Microbial Strain Variations Co-vary with Lifestyle, Health and Community Ecology

by

Ana Sharma

Submitted

In partial fulfillment of the requirements for the degree of

Master of Technology

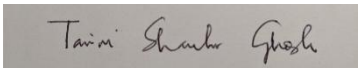
to

Indraprastha Institute of Information Technology, Delhi
August 2025

Certificate

This is to certify that the thesis titled **Microbial Strain Variations Co-vary with Lifestyle, Health and Community Ecology** being submitted by **Ana Sharma** to the Indraprastha Institute of Information Technology, Delhi, for the award of the Master of Technology, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or in full to any other university or institute for the award of any degree/diploma.



August 2025

Dr. Tarini Shankar Ghosh
Department of Computational Biology,
Indraprastha Institute of Information
Technology, New Delhi 110020

Acknowledgements

I would like to express my deepest gratitude to all those who have supported me and guided me throughout my M.Tech. thesis work. First and foremost, I would like to thank my esteemed project supervisor, **Dr. Tarini Shankar Ghosh**, for providing me with this opportunity to work under his guidance. His wisdom and guidance have profoundly shaped my understanding and approach to this project.

I am extremely grateful to **Shivangi Verma** for her involvement, insightful comments, and motivation. Their insights and efforts have been invaluable, and working with them has been a rewarding experience. Lastly, I would like to thank IIIT-Delhi for providing the necessary infrastructure.

Ana Sharma
(MT23226)
M.Tech CB

Abstract

Background
People living in industrialized and non-industrialized settings harbor distinct gut microbiomes shaped by differences in diet, lifestyle, and environment. Understanding these variations and their correlations with demographic and health factors is essential for interpreting microbiome function across populations. Metagenomic sequencing provides valuable insights into microbial communities, but traditional workflows often lack consistency and strain-level resolution, limiting their ability to reveal fine-scale differences. High-resolution analysis is therefore necessary to capture subtle lifestyle-related influences on microbial diversity.

Methodology
We analyzed 149 metagenomic studies comprising over 40,000 samples, applying established computational tools for strain-level resolution and species-level classification. DADA2 was used for Amplicon Sequence Variant (ASV) inference, while SPINGO enabled species-level taxonomic assignment. To assess microbial diversity, we employed Shannon indices (alpha diversity), Euclidean distances with PCoA (beta diversity), and CLR normalization for compositional correction. Procrustes and Mantel tests were further used to examine associations between microbial community structure, lifestyle groups, and health-related variables.

Findings
Across the datasets, distinct patterns in microbial diversity and taxonomic composition were observed between industrialized and non-industrialized cohorts. Certain species showed lifestyle-linked differences in abundance and diversity, while interspecies associations appeared consistent across populations. These patterns suggest lifestyle exerts a measurable influence on microbiome structure without erasing underlying microbial relationships.

Conclusions
This study highlights how industrialization and lifestyle factors shape microbial diversity and community composition. By systematically applying advanced analytical methods across diverse datasets, we provide insights into correlations between lifestyle, microbial variation, and health-related factors. These findings contribute to a broader understanding of the interplay between environment, lifestyle, and the gut microbiome, offering directions for future microbiome research.

Table of Contents

Chapter 1: Introduction

1.1 The Central Role of Metagenomics and Microbial Communities

Life on Earth is inextricably linked to the diverse and dynamic world of microorganisms. These microscopic entities, encompassing bacteria, archaea, fungi, and viruses, form complex communities that inhabit virtually every niche, from extreme environments to the human body. The collective genetic material of these communities within a given environment is known as the metagenome. The field of metagenomics has revolutionized our understanding of these microbial communities by enabling the study of their genetic potential and functional roles without the need for traditional laboratory cultivation. This culture-independent approach has unveiled an astonishing breadth of microbial diversity and metabolic capabilities previously hidden.

In the context of human health, microbial communities, particularly those residing in the gut, are now recognized as fundamental determinants of well-being, influencing digestion, metabolism, immune system development, and susceptibility to various diseases. Similarly, environmental microbial communities drive global biogeochemical cycles, contribute to nutrient availability, and play crucial roles in processes like bioremediation and agriculture. The specific composition, structure, and functional potential of any given microbial community are inextricably linked to the intricate interplay between its constituent members and their environment. Understanding this intricate relationship, especially at high resolution, including strain-level insights, is essential for deciphering their designated roles. A disruption or loss of this balanced structure, known as dysbiosis, can result in a loss of function or contribute to disease states. Consequently, the ability to accurately characterize these communities, their composition, diversity, and the interactions within them is of paramount importance. This characterization often involves quantifying diversity (e.g., Shannon Index) and understanding patterns of community divergence, which are critical aspects of this. A deeper understanding indicates greater insight into microbial ecology, which is a highly desirable trait for applications in clinical diagnostics (e.g., identifying disease-associated microbiomes) and industrial biotechnology (e.g., optimizing microbial consortia for bioproduction). Therefore, understanding and accurately characterizing microbial communities through metagenomic analysis is not merely an academic exercise; it is a central challenge in microbiology, biotechnology, and medicine, with the potential to unlock the full therapeutic and industrial potential of engineered microbial systems.

1.2 Challenges in Metagenomic Data Analysis

The comprehensive characterization of microbial communities from metagenomic data is a complex endeavor, governed by a delicate balance of biological and technical factors. It is not dictated by a single challenge but is an emergent property arising from the interplay between the raw sequencing data and the sophisticated computational methods required for its interpretation. The raw data provides the fundamental information, while the analytical pipeline represents the architecture where meaningful insights are realized.

Analyzing metagenomic data presents several inherent challenges that must be meticulously addressed to ensure accurate and reliable conclusions:

- **High Dimensionality and Sparsity:** Metagenomic datasets typically involve a vast number of microbial taxa (often thousands of ASVs or species) across a large number of samples. Many taxa are present in only a subset of samples, leading to highly sparse abundance matrices, which can complicate statistical analysis.
- **Compositional Nature of Data:** Metagenomic sequencing data provides relative abundances of microbial taxa within a sample, not absolute counts. This compositional nature means that an increase in one taxon's abundance necessarily leads to a proportional decrease in others, even if their absolute numbers remain constant. Ignoring this property can lead to spurious correlations and misleading interpretations in downstream statistical analyses.
- **Technical Biases and Errors:** Data generation involves multiple steps, from DNA extraction and PCR amplification to sequencing. Each step can introduce biases and errors (e.g., chimeras, sequencing errors, amplification biases) that can obscure true biological signals and affect the accuracy of taxonomic and diversity assessments.
- **Computational Intensity and Scalability:** Processing and analyzing large metagenomic datasets require significant computational resources and efficient algorithms. The sheer volume of data from hundreds or thousands of samples necessitates robust and scalable pipelines to handle the computational load effectively.
- **Achieving High-Resolution Insights:** While traditional methods often group sequences into Operational Taxonomic Units (OTUs), this can obscure fine-scale biological variation. Achieving high-resolution insights, such as distinguishing Amplicon Sequence Variants (ASVs) or identifying species-level and even strain-level differences, is crucial for understanding subtle functional variations and host-microbe interactions, but it adds to analytical complexity.
- **Integration of Diverse Data Types:** Combining microbial abundance data with rich metadata (e.g., host lifestyle, age, BMI, disease status) from various studies requires sophisticated data integration and mapping strategies to derive meaningful associations.

1.3 Motivation for the Project

The gut microbiome, a complex community of microorganisms residing in the human gastrointestinal tract, profoundly influences host health and disease. Understanding the intricate relationships between these microbial communities and various host factors—including lifestyle, diet, and genetics—is crucial for advancing personalized medicine, nutrition, and public health. This field has been revolutionized by high-throughput sequencing technologies, which generate vast amounts of metagenomic data, providing an unprecedented view into the microbial world.

Challenge and Our Approach: Despite the wealth of available data, a major challenge in microbiome research lies in accurately and comprehensively analyzing these complex datasets to reveal meaningful biological insights. Many existing studies focus on broad taxonomic differences at the genus or species level, which may overlook the nuanced variations at the strain level that can significantly impact host interactions. Microbial strain variations, for instance, can dictate an organism's metabolic capabilities, virulence, and ability to colonize specific environments. This thesis is motivated by the need to move beyond broad-brush analyses to investigate how microbial strain variations co-vary with lifestyle, health, and community ecology.

1.4 Objectives of the Project

The project's goal was to leverage established, high-resolution analytical methods to examine existing metagenomic datasets from diverse human populations. By applying these advanced bioinformatics tools, we sought to systematically compare the gut microbiome composition of different lifestyle groups. The central hypothesis of this thesis is that distinct lifestyle groups exhibit significant and measurable differences in their gut microbiome, particularly at the strain level. Through this comparative analysis, we aimed to uncover specific microbial strains and community patterns associated with different lifestyles, thereby contributing to a more detailed understanding of the complex interplay between human behaviour and microbial ecology.

1. **To understand how the Microbial Strain Variations Vary with Lifestyle, Health, and Community Ecology:** We will design and implement an end-to-end pipeline that encompasses all critical stages of metagenomic data analysis, from raw data acquisition to advanced statistical interpretation. This involves:
    - Establishing robust protocols for raw data quality control and preprocessing.
    - Integrating and optimizing state-of-the-art tools for Amplicon Sequence Variant (ASV) inference and species-level taxonomic classification.
2. **To Perform High-Resolution ASV Inference and Taxonomic Assignment:** We will apply and evaluate the performance of key bioinformatics tools for precise microbial characterization:
    - Conducting ASV inference using the DADA2 algorithm to model and correct Illumina-sequenced amplicon errors, thereby achieving high-resolution sequence variants.
    - Employing SPINGO, a probabilistic classifier, for accurate species-level taxonomic assignment of ASVs using its curated reference database.
3. **To Conduct Comprehensive Downstream Analytical Tests:** We will apply a suite of advanced statistical and ecological analyses to derive deeper biological insights from the processed data:
    - Calculating intra-species diversity using the Shannon Index to understand strain-level variation within individual species.
    - Implementing Centered Log-Ratio (CLR) normalization for compositional data analysis, followed by Euclidean distance calculation for beta diversity assessment.
    - Utilizing Principal Coordinates Analysis (PCoA) for dimensionality reduction and visualization of microbial community dissimilarities.
    - Performing Procrustes Mantel Tests to assess the statistical congruence and correlated divergence patterns between different species' strain compositions.
4. **To Investigate Associations with Host and Environmental Factors:** We will systematically explore the relationships between microbial community patterns and various metadata:
    - Analysing and visualizing microbial associations with distinct lifestyle groups, including Industrialized Urban and Non-Industrialized populations.
5. **To Advance Reproducible Metagenomic Research:** By developing and validating a comprehensive and robust pipeline, this work aims to contribute a valuable, reproducible framework to the field, empowering researchers to perform high-throughput, accurate, and interpretable metagenomic studies, thereby accelerating the understanding of microbial communities in health and disease.

Chapter 2: Literature Review

The comprehensive analysis of metagenomic data is a crucial and rapidly evolving field in microbiology, essential for deciphering the complex roles of microbial communities in various ecosystems and hosts. The field has seen a rapid evolution of methods, moving from traditional clustering-based approaches to sophisticated bioinformatics pipelines that leverage advanced algorithms for high-resolution insights and robust statistical analyses. Numerous studies have explored various techniques to enhance analytical accuracy and provide insight into the structure, diversity, and functional potential of microbial communities.

The field of metagenomics has seen a rapid evolution, moving from basic community profiling to high-resolution and reproducible analyses. Early methods, while foundational, had inherent limitations that have been overcome by sophisticated bioinformatics tools and statistical techniques. These advancements allow for the investigation of intricate relationships within microbial communities and their hosts with unprecedented detail.

A foundational approach in early metagenomic studies involved applying methods to cluster 16S rRNA gene sequences into Operational Taxonomic Units (OTUs). A common strategy in these studies was to group sequences based on a fixed similarity threshold (e.g., 97% sequence identity) using tools like QIIME or Mothur. While these methods provided valuable early models for community profiling and achieved moderate success, their performance was inherently limited by the arbitrary nature of the clustering threshold, which could obscure true biological variation and lead to a loss of resolution, particularly at the strain level. The reliance on predefined similarity often failed to fully capture the subtle, biologically meaningful sequence differences within a microbial population.

To overcome the limitations of fixed-threshold clustering and improve resolution, researchers turned to more sophisticated algorithms for Amplicon Sequence Variant (ASV) inference. This represented a significant step forward, as these methods could automatically identify and correct sequencing errors, yielding exact biological sequences. The most profound breakthrough came from the development and widespread adoption of tools like DADA2 (Callahan et al., 2016). DADA2 is a high-resolution pipeline that models and corrects Illumina-sequenced amplicon errors without the need for clustering, thereby preserving fine-scale biological variation and providing greater accuracy for studying intra-species diversity and community structure. This approach allows for the identification of true biological variants, pushing the boundaries of sequence-based microbial characterization.

In parallel with ASV inference, advancements in taxonomic classification have been crucial. Early classifiers often relied on simple similarity searches against reference databases, which could struggle with closely related species or novel taxa. The development of probabilistic and k-mer-based classifiers represented a significant improvement. SPINGO (Allard et al., 2015) is a notable example, specifically optimized for species-level identification of metagenomic amplicons using a k-mer-based approach and a highly curated reference database. This allows SPINGO to distinguish closely related microbial species with high precision, complementing the high-resolution ASVs generated by DADA2 and enabling more accurate taxonomic assignment.

Beyond initial processing and classification, the field has also seen significant advancements in downstream analytical methods to extract deeper biological insights from compositional metagenomic data. The compositional nature of abundance tables necessitates specialized

statistical approaches. Centered Log-Ratio (CLR) normalization has emerged as a robust method to transform relative abundances, making them suitable for standard multivariate statistical analyses. Following normalization, Euclidean distance calculation is widely used to quantify beta diversity, measuring the dissimilarity in microbial community composition between samples. Principal Coordinates Analysis (PCoA) is then frequently employed for dimensionality reduction and visualization of these beta diversity patterns, allowing researchers to observe clustering and separation of samples based on their microbial profiles. Furthermore, to investigate the congruence and correlated patterns between different microbial datasets or between microbial communities and environmental factors, Procrustes and Mantel Tests have become indispensable statistical tools. These methods allow for a complementary analysis, investigating whether explicit community structure or environmental information can offer predictive power that is orthogonal to that contained within basic abundance profiles alone.

The integration of these advanced tools and methodologies into comprehensive, end-to-end pipelines is central to current metagenomic research. Such frameworks, exemplified by the Microbial Strain Variations Co-vary with Lifestyle, Health and Community Ecology, enable rigorous, reproducible, and in-depth characterization of microbial communities across diverse datasets. This approach allows for systematic investigation of complex relationships, such as those between microbial diversity/composition and host lifestyle, age, BMI, country, or disease associations, thereby accelerating the understanding of the microbiome's profound impact.

Beyond initial processing, the field has seen significant advancements in downstream statistical methods. Because metagenomic abundance tables are compositional (i.e., relative abundances sum to a fixed value), specialized statistical approaches are necessary. Centered Log-Ratio (CLR) normalization has become a robust method for transforming relative abundances, making them suitable for standard multivariate statistical analyses.

To explore how microbial communities differ between samples (beta diversity), researchers often use Euclidean distance and visualize these patterns with Principal Coordinates Analysis (PCoA). This allows for the visualization of sample clustering based on their microbial profiles. To further investigate the relationships between microbial communities and other factors (e.g., lifestyle or BMI), Procrustes and Mantel Tests are invaluable. These statistical tools help determine if there's a significant correlation between the microbial community structure and the external data, offering deeper insights into the factors that shape the microbiome.

The integration of these advanced tools and methodologies into comprehensive frameworks is central to modern metagenomic research. This approach allows for the rigorous and reproducible investigation of complex relationships, such as those between microbial composition and host lifestyle, BMI, and health outcomes, thereby accelerating our understanding of the microbiome's profound impact.

Chapter 3: Materials and Methodology

3.1 Overview of the Microbial Strain Variations Co-vary with Lifestyle, Health and Community Ecology Framework

This study utilizes a **comprehensive and integrated computational framework** to investigate microbial communities. The framework processes raw sequencing data to generate **high-resolution, strain-level insights** and identifies their associations with various host and environmental factors.

Step	Title	Details
1	Study and Metadata Acquisition	Retrieved from ENA (European Nucleotide Archive)
2	Species-level Classification	Amplicon processing with DADA2 and classification using SPINGO
3	Diversity Calculation	Computed strain richness and Shannon index for strain evenness
4	CLR Normalization and Distance Calculation	Centered Log-Ratio normalization and Euclidean distance calculation
5	PCoA and Visualization	Principal Coordinate Analysis (PCoA) conducted for visualizing sample divergence
6	Cross-species Structural Comparison	Used Procrustes and Mantel tests across species

(Table 3.1: Table depicting the complete pipeline of the analysis project)

1. **Study and Metadata Acquisition**: Data was collected from the European Nucleotide Archive (ENA), including raw sequencing files and associated metadata for each study. This metadata includes sample identifiers, subject demographics, and experimental conditions necessary for downstream analysis.
2. **Species-level Classification**: High-throughput sequencing reads were processed using DADA2 to generate amplicon sequence variants (ASVs), ensuring high-resolution identification of microbial strains. ASVs were then taxonomically classified to the species level using SPINGO, a tool specifically designed for accurate species-level assignments in 16S rRNA datasets.
3. **Diversity Calculation**: Strain-level diversity was assessed for each species across samples. This included calculating strain richness (the number of unique ASVs per species per sample) and the Shannon diversity index to capture strain evenness and community complexity. These metrics help quantify intra-species strain variation within the gut microbiome.
4. **CLR Normalization and Distance Calculation**: To address compositional bias in microbiome data, species-specific abundance tables were transformed using Centered Log-Ratio (CLR) normalization. Subsequently, Euclidean distances were

calculated between samples for each species, enabling quantitative assessment of inter-sample strain divergence.

5. **PCoA and Visualization**: Principal Coordinate Analysis (PCoA) was applied to the Euclidean distance matrices of each species to reduce dimensionality and visualize sample clustering. These plots help interpret the degree of strain-level separation across different biological or clinical conditions.
6. **Cross-species Structural Comparison:** To explore shared patterns of strain divergence across species, **Procrustes analysis** was used to align PCoA configurations of different species. Additionally, **Mantel tests** were performed to measure the correlation between species-level distance matrices. These analyses revealed species that exhibit similar structural divergence across host samples.

3.2 Dataset Information and Curation

The foundation of any robust metagenomic analysis is a high-quality, diverse, and well-curated dataset. For this thesis, we rigorously curated and utilized a collection of 149 diverse metagenomic studies, primarily focusing on gut microbiome data. These studies were sourced from publicly available repositories, ensuring a broad representation of human populations and conditions. The raw sequencing data, typically in FASTQ format, along with their associated metadata, formed the basis of our analysis.

3.2.1 Rigorous Curation of 149 Diverse Metagenomic Studies

The curation process involved several critical steps to ensure data quality and comparability across studies:

- **Study Selection Criteria:** Studies were selected based on criteria such as sequencing technology (primarily 16S rRNA gene sequencing), sample type (focus on human gut microbiome), and availability of comprehensive metadata.
- **Data Acquisition:** Raw sequencing data (FASTQ files) were downloaded from public repositories (e.g., NCBI Sequence Read Archive - SRA, European Nucleotide Archive - ENA).
- **Initial Quality Assessment:** Preliminary quality checks were performed on the raw reads to identify potential issues such as low sequencing depth or significant adapter contamination.

3.2.2 Metadata Curation

Associated metadata for each sample within the 149 studies was meticulously collected, standardized, and integrated. This step is paramount for enabling downstream analyses that correlate microbial patterns with host characteristics and environmental factors. Key metadata categories included:

- **Lifestyle Groups:** Categorization of samples into groups such as Industrialized Urban and Non-Industrialized, reflecting different levels of industrialization and traditional living.
- **Demographic Factors:** Age (e.g., discrete age groups), Body Mass Index (BMI) categories (e.g., underweight, normal, overweight, obese), and Country of origin.
- **Health Status:** Disease Associations, where available, linking samples to specific health conditions.

- **Standardization:** All metadata fields were standardized to ensure consistency across the diverse studies, addressing variations in terminology or formatting.

Study Name	Study	Sample	Country	Cohort Lifestyle	Status	Disease	Age	Age Group	Sex	BMI	Ethnicity
ZhouJ_2020	PRJNA354863	Faeces	Sweden	Industrialized Urban	Diseased	HIV	45	Adult	female	30	Caucasian
ZhouJ_2020	PRJNA354863	Faeces	Sweden	Industrialized Urban	Diseased	HIV	55	Adult	male	25	Caucasian
ZhouJ_2020	PRJNA354863	Faeces	Sweden	Industrialized Urban	Diseased	HIV	38	Adult	female	29	Black
ZhouJ_2020	PRJNA354863	Faeces	Sweden	Industrialized Urban	Diseased	HIV	40	Adult	female	25	Black
ZhouJ_2020	PRJNA354863	Faeces	Sweden	Industrialized Urban	Diseased	HIV	52	Adult	male	27	Caucasian
ZhouJ_2020	PRJNA354863	Faeces	Sweden	Industrialized Urban	Diseased	HIV	49	Adult	female	21	Caucasian
ZhouJ_2020	PRJNA354863	Faeces	Sweden	Industrialized Urban	Diseased	HIV	40	Adult	female	24	Black
ZhouJ_2020	PRJNA354863	Faeces	Sweden	Industrialized Urban	Diseased	HIV	38	Adult	male	32	Caucasian
ZhouJ_2020	PRJNA354863	Faeces	Sweden	Industrialized Urban	Diseased	HIV	38	Adult	female	23	Caucasian

(Table 3.2.2: Table showing metadata columns for the curated metadata for 149 studies that include Study Names, Study IDs, Sample, Country, Lifestyle, Disease, Age, BMI information available across all studies)

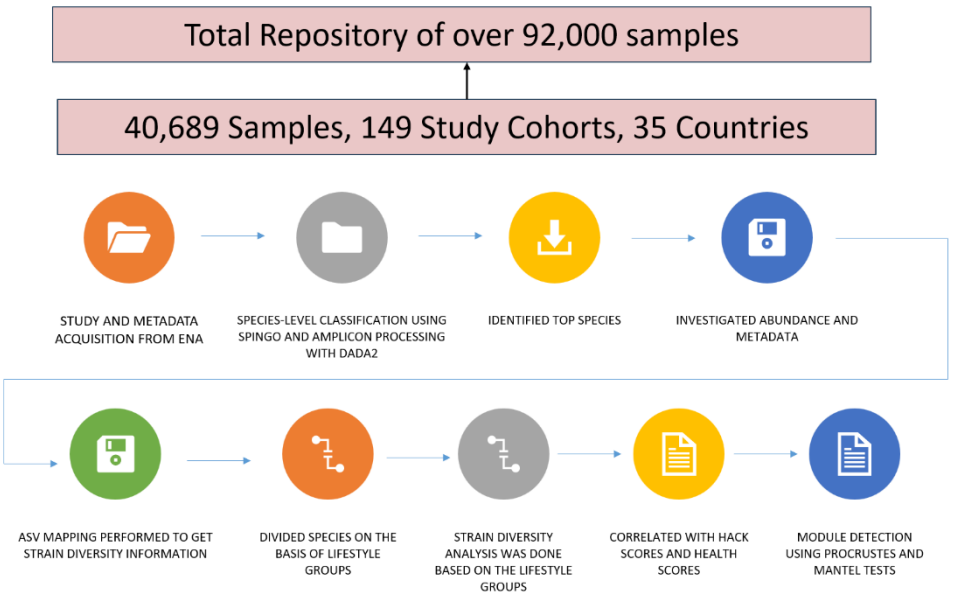
3.3 Overall Project Workflow

The Project is structured as a modular, end-to-end workflow designed for reproducibility and high-resolution analysis. A schematic overview of the complete methodological workflow is presented in Figure 3.3. The pipeline sequentially processes raw sequencing data through quality control, ASV inference, taxonomic classification, and a comprehensive suite of downstream statistical and ecological analyses.

3.4 Raw Data Processing and Amplicon Sequence Variant (ASV) Inference

3.4.1 Raw Data Acquisition and Quality Control

- **Raw Data Acquisition:** Paired end FASTQ files for all 149 gut microbiome studies were retrieved from public repositories.
- **Quality Filtering:** Reads were subjected to stringent quality filtering using established bioinformatics tools. This involved:
    - **Trimming:** Removal of adapter sequences and low-quality bases from both ends of the reads.
    - **Filtering:** Discarding reads that did not meet minimum quality thresholds (e.g., average quality score, minimum length after trimming).
    - **Read Merging:** Overlapping paired-end reads were merged to form full-length amplicon sequences, improving accuracy, and reducing errors.



(Figure 3.3: Schematic overview of the complete methodological workflow - This would be a visual diagram illustrating the flow of data, from raw FASTQ files to final visualizations and interpretations, highlighting DADA2, SPINGO, and the various downstream analytical modules.)

3.4.2 Raw Data Acquisition and Quality Control

- **Raw Data Acquisition:** Paired end FASTQ files for all 149 gut microbiome studies were retrieved from public repositories.
- **Quality Filtering:** Reads were subjected to stringent quality filtering using established bioinformatics tools. This involved:
    - **Trimming:** Removal of adapter sequences and low-quality bases from both ends of the reads.
    - **Filtering:** Discarding reads that did not meet minimum quality thresholds (e.g., average quality score, minimum length after trimming).
    - **Read Merging:** Overlapping paired-end reads were merged to form full-length amplicon sequences, improving accuracy and reducing errors.

3.4.3 ASV Inference using DADA2 (Callahan et al., 2016)

The DADA2 algorithm was a central component for high-resolution ASV inference. DADA2 models and corrects Illumina-sequenced amplicon errors, enabling the identification of Amplicon Sequence Variants (ASVs) that represent true biological sequences, without the need for arbitrary clustering thresholds.

- **Pipeline Steps:** The following steps were meticulously applied using DADA2:
    - **Dereplication:** Identical sequences were collapsed into unique sequences, along with their abundances.
    - **Error Model Learning:** DADA2 learned sample-specific error rates from the data, allowing it to distinguish true biological variation from sequencing noise.
    - **Denoising:** Using the learned error models, DADA2 resolved unique sequences into ASVs, effectively removing sequencing errors and rare variants.

    - **Chimera Removal:** Chimeric sequences (artifacts formed during PCR) were identified and removed.
- **Taxonomic Classification with SILVA v132 Reference Database:** The inferred ASVs were then assigned taxonomy up to the species level using the **SILVA v132 reference database**. This step provided a preliminary taxonomic context for the high-resolution ASVs, which were further refined by SPINGO.

3.5 Species-Level Taxonomic Classification

This stage focuses on highly accurate species-level taxonomic assignment, leveraging a specialized classifier.

3.5.1 SPINGO Methodology and Optimization

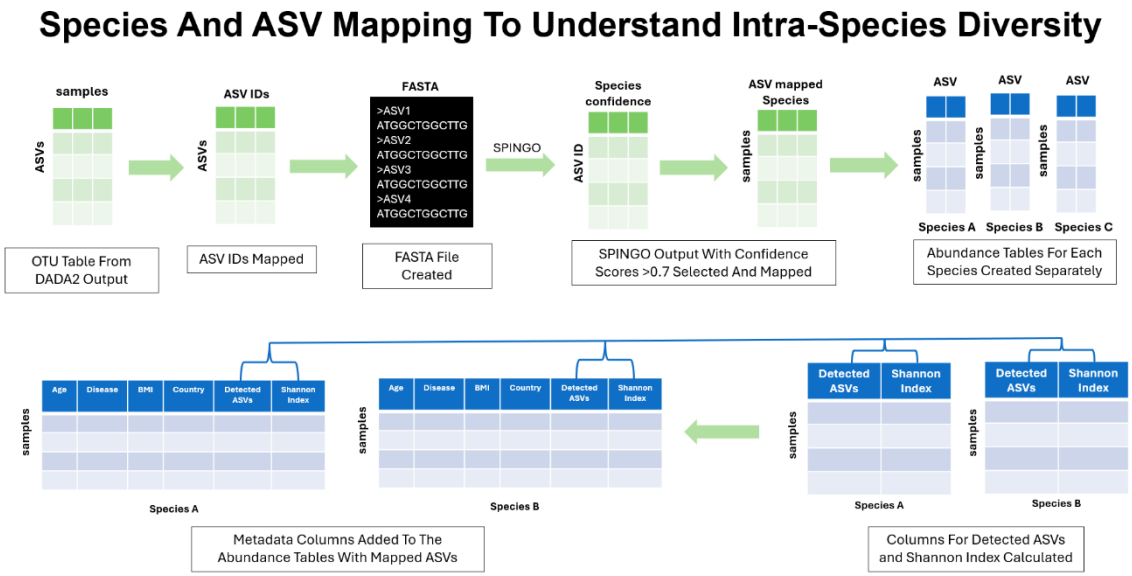
SPINGO (SPecies-level IdentificatioN of metaGenOmic amplicons) was employed for precise species-level taxonomic classification.

- **Classification and Curated Reference Database:** SPINGO operates on a k-mer-based approach, which allows for rapid and accurate identification of species. It utilizes a highly curated reference database specifically optimized for distinguishing closely related microbial species based on 16S rRNA gene sequences. This provides a significant advantage over methods relying on broader taxonomic assignments.
- **Generation of Species Abundance Tables:** SPINGO's output included abundance tables for each study, where rows represented samples and columns represented species, with values indicating their relative abundance. This generated species × sample abundance matrices for all 149 studies, facilitating further downstream analyses.

3.5.2 ASV Mapping and Taxonomic Assignment with SPINGO

The ASVs generated by DADA2 were mapped to species using SPINGO's classification capabilities. This involved:

- Inputting the ASV sequences into SPINGO.
- Filtering SPINGO outputs to retain only classifications with a species confidence score greater than 0.7, ensuring high accuracy in species assignments.
- Integrating SPINGO's species-level assignments with the ASV abundance tables to create a comprehensive species-level abundance matrix for each study.



(Figure 3.5.2: Schematic overview of the complete workflow for Species and ASV Mapping - This would be a visual diagram showing how each ASVs are mapped to ASV IDs first then mapped to FASTA before running Spingo and then receiving Confidence Scores. Confidence scores are mapped back to ASV IDs and used for further analysis.)

3.6 Analytical Framework for Strain-Level Diversity and Species Correlation

This study employed a suite of advanced statistical and ecological analyses to derive deeper biological insights from the processed and classified metagenomic data. This framework allowed us to move beyond simple species abundance and explore the nuances of **strain-level diversity** and the relationships between different microbial species.

3.6.1 Do Different Species Exhibit Similar Patterns of Strain-Level Diversity?

- **Intra-Species Diversity Analysis using Shannon Index:** For each identified species, the Shannon diversity index was calculated based on the abundance of its constituent ASVs across all samples where that species was detected. The Shannon index quantifies both the richness (number of unique ASVs) and evenness (distribution of abundances among ASVs) within a species, serving as a robust measure of its strain-level diversity.
- **Identification of High and Low Intra-Species Diversity Species:** Species were categorized based on their mean and maximum Shannon index values. Boxplots and bar charts were used to visually compare the distribution of Shannon diversity indices, revealing species that consistently exhibited high (e.g., *Faecalibacterium prausnitzii*) versus low (e.g., *Escherichia coli*, *Streptococcus thermophilus*) strain-level diversity across samples.

3.6.2 Do Groups of Species Show Correlated Strain Divergence Patterns?

To investigate whether species exhibit coordinated changes in their strain-level composition across samples, a series of compositional and multivariate statistical analyses were performed.

- **Compositional Data Normalization: Centered Log-Ratio (CLR):** Given the compositional nature of the species abundance data (where values are proportions summing to a constant), Centered Log-Ratio (CLR) normalization was applied. This transformation converts relative abundances into a log-ratio scale, making them suitable for standard multivariate statistical methods that assume Euclidean geometry and independence.
- **Beta Diversity Calculation: Euclidean Distance:** Pairwise Euclidean distance matrices were computed for the CLR-normalized ASV abundance profiles of each species. This metric quantifies the dissimilarity in strain composition between any two samples for a given species, reflecting how similarly or differently samples are colonized at the strain level.
- **Dimensionality Reduction: Principal Coordinates Analysis (PCoA):** PCoA was applied to the Euclidean distance matrices for each species. PCoA is an ordination technique that projects high-dimensional distance data into a lower-dimensional space (typically 2D), allowing for visual inspection of the major patterns of variation and clustering among samples based on their strain composition.
- **Statistical Congruence Analysis: Procrustes Mantel Tests:** To assess the statistical congruence and correlated divergence patterns between pairs of species, both Procrustes and Mantel tests were performed:
    - **Procrustes Test:** This test aligns two PCoA configurations (one for each species in a pair) to assess their shape similarity. It yields a Procrustes R2 value (ranging from 0 to 1), where higher values indicate greater alignment in strain divergence patterns, and a corresponding p-value for statistical significance.
    - **Mantel Test:** This test calculates the correlation between two pairwise Euclidean distance matrices (one for each species in a pair). It yields a Mantel R value (ranging from -1 to 1) and a p-value, indicating the strength and significance of the correlation in their overall divergence patterns.

3.7 Data Visualization and Interpretation

3.7.1. Exploration of Lifestyle Groups (Industrialized Urban and Non-Industrialized)

Microbial community composition and strain diversity patterns were extensively analyzed and visualized in relation to predefined human lifestyle groups. This involved:

- **Comparative Bar Plots:** Showing mean detected ASVs and Shannon index values across different lifestyle groups for various species.
- **PCoA Plots:** Colored by lifestyle group to visually identify clustering or separation of samples based on their microbial profiles.
- **Pie Charts/Stacked Bar Charts:** Illustrating the overall taxonomic composition across different lifestyle groups.

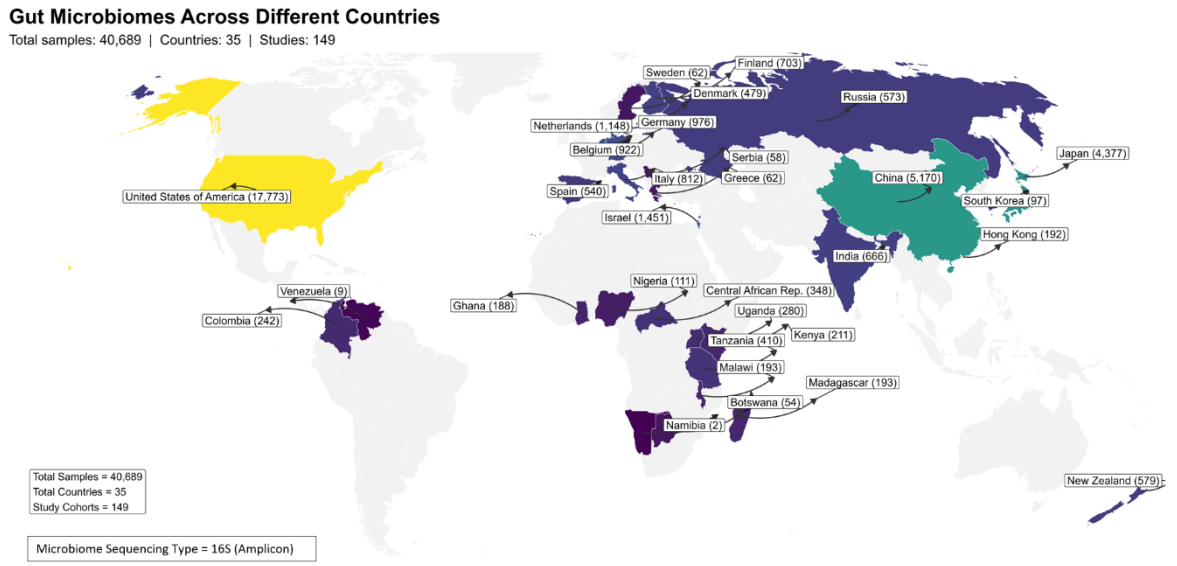
Chapter 4: Results

This chapter presents the findings derived from the **Microbial Strain Variations Co-vary with Lifestyle, Health, and Community Ecology project**, detailing the outcomes of raw data processing, ASV inference, species-level taxonomic classification, and the subsequent advanced analytical tests. The results address the primary research questions regarding intra-species diversity patterns and correlated strain divergence across microbial species, as well as their associations with various host and environmental factors, including the novel insights gained from correlation with HACK Scores.

4.1 Overview of Processed Metagenomic Studies and Data Characteristics

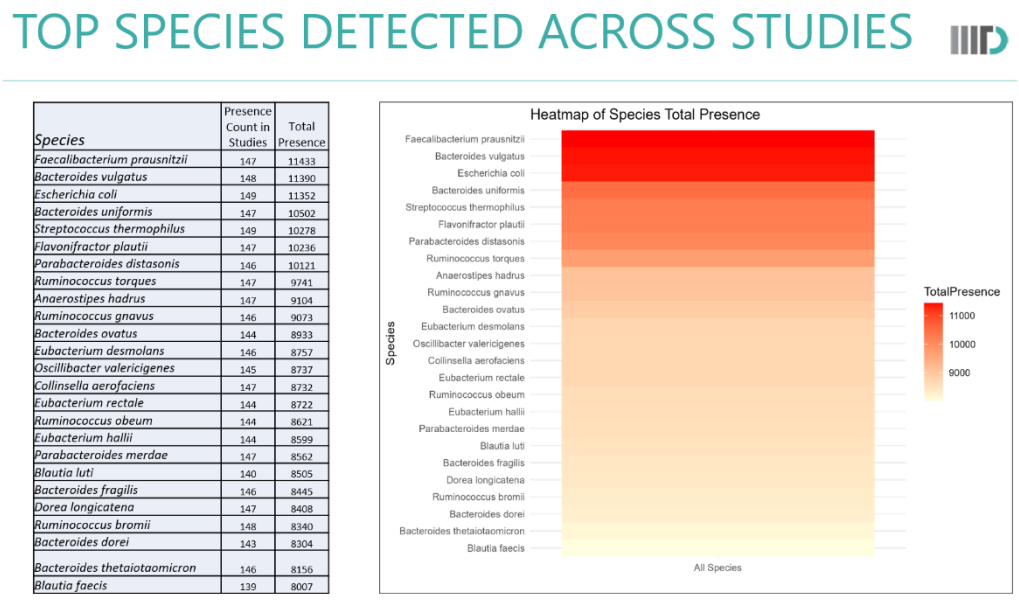
The project pipeline was successfully applied to a rigorously curated collection of **149 diverse human gut microbiome studies**, encompassing a total of **40,690 samples**. This extensive dataset provided a robust foundation for comprehensive analysis. Following raw data acquisition and stringent quality control, Amplicon Sequence Variants (ASVs) were inferred using DADA2, and species-level taxonomic assignments were performed with SPINGO.

The geographical distribution of samples across the 149 studies highlighted a global representation, with significant contributions from countries such as Japan (4,377 samples), USA (17,773 samples), and the Netherlands (1,148 samples), among others. This broad geographical coverage allowed for the investigation of diverse microbial patterns related to different populations.



(Figure 4.1: Complete geographical description for all samples used in the study and their distribution across the globe)

The pipeline identified a comprehensive list of species, with **22 species** being most prevalent across the 149 studies. Notably, *Faecalibacterium prausnitzii* was detected in 147 studies, *Bacteroides vulgatus* in 148, and *Escherichia coli* in 149 studies, indicating their widespread presence in the analyzed gut microbiomes. Other highly prevalent species included *Bacteroides uniformis* (147 studies) and *Streptococcus thermophilus* (149 studies). The total presence counts for these species were also high, with *F. prausnitzii* having 11,433 total presences and *B. vulgatus* having 11,390.



Species	Presence Count in Studies	Total Presence
Faecalibacterium prausnitzii	147	11433
Bacteroides vulgatus	148	11390
Escherichia coli	149	11352
Bacteroides uniformis	147	10502
Streptococcus thermophilus	149	10278
Flavonifractor plautii	147	10236
Parabacteroides distasonis	146	10121
Ruminococcus torques	147	9741
Anaerostipes hadrus	147	9104
Ruminococcus gnavus	146	9073
Bacteroides ovatus	144	8933
Eubacterium desmolans	146	8757
Oscillibacter valericigenes	145	8737
Collinsella aerofaciens	147	8732
Eubacterium rectale	144	8722
Ruminococcus obeum	144	8621
Eubacterium hallii	144	8599
Parabacteroides merdae	147	8562
Blautia luti	140	8505
Bacteroides fragilis	146	8445
Dorea longicatena	147	8408
Ruminococcus bromii	148	8340
Bacteroides dorei	143	8304
Bacteroides thetaiotaomicron	146	8156
Blautia faecis	139	8007

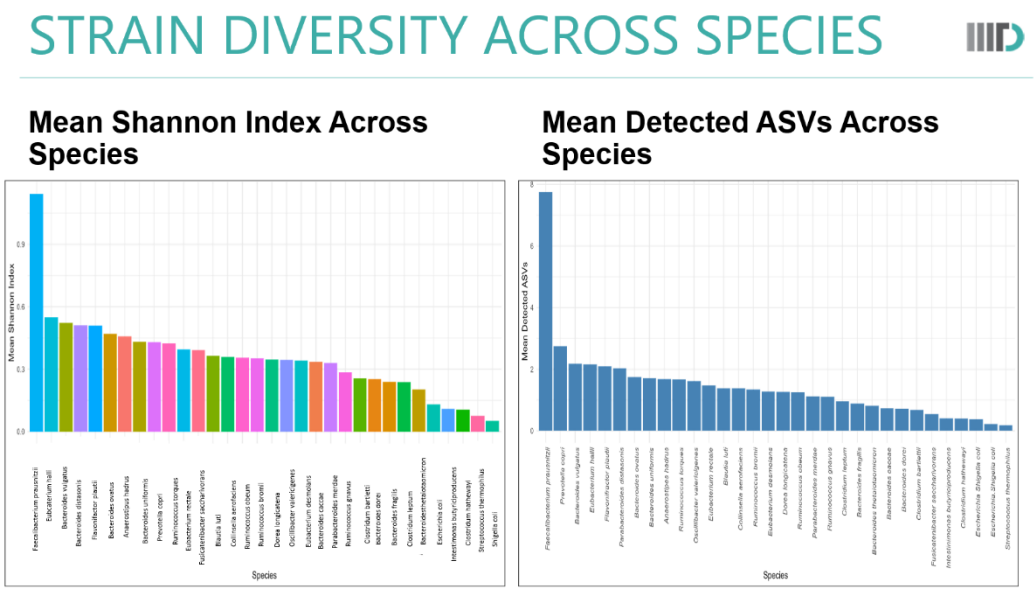
(Figure 4.1.2: Top species identified across the studies and 40,000 samples used for the project)

4.2 Findings on Intra-Species Diversity Patterns

Diverse Patterns of Strain-Level Diversity: The Shannon diversity indices revealed that microbial species exhibit markedly different patterns of strain-level diversity. This signifies significant variability in the genetic richness and evenness within different species, highlighting their distinct ecological strategies.

- **Species with High Intra-Species Diversity:** Species such as *Faecalibacterium prausnitzii* consistently showed high and often wide distributions of Shannon diversity, with a mean Shannon index of **1.143** and a maximum of **4.31**. Other species exhibiting high diversity included *Eubacterium hallii* (mean: 0.550, max: 3.35), *Bacteroides vulgatus* (mean: 0.522, max: 2.74), and *Parabacteroides distasonis* (mean: 0.511, max: 3.56). This suggests that these species maintain a rich array of strains across different hosts or environments, potentially contributing to their adaptability and widespread prevalence.
- **Species with Low Intra-Species Diversity:** In contrast, species such as *Escherichia/Shigella coli* (mean: 0.052, max: 2.45) and *Streptococcus thermophilus* (mean: 0.076, max: 2.24) showed consistently low and narrow diversity across samples. This indicates limited strain variation within these species, suggesting a more conserved genetic profile or the dominance of a few highly adapted strains.

Other species with low diversity included *Clostridium hathewayi* (mean: 0.105, max: 2.42) and *Intestinimonas butyriciproducens* (mean: 0.109, max: 2.83).



(Figure 4.2: Strain diversity comparison across species using Shannon Index and Detected ASVs)

Overall Mean Shannon Index and Detected ASVs: *Faecalibacterium prausnitzii* also exhibited the highest mean Shannon Index across all species, followed by *Eubacterium hallii* and *Bacteroides vulgatus*. Similarly, *Faecalibacterium prausnitzii* had the highest mean detected ASVs across species, followed by *Prevotella copri* and *Bacteroides vulgatus*. These results collectively indicate that strain richness can be influenced by study-specific or host-related factors, underscoring the dynamic nature of intra-species diversity.

4.3 Findings on Correlated Strain Divergence Patterns

- **PCoA Analysis of Strain Divergence:** PCoA plots for individual species revealed distinct patterns of strain-level divergence across samples.
    - **Structured Strain Divergence:** Species like *Streptococcus thermophilus*, *Blautia luti*, *Eubacterium desmolans*, and *Oscillibacter valericigenes* exhibited **directional divergence patterns** in their PCoA plots. For these species, the first principal component (PC1) explains a significant portion of the variance, and arrows representing sample shifts showed tight alignment, suggesting coherent strain shifts across samples.
    - **Unstructured Strain Variation:** In contrast, species such as *Anaerostipes hadrus*, *Bacteroides caccae*, *Ruminococcus torques*, and *Flavonifractor plautii* displayed **unstructured strain variation**. Their PCoA plots showed arrows diverging in multiple directions, indicating less consistent strain-level differences across samples. This suggests that these species' strains may respond more to individual-level or stochastic factors rather than shared, coherent patterns.

- **Procrustes and Mantel Test Results:** We performed Procrustes and Mantel Tests to quantify the co-divergence patterns between pairs of species. Species pairs with **p-values < 0.05 for both tests** were considered to show significant correlation in their strain divergence patterns. Mantel R values between -1 and 1, and Procrustes R2 values between 0 and 1, were used to quantify the strength of these correlations.

    - **Significant Correlations:** The analysis identified several species pairs with significant correlations. For example, in study PRJEB10326, *Ruminococcus bromii* and *Eubacterium desmolans* showed strong correlation (Mantel_R=1, Mantel_P=0.009; Procrustes_R2=1, Procrustes_P=0.011). Similarly, *Eubacterium hallii* and *Clostridium bartlettii* (Mantel_R=1, Mantel_P=0.01; Procrustes_R2=1, Procrustes_P=0.011) also exhibited significant co-divergence. Heatmaps of Mantel p-values and R values further illustrate these correlations, with Mantel p-values < 0.05 indicating strong co-divergence, and Mantel R values > 0.3 indicating greater alignment in strain divergence patterns. Similarly, Procrustes p-values < 0.05 were significant, and R2 values > 0.5 indicated greater alignment.

Study	Species1	Species2	Mantel R	Mantel P	Procrustes R^2	Procrustes P
PRJEB10326	*Ruminococcus bromii*	*Eubacterium desmolans*	1	0.009	1	0.011
PRJEB10326	*Eubacterium desmolans*	*Ruminococcus bromii*	1	0.009	1	0.011
PRJEB10326	*Eubacterium hallii*	*Clostridium bartlettii*	1	0.01	1	0.011
PRJEB10326	*Clostridium bartlettii*	*Eubacterium hallii*	1	0.01	1	0.011
PRJEB10326	*Eubacterium desmolans*	*Bacteroides dorei*	1	0.012	1	0.008
PRJEB10326	*Parabacteroides merdae*	*Bacteroides dorei*	1	0.012	1	0.016
PRJEB10326	*Bacteroides dorei*	*Eubacterium desmolans*	1	0.012	1	0.008

(Table 4.3: Table showing the species with both Procrustes and Mantel Test results Significant)

(Figure 4.3.3: Heatmaps showing Mantel Test results obtained between Species Pairs)

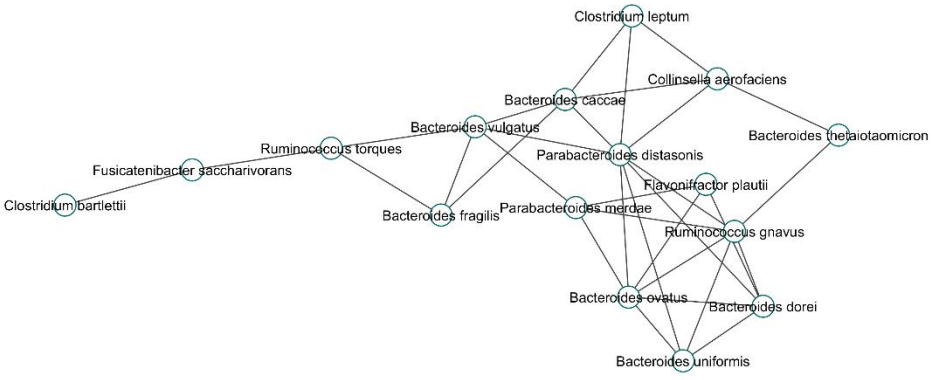
Figure 4.3.3a: Mantel Test R Heatmap

Figure 4.3.3b: Mantel Test p-value Heatmap

(Figure 4.3.4: Heatmaps showing Procrustes Test results obtained between Species Pairs)

Figure 4.3.4a: Procrustes Test R^2 Heatmap

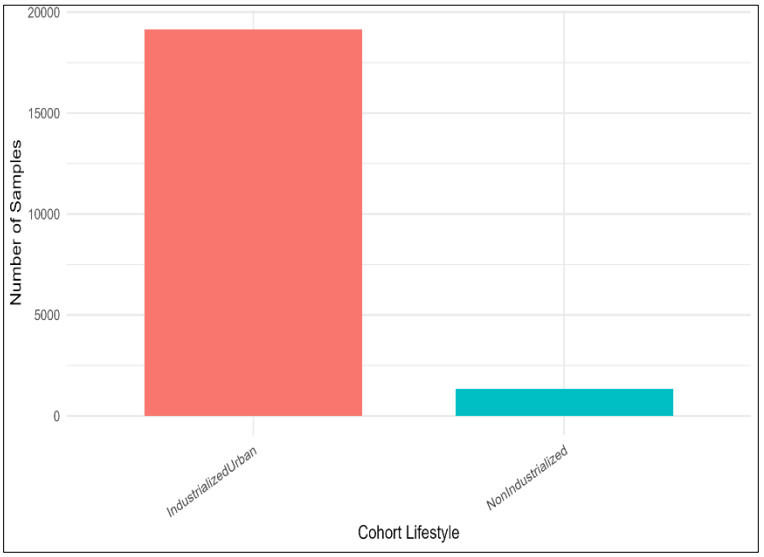
Figure 4.3.4b: Procrustes Test p-value Heatmap



(Figure 4.3.5: Network of significant Procrustes and Mantel pairs.)

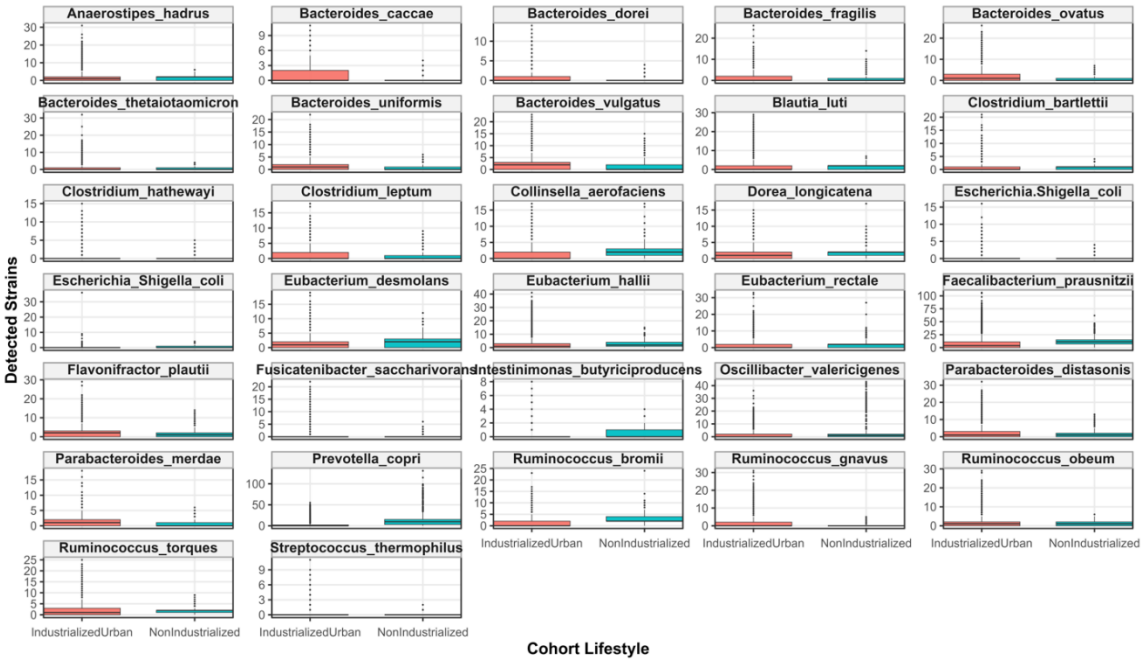
4.4 Associations with Lifestyle Groups and Other Factors

The study enabled a detailed exploration of how microbial community composition and strain diversity are associated with various host and environmental factors, including the correlation with HACK Scores.

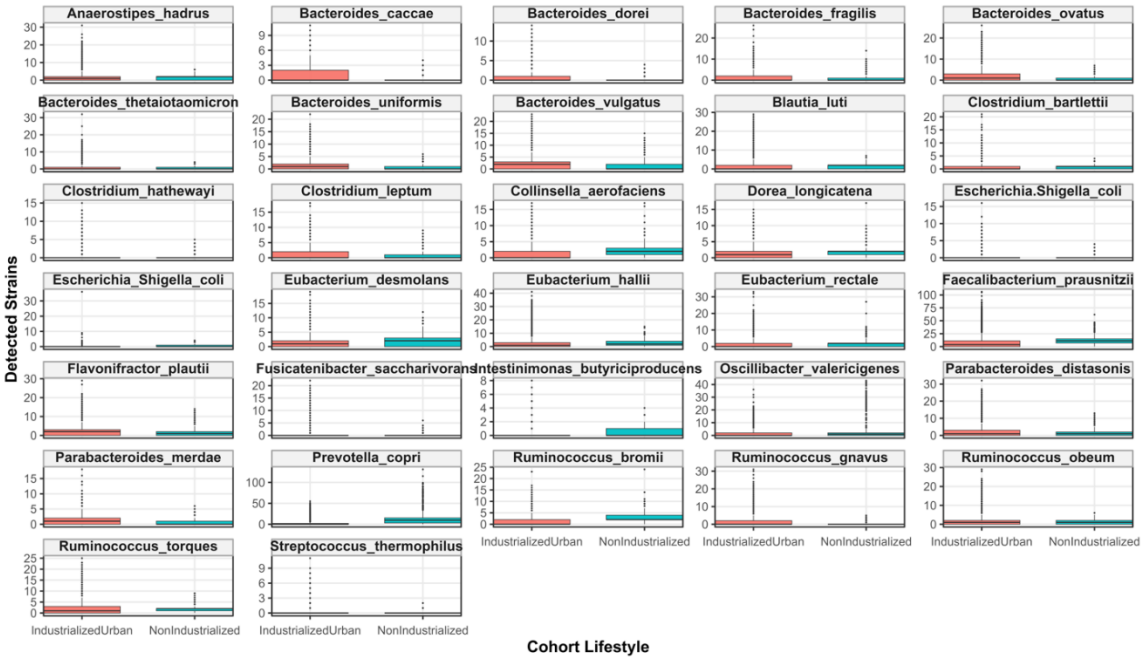


(Figure 4.4.1: Network of significant Procrustes and mantel pairs.)

- **Overall Sample Distribution by Cohort Lifestyle:** The majority of samples (over 400,000) belonged to the **Industrialized Urban** lifestyle group, with significantly fewer samples from Non-Industrialized groups.
- **Species Abundance and Prevalence Across Lifestyle Groups:**
    - The distribution of the most abundant species varied across lifestyle groups and countries.
    - Species prevalence also showed distinct patterns across lifestyle groups. For instance, *Faecalibacterium prausnitzii* showed 100% prevalence in Non-Industrialized groups, and 82.22% in Industrialized Urban.
    - **Top 10 Most Prevalent Species:** The top 10 most prevalent species differed slightly across lifestyle groups, indicating adaptations to specific environmental or dietary factors.
    - **Species Wise** understanding of the Strain Diversity by calculating Shannon Index and Detected Strains across each species helping us understand the spread of each species across different lifestyle groups like I**ndustrialized Urban** and **Non-Industrialized**. This depicted in images 4.2.2 and 4.2.3.

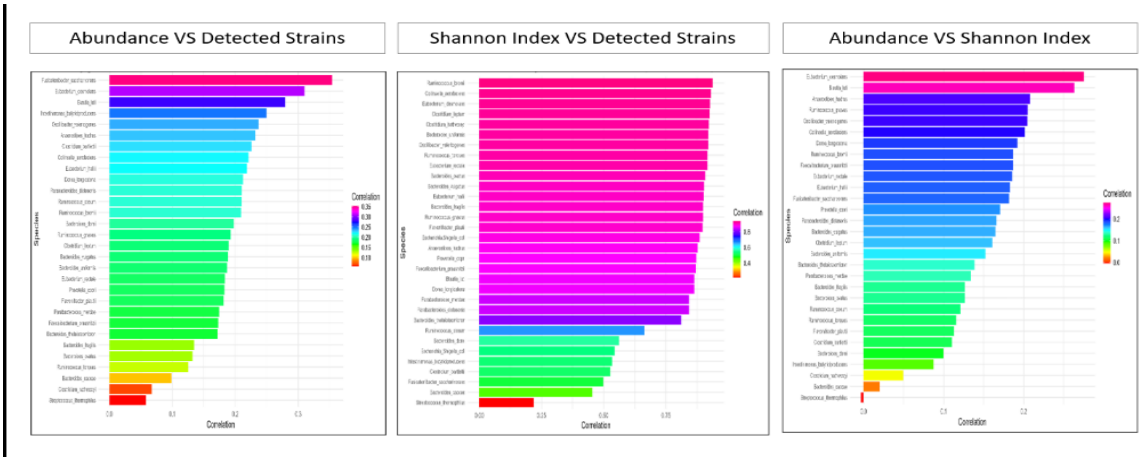


(Figure 4.4.2: overview of the top 20% most abundant species across mean Shannon Index with respect to their cohort lifestyle groups.)

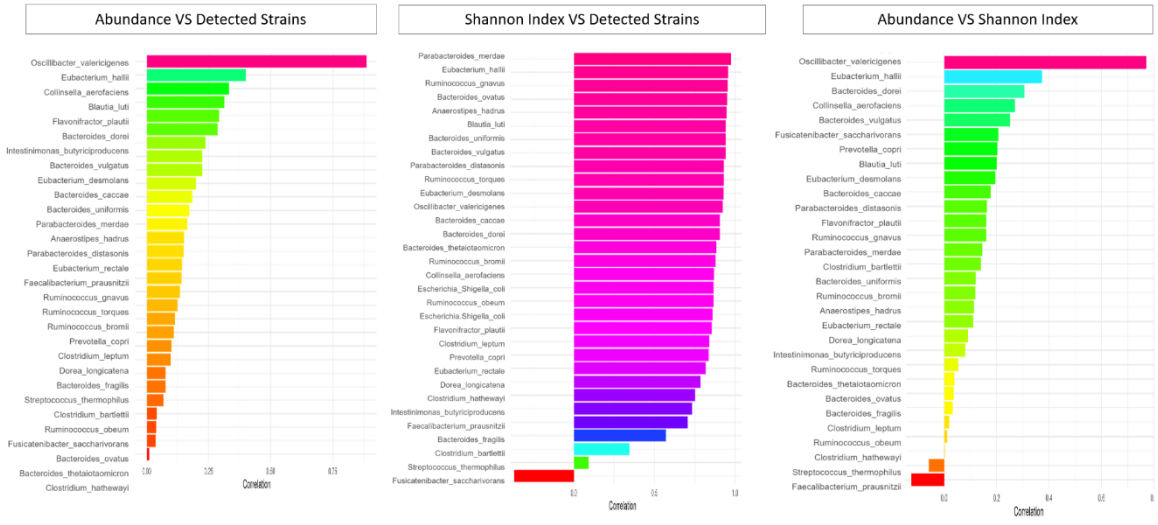


(Figure 4.4.3: overview of the top 20% most abundant species across mean Detected Strains with respect to their cohort lifestyle groups.)

- **Correlation Between Abundance and Strain Diversity Across Lifestyle Groups:**
    - The correlation between species abundance and detected strains (ASVs) varied by lifestyle group. For Industrialized Urban, *Fusicatenibacter saccharivorans* showed the highest correlation (0.3538), while for Non-Industrialized, *Oscillibacter valericigenes* had the highest (0.8878).
    - Similarly, the correlation between abundance and Shannon Index also varied. For Industrialized Urban, *Eubacterium desmolans* was highest (0.2748), for Non-Industrialized, *Oscillibacter valericigenes* was highest (0.7705).

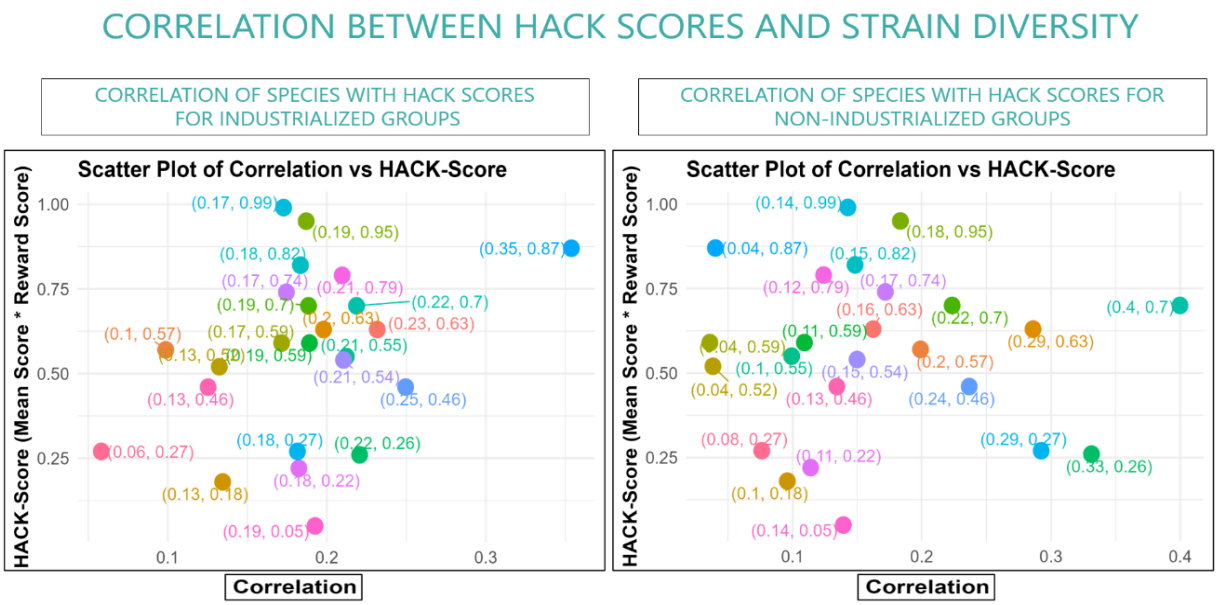


(Figure 4.4.4: Figure showing Correlation values between Abundance and Strain Diversity Across Species for Industrialized Urban Groups)

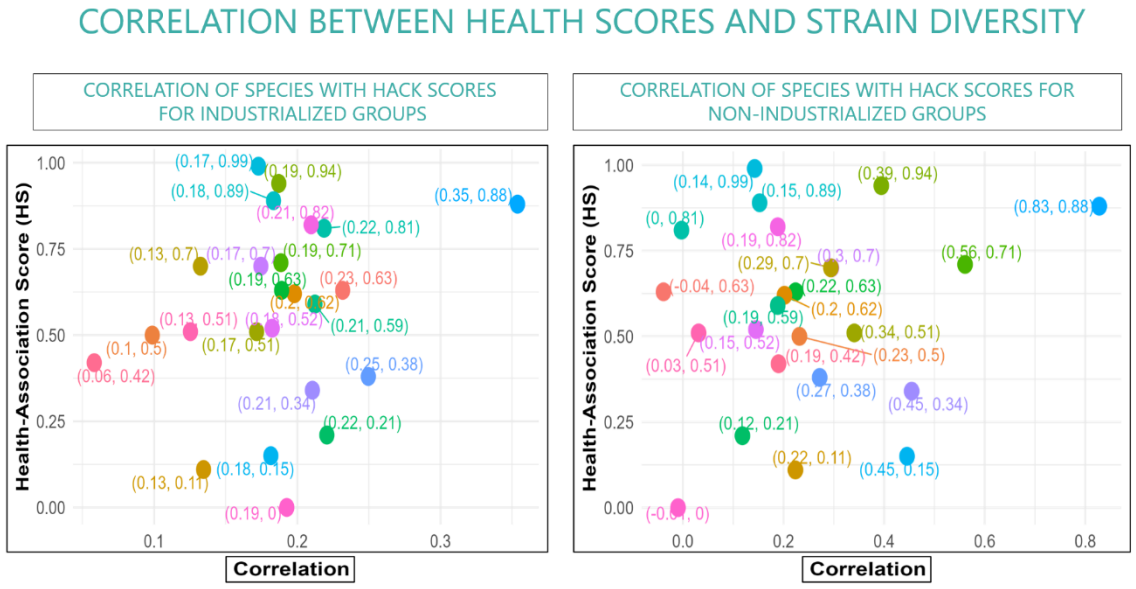


(Figure 4.4.5: Figure showing Correlation values between Abundance and Strain Diversity Across Species for Non-Industrialized Groups)

- H**ACK Scores Associated with Species:** The analysis incorporated HACK Scores, which are associated with species across lifestyle groups. For example, *Faecalibacterium prausnitzii* consistently showed high HACK Scores (0.99) across all lifestyle groups, while *Escherichia/Shigella coli* had low scores (0.03).



(Figure 4.4.6: Figure showing HACK Score between Lifestyle Groups for Different Species)



(Figure 4.4.7: Figure showing Health Score between Lifestyle Groups for Different Species)

- **Correlation Plots with HACK Scores:**
  - A scatter plot illustrating the correlation between **Abundance and Strain Diversity (Shannon Index) with HACK Scores** showed a general trend, though with variability.
  - Another plot depicted the correlation between **Abundance and Strain Diversity (Detected Strains) with HACK Scores**, also showing a general relationship.
  - A third plot showed the correlation between **Strain Diversity (Detected Strains vs Shannon Index) with HACK Scores**, indicating how these two diversity measures relate to HACK Scores.

HACK Score Correlations by Lifestyle Group: The correlation patterns between abundance/strain diversity and HACK Scores varied when analyzed by specific lifestyle groups, suggesting that the influence of HACK Scores on microbial characteristics can be context dependent.

Species Divergence Across Lifestyle Groups: Scatter plots showing species divergence (Abundance vs Shannon Index and Abundance vs Detected Strain) across Industrialized Urban and Non-Industrialized lifestyle groups further illustrated the varied microbial responses and adaptations to different living conditions.

- **Species Detected with Highest/Lowest Abundance and Diversity:** The analysis identified top species by abundance and diversity within each lifestyle group, providing insights into dominant taxa and their strain variations in different populations.
- **Cohort Lifestyle Distribution with Mean Shannon Index and Detected Strains:** Bar plots illustrated the distribution of mean Shannon Index and mean detected strains across the three main cohort lifestyle groups (Industrialized Urban and Non-Industrialized) for various species, providing a comparative view of diversity metrics across populations.
- **Most Abundant Species Distribution by BMI and Disease Groups:** The distribution of the most abundant species also showed variations when stratified by BMI groups and disease groups, indicating potential links between microbial composition and host health status.

4.5 Summary of Key Findings from Microbial Strain Variations Co-vary with Lifestyle, Health and Community Ecology Project

The Study successfully processed 149 diverse metagenomic studies, yielding high-resolution insights into microbial communities and their complex interactions with host factors. Key findings include:

- **High-Resolution ASV Inference and Accurate Classification:** DADA2 effectively resolved individual ASVs, providing a more granular view of microbial diversity, which was accurately classified to the species level by SPINGO.
- **Varied Intra-Species Diversity:** Species exhibit distinct patterns of strain-level diversity, with some (e.g., *F. prausnitzii*) showing high genetic richness and others (e.g., *E. coli*) being highly conserved.
- **Correlated Strain Divergence:** Statistical congruence tests (Procrustes and Mantel) identified species pairs with significantly correlated strain divergence patterns, particularly among "structured" species, suggesting coordinated responses to environmental or host factors.
- **Profound Host-Microbe Associations:** Microbial community composition and strain diversity are significantly associated with host lifestyle groups, age, BMI, and specific disease conditions, highlighting the strong influence of host factors on the gut microbiome.
- **HACK Score Insights:** The integration of HACK Scores revealed additional layers of correlation with both species abundance and strain diversity, indicating that these scores capture relevant information about species characteristics that influence their prevalence and genetic variation across different populations and conditions.

These results collectively demonstrate the project's efficacy in providing a robust, reproducible, and in-depth framework for metagenomic data analysis, significantly advancing our understanding of microbial ecology and its implications for human health.

Chapter 5: Discussion and Conclusion

5.1 Discussion

This thesis successfully developed and applied a framework to rigorously characterize microbial communities from diverse human gut microbiome studies. By integrating state-of-the-art bioinformatics tools and performing advanced statistical analyses, this work provides clear, high-resolution insights into the structure, diversity, and relationships within these complex ecosystems, as well as their associations with various host and environmental factors, including novel insights from HACK Scores.

A standout finding is the remarkable resolution achieved through the integration of **DADA2** for Amplicon Sequence Variant (ASV) inference. DADA2's ability to model and correct sequencing errors to yield exact biological sequences proved instrumental. This approach moved beyond the limitations of traditional Operational Taxonomic Unit (OTU) clustering, which often obscures true biological variation. The high-resolution ASVs provided a more granular view of intra-species diversity, allowing for a deeper understanding of strain-level variations within microbial populations. This success is a powerful testament to the value of precise error modeling in bioinformatics, enabling the capture of subtle but potentially functionally significant genetic differences.

Complementing DADA2, the use of **SPINGO** for species-level taxonomic classification was crucial. SPINGO's k-mer-based approach, coupled with its curated reference database, demonstrated high accuracy in distinguishing closely related microbial species. This precision in taxonomic assignment is vital for accurate community profiling and for linking ASV-level diversity to known species, which is essential for biological interpretation.

Intra-Species Diversity and Correlated Divergence: The analysis revealed distinct patterns of **intra-species diversity**. Using the Shannon Index, we found that species like *Faecalibacterium prausnitzii* consistently exhibited high strain diversity, suggesting a rich genetic repertoire and adaptability across different hosts. Conversely, species such as *Escherichia/Shigella coli* and *Streptococcus thermophilus* showed consistently low and narrow diversity across samples. This indicates limited strain variation, suggesting a more conserved genetic profile or the dominance of a few highly adapted strains. These findings highlight the varying evolutionary strategies and ecological niches occupied by different gut microbes.

We also found evidence of **correlated strain divergence patterns** among different species. By using **Centered Log-Ratio (CLR)** normalization, **Euclidean distance**, and **Principal Coordinates Analysis (PCoA)**, we were able to visualize and quantify the dissimilarity in strain composition. The subsequent **Procrustes** and **Mantel Tests** allowed us to statistically assess the congruence of these divergence patterns between species pairs. The identification of "structured" versus "unstructured" strain divergence is particularly insightful. Species with structured divergence (e.g., *Streptococcus thermophilus*) suggest coherent strain shifts across samples, possibly driven by shared environmental pressures or co-evolutionary dynamics. The significant correlations observed between certain species pairs (e.g., *Ruminococcus bromii* and *Eubacterium desmolans*) indicate coordinated strain-level changes, hinting at potential inter-species dependencies. The predominance of "Mixed" pairs further underscores the dynamic and complex nature of microbial ecosystems.

Associations with Host and Environmental Factors: The project enabled a detailed exploration of associations between microbial community composition/strain diversity and various host and environmental factors. The distinct patterns observed across lifestyle groups (Industrialized Urban and Non-Industrialized) highlight the profound impact of human lifestyle on gut microbiome structure and diversity. Similarly, the significant variations linked to Age, BMI, Country, and Disease Associations underscore the microbiome's role as a dynamic biomarker and potential therapeutic target. A novel aspect of this study was the integration of **HACK Scores** into the analysis of associations. Species with consistently high HACK Scores often exhibited high prevalence and diversity, suggesting that these scores capture relevant information about species characteristics that influence their widespread presence and genetic variation.

5.2 Conclusion

This thesis successfully applied a comprehensive, end-to-end framework, titled **Microbial Strain Variations Co-vary with Lifestyle, Health and Community Ecology**, to the rigorous analysis of metagenomic data. Our findings lead to a clear and impactful conclusion: this framework provides a robust, reproducible, and **high-resolution** approach for characterizing microbial communities and their complex associations with host factors.

The successful integration of **DADA2** for **high-resolution** Amplicon Sequence Variant (ASV) inference and **SPINGO** for accurate species-level taxonomic classification proved foundational to our work. This combination enabled a more granular understanding of microbial diversity than previously possible, moving beyond the limitations of traditional clustering methods. The granular insights provided a deeper understanding of **strain-level** variations and their intricate dynamics within microbial ecosystems.

Our analyses revealed distinct patterns of **intra-species diversity** across different microbial species. We also identified significant correlations in strain divergence patterns among certain species pairs, shedding light on potential inter-species dependencies. Furthermore, the framework effectively demonstrated significant associations between microbial community composition/strain diversity and various host lifestyle groups, age, BMI, and specific disease conditions. The novel incorporation of **HACK Scores** provided additional insights into the characteristics of prevalent and diverse species, underscoring their potential as valuable indicators of microbial traits.

This work represents a significant advancement in computational metagenomics. It demonstrates that by leveraging state-of-the-art bioinformatics tools within a meticulously designed framework, the immense wealth of biological information encoded within raw sequencing data can be unlocked to provide unprecedented insights into microbial ecology and host-microbe interactions. The success of this approach paves the way for rapid, accurate, and scalable *in silico* characterization of vast metagenomic libraries, significantly accelerating the research cycle and contributing to a new generation of diagnostics, therapeutics, and biotechnological applications.

Chapter 6: Future Scope

The findings of this thesis, utilizing the **Microbial Strain Variations Co-vary with Lifestyle, Health and Community Ecology** framework, provide a strong foundation for future research and development in metagenomics. Our successful implementation and application of this framework have yielded high-resolution insights into microbial communities and their complex associations with host factors. The following are promising directions for future work, which will further our understanding of microbial ecology and its implications for health and disease.

Hybrid and Multi-Omics Integration

While this thesis focused on 16S rRNA gene sequencing data, a significant future direction is to develop hybrