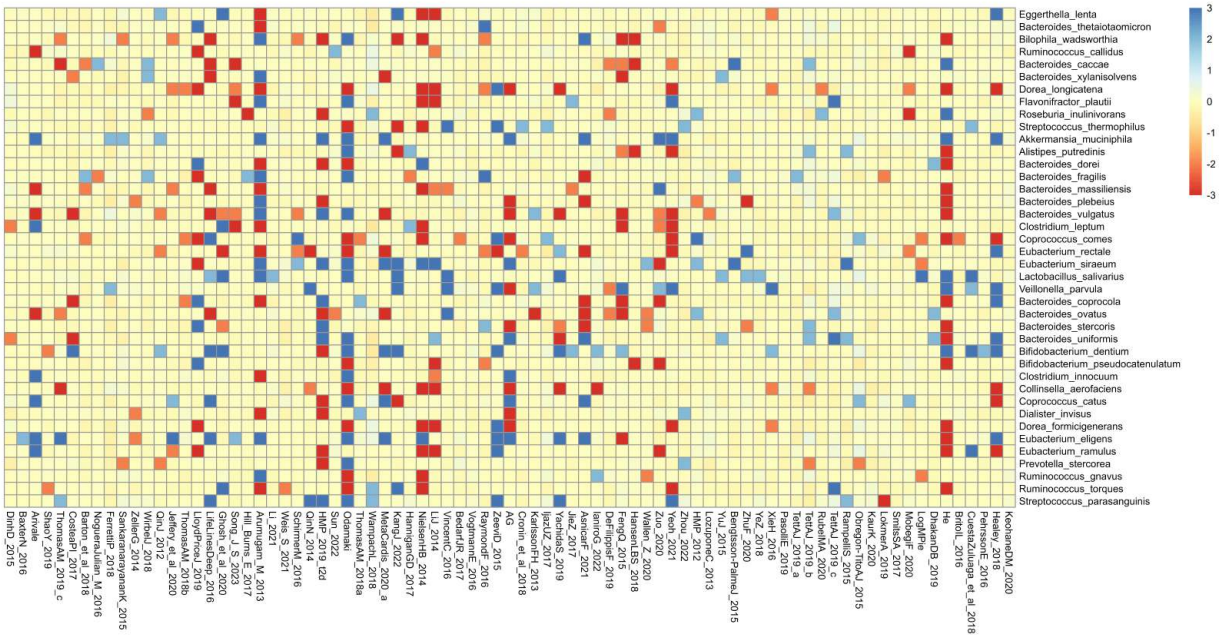

Figure 3.14: The directionality heatmap displays the direction of associations between microbial species and age, based on the values of correlation coefficients (r). Each row in the heatmap corresponds to a microbial species, while each column represents age. Positive associations are depicted with shades of blue, indicating instances where higher microbial abundance corresponds to older age groups. Conversely, negative associations are depicted with shades of red, suggesting associations where higher microbial abundance corresponds to younger age groups. Species involved in this plot have some association with the drugs.

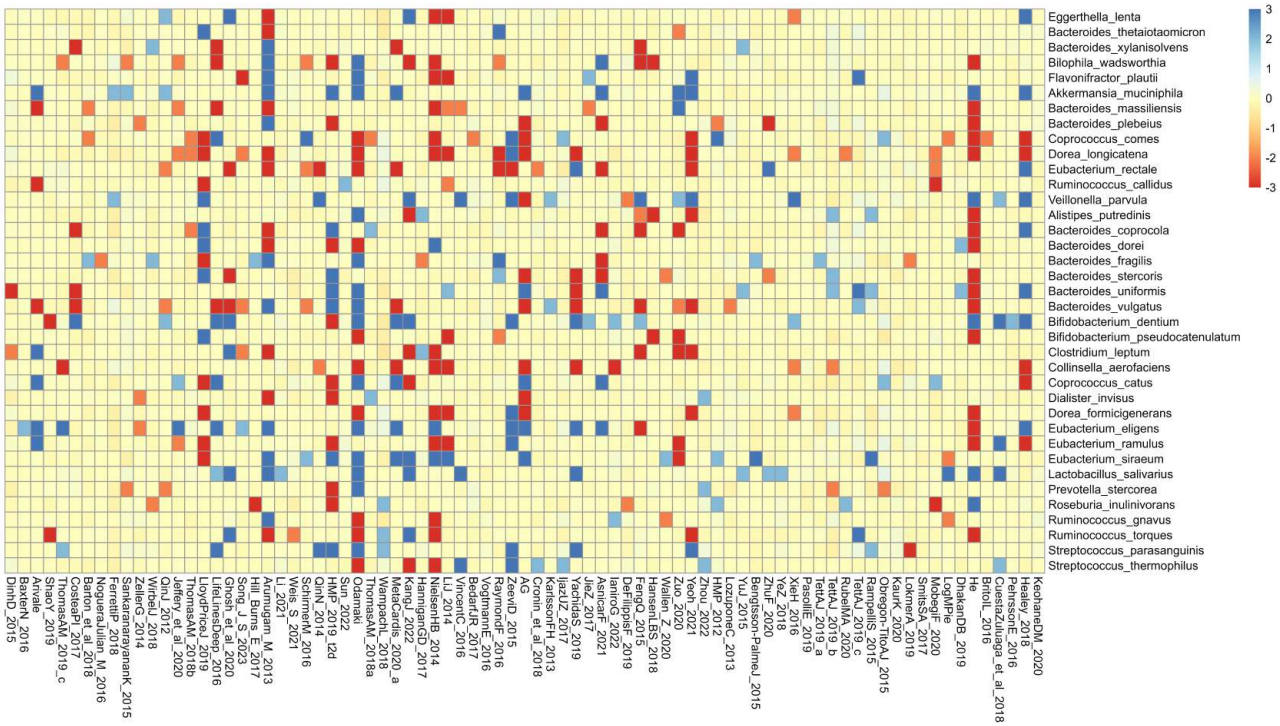

Figure 3.15: The directionality heatmap displays the direction of associations between microbial species and age, based on the values of correlation coefficients (r). Each row in the heatmap corresponds to a microbial species, while each column represents age. Positive associations are depicted with shades of blue, indicating instances where higher microbial abundance corresponds to older age groups. Conversely, negative associations are depicted with shades of red, suggesting associations where higher microbial abundance corresponds to younger age groups. Species involved in this plot have positive association with the drugs.

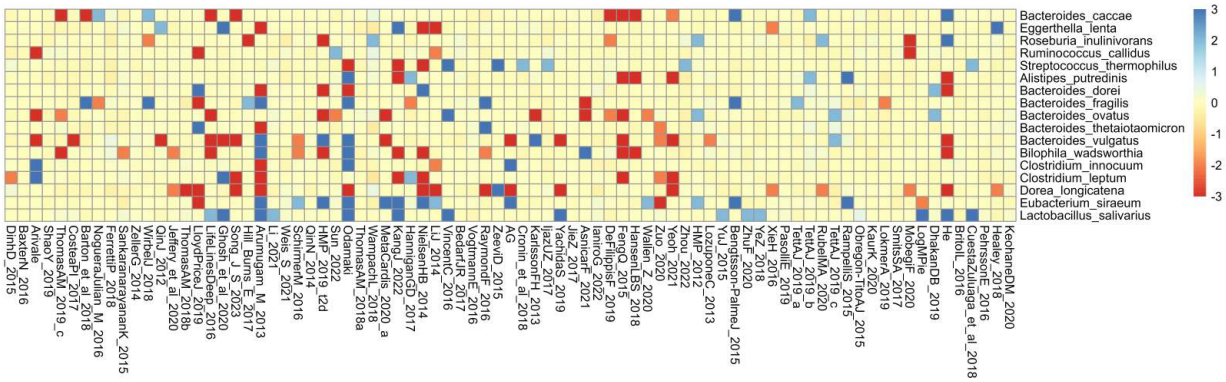

Figure 3.16: The directionality heatmap displays the direction of associations between microbial species and age, based on the values of correlation coefficients (r). Each row in the heatmap corresponds to a microbial species, while each column represents age. Positive associations are depicted with shades of blue, indicating instances where higher microbial abundance corresponds to older age groups. Conversely, negative associations are depicted with shades of red, suggesting associations where higher microbial abundance corresponds to younger age groups. Species involved in this plot have negative association with the drugs.

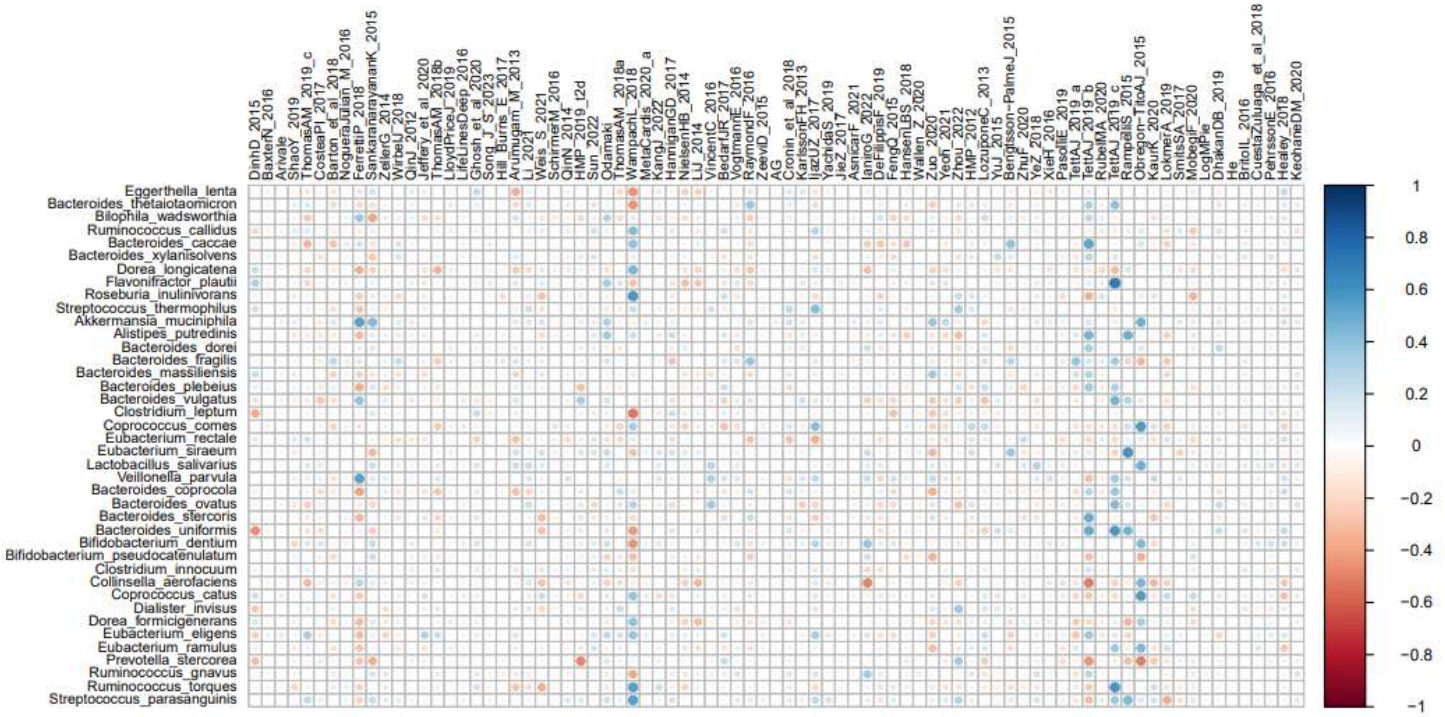

Figure 3.17: This correlation plot illustrates the associations between microbial species and age, depicting correlation coefficients (r values) as a heatmap. Each datapoint in the heatmap represents the r values (correlation coefficient) between a microbial species and age, with blue colour indicating positive correlations and negative correlation is indicated by red colour. This plot provides insights about the strength and directionality of associations between microbial species abundance and age. Species included in this plot have some association with drugs.

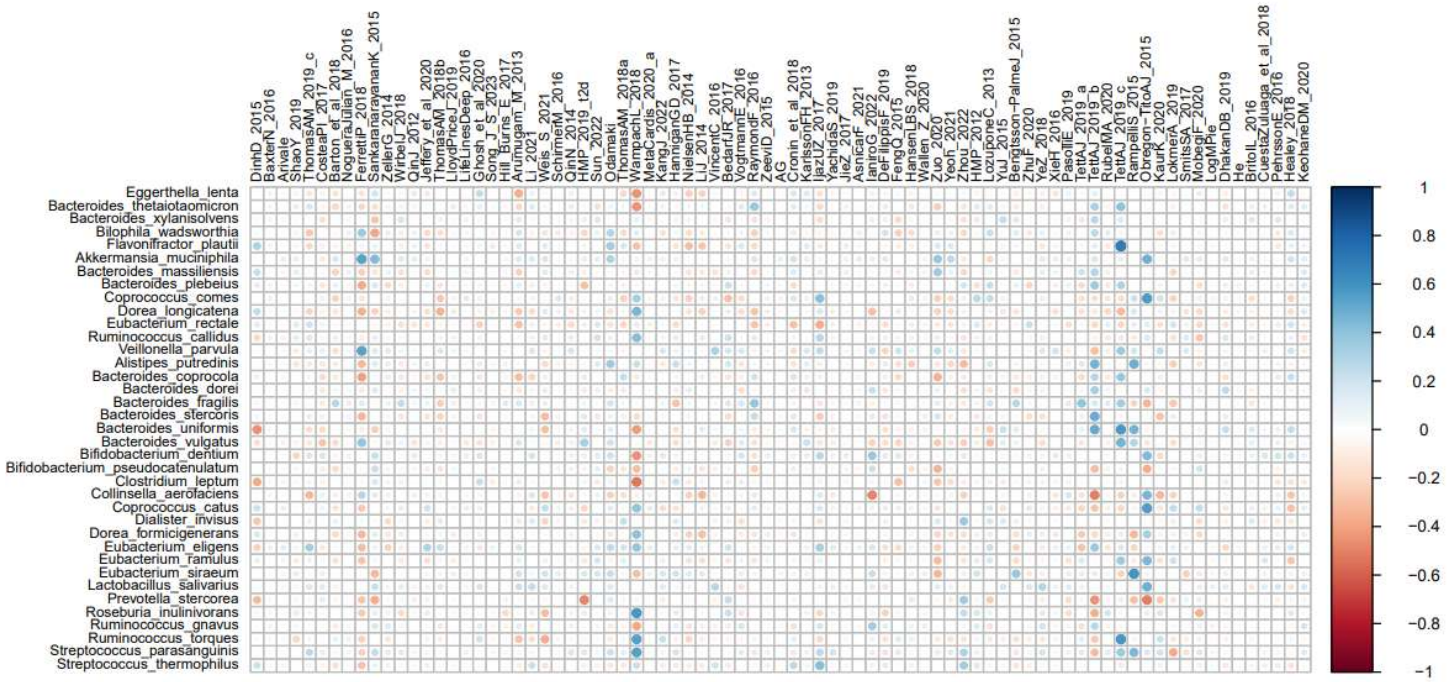

Figure 3.18: This correlation plot illustrates the associations between microbial species and age, depicting correlation coefficients (r values) as a heatmap. Each datapoint in the heatmap represents the r values (correlation coefficient) between a microbial species and age, with blue colour indicating positive correlations and negative correlation is indicated by red colour. This plot provides insights about the strength and directionality of associations between microbial species abundance and age. Species included in this plot have positive association with drugs..

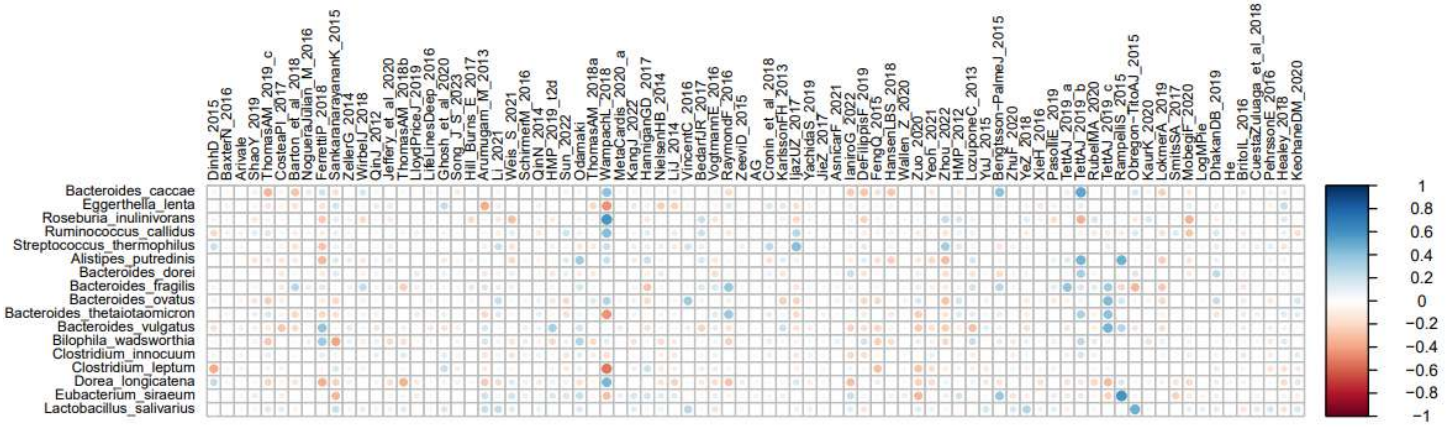

Figure 3.19: This correlation plot illustrates the associations between microbial species and age, depicting correlation coefficients (r values) as a heatmap. Each datapoint in the heatmap represents the r values (correlation coefficient) between a microbial species and age, with blue colour indicating positive correlations and negative correlation is indicated by red colour. This plot provides insights about the strength and directionality of associations between microbial species abundance and age. Species included in this plot have negative association with drugs.

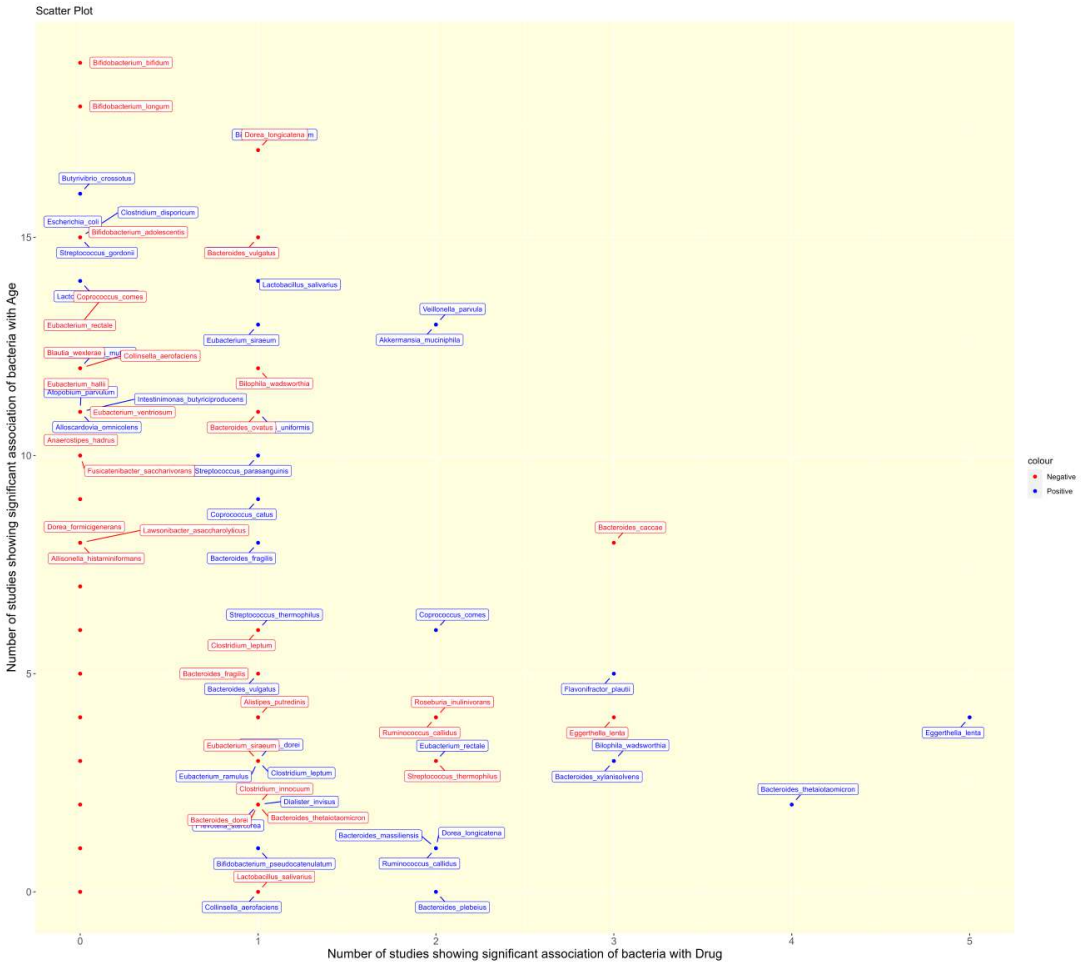

Figure 3.20: The scatter plot visualizes the dissimilarities in microbial community composition across different samples, considering associations with both age and drug exposure. Each point in the plot represents a sample, with the x-axis indicating the number of positively associated drugs and positive age associations, and the y-axis representing the number of negatively associated drugs and negative age associations. Species are color-coded based on their respective associations, allowing for the identification of patterns and potential interactions between drug effects and age-related microbial shifts.

Chapter 4

Conclusion and Future Scope

4.1 Conclusion

In conclusion, this thesis has significantly contributed to our understanding of the complex interplay between gut microbiota, drug exposure, age, and lifestyle. Through meticulous analysis of diverse datasets from numerous studies, we have identified profound associations between microbial species and drug exposure, elucidating specific gut microbiome clusters that respond distinctively to different drug structures. Additionally, by leveraging the concept of the Hack score, we have delineated drug modules, wherein drugs associated with low Hack scores exhibit detrimental effects on beneficial bacterial species, while those associated with high scores confer potential therapeutic benefits. These drug-responsive modules also show differential enrichment of markers of health and disease, indicating that while certain drugs have a minimal or beneficial impact on the gut microbiomes, others can potentially drive the gut microbiome to a detrimental state. These findings underscore the pivotal role of gut microbiota in modulating drug responses and highlight the importance of considering microbiome dynamics in drug development and personalized medicine initiatives.

Furthermore, our investigation has shed light on the association between microbial species and industrialized lifestyles. We have observed that species exhibiting significant associations with drugs are also commonly associated with industrialized lifestyles, suggesting potential links between environmental factors and drug-microbiome interactions. This insight underscores the need for comprehensive approaches that account for both genetic and environmental influences on the gut microbiome, particularly in the context of drug response variability and personalized healthcare strategies.

Moreover, while age is recognized as a crucial factor influencing microbial composition, our analysis revealed no significant correlation between age and species associated

with drugs. This finding suggests that age-related microbial shifts may not directly influence drug-microbiome interactions. However, further research is warranted to elucidate the intricate relationships between age, microbial composition, and drug response variability. Overall, the insights gained from this thesis have significant implications for precision medicine, highlighting the importance of personalized approaches that consider the complex interplay between host factors, lifestyle, and drug exposure in shaping gut microbiome dynamics and therapeutic outcomes.

4.2 Future Perspectives

The findings presented in this thesis open up exciting avenues for future research in the field of microbiome-drug interactions. While our analysis encompassed four diverse datasets, there exists vast potential for further exploration with even more expansive and varied datasets, encompassing diverse populations, clinical conditions, and drug exposures. Expanding the scope of analysis to include additional datasets will enhance the robustness and generalizability of our findings, allowing for the identification of novel drug-microbiome associations and the validation of existing correlations across different populations and contexts. Additionally, integrating multi-omics data, including metagenomics, metabolomics, and host genetics, holds promise for elucidating the mechanistic underpinnings of drug-microbiome interactions and their impact on human health. Furthermore, longitudinal studies tracking microbiome dynamics before and after drug exposure, coupled with comprehensive clinical metadata, will provide invaluable insights into the temporal dynamics of microbiome responses to drug interventions and their implications for therapeutic efficacy and adverse effects. Leveraging advanced computational and statistical methods, such as machine learning and network analysis, will enable the development of predictive models for personalized medicine, facilitating the identification of optimal drug therapies tailored to individual microbiome profiles and clinical characteristics. Overall, the future of microbiome research in the context of drug interactions is ripe with opportunities for innovation and discovery, promising transformative insights into human health and disease management.

Bibliography

[1] Q. Zhao, Y. Chen, W. Huang, H. Zhou, and W. Zhang, "Drug-microbiota interactions: an emerging priority for precision medicine," *Signal Transduction and Targeted Therapy*, vol. 8, no. 1, p. 386, Oct. 2023. [Online]. Available: https://www.nature.com/articles/s41392-023-01619-w

[2] R. Saad, M. R. Rizkallah, and R. K. Aziz, "Gut Pharmacomicrobiomics: the tip of an iceberg of complex interactions between drugs and gut-associated microbes," *Gut Pathogens*, vol. 4, no. 1, p. 16, Dec. 2012. [Online]. Available: https://gutpathogens.biomedcentral.com/articles/10.1186/1757-4749-4-16

[3] J. A. Gilbert, R. A. Quinn, J. Debelius, Z. Z. Xu, J. Morton, N. Garg, J. K. Jansson, P. C. Dorrestein, and R. Knight, "Microbiome-wide association studies link dynamic microbial consortia to disease," *Nature*, vol. 535, no. 7610, pp. 94–103, Jul. 2016. [Online]. Available: https://www.nature.com/articles/nature18850

[4] S. K. Forslund, R. Chakaroun, M. Zimmermann-Kogadeeva, L. Markó, J. Aron-Wisnewsky, T. Nielsen, L. Moitinho-Silva, T. S. B. Schmidt, G. Falony, S. Vieira-Silva, S. Adriouch, R. J. Alves, K. Assmann, J.-P. Bastard, T. Birkner, R. Caesar, J. Chilloux, L. P. Coelho, L. Fezeu, N. Galleron, G. Helft, R. Isnard, B. Ji, M. Kuhn, E. Le Chatelier, A. Myridakis, L. Olsson, N. Pons, E. Prifti, B. Quinquis, H. Roume, J.-E. Salem, N. Sokolovska, V. Tremaroli, M. Valles-Colomer, C. Lewinter, N. B. Søndertoft, H. K. Pedersen, T. H. Hansen, J. P. Gøtze, L. Køber, H. Vestergaard, T. Hansen, J.-D. Zucker, S. Hercberg, J.-M. Oppert, I. Letunic, J. Nielsen, F. Bäckhed, S. D. Ehrlich, M.-E. Dumas, J. Raes, O. Pedersen, K. Clément, M. Stumvoll, and P. Bork, "Combinatorial, additive and dose-dependent drug–microbiome associations," *Nature*, vol. 600, no. 7889, pp. 500–505, Dec. 2021, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41586-021-04177-9

[5] C. L. Nguyen, K. A. Markey, O. Miltiadous, A. Dai, N. Waters, K. Sadeghi, T. Fei, R. Shouval, B. P. Taylor, C. Liao, J. B. Slingerland, A. E. Slingerland,

A. G. Clurman, M. A. Maloy, L. Bohannon, P. A. Giardina, D. G. Brereton, G. K. Armijo, E. Fontana, A. Gradissimo, B. Gyurkocza, A. D. Sung, N. J. Chao, S. M. Devlin, Y. Taur, S. A. Giralt, M.-A. Perales, J. B. Xavier, E. G. Pamer, J. U. Peled, A. L. Gomes, and M. R. Van Den Brink, "High-resolution analyses of associations between medications, microbiome, and mortality in cancer patients," *Cell*, vol. 186, no. 12, pp. 2705–2718.e17, Jun. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0092867423005263

[6] S. M. Bahr, B. C. Tyler, N. Wooldridge, B. D. Butcher, T. L. Burns, L. M. Teesch, C. L. Oltman, M. A. Azcarate-Peril, J. R. Kirby, and C. A. Calarge, "Use of the second-generation antipsychotic, risperidone, and secondary weight gain are associated with an altered gut microbiota in children," *Translational Psychiatry*, vol. 5, no. 10, pp. e652–e652, Oct. 2015. [Online]. Available: https://www.nature.com/articles/tp2015135

[7] S. Fromentin, S. K. Forslund, K. Chechi, J. Aron-Wisnewsky, R. Chakaroun, T. Nielsen, V. Tremaroli, B. Ji, E. Prifti, A. Myridakis, J. Chilloux, P. Andrikopoulos, Y. Fan, M. T. Olanipekun, R. Alves, S. Adiouch, N. Bar, Y. Talmor-Barkan, E. Belda, R. Caesar, L. P. Coelho, G. Falony, S. Fellahi, P. Galan, N. Galleron, G. Helft, L. Hoyles, R. Isnard, E. Le Chatelier, H. Julienne, L. Olsson, H. K. Pedersen, N. Pons, B. Quinquis, C. Rouault, H. Roume, J.-E. Salem, T. S. B. Schmidt, S. Vieira-Silva, P. Li, M. Zimmermann-Kogadeeva, C. Lewinter, N. B. Søndertoft, T. H. Hansen, D. Gauguier, J. P. Gøtze, L. Køber, R. Kornowski, H. Vestergaard, T. Hansen, J.-D. Zucker, S. Hercberg, I. Letunic, F. Bäckhed, J.-M. Oppert, J. Nielsen, J. Raes, P. Bork, M. Stumvoll, E. Segal, K. Clément, M.-E. Dumas, S. D. Ehrlich, and O. Pedersen, "Microbiome and metabolome features of the cardiometabolic disease spectrum," *Nature Medicine*, vol. 28, no. 2, pp. 303–314, Feb. 2022. [Online]. Available: https://www.nature.com/articles/s41591-022-01688-4

[8] T. S. Ghosh, S. Rampelli, I. B. Jeffery, A. Santoro, M. Neto, M. Capri, E. Giampieri, A. Jennings, M. Candela, S. Turroni, E. G. Zoetendal, G. D. A. Hermes, C. Elodie, N. Meunier, C. M. Brugere, E. Pujos-Guillot, A. M. Berendsen, L. C. P. G. M. De Groot, E. J. M. Feskins, J. Kaluza, B. Pietruszka, M. J. Bielak, B. Comte, M. Maijo-Ferre, C. Nicoletti, W. M. De Vos, S. Fairweather-Tait, A. Cassidy, P. Brigidi, C. Franceschi, and P. W. O'Toole, "Mediterranean diet intervention alters the gut microbiome in older people reducing frailty and improving health status: the NU-AGE 1-year dietary intervention across five

European countries," *Gut*, vol. 69, no. 7, pp. 1218–1228, Jul. 2020. [Online]. Available: https://gut.bmj.com/lookup/doi/10.1136/gutjnl-2019-319654

[9] T. S. Ghosh, F. Shanahan, and P. W. O'Toole, "Toward an improved definition of a healthy microbiome for healthy aging," *Nature Aging*, vol. 2, no. 11, pp. 1054–1069, Nov. 2022. [Online]. Available: https://www.nature.com/articles/s43587-022-00306-9

[10] N. Nagata, S. Nishijima, T. Miyoshi-Akiyama, Y. Kojima, M. Kimura, R. Aoki, M. Ohsugi, K. Ueki, K. Miki, E. Iwata, K. Hayakawa, N. Ohmagari, S. Oka, M. Mizokami, T. Itoi, T. Kawai, N. Uemura, and M. Hattori, "Population-level Metagenomics Uncovers Distinct Effects of Multiple Medications on the Human Gut Microbiome," *Gastroenterology*, vol. 163, no. 4, pp. 1038–1052, Oct. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0016508522007326

[11] H. Cooper, L. V. Hedges, and J. C. Valentine, Eds., *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation, Jun. 2019. [Online]. Available: http://www.jstor.org/stable/10.7758/9781610448864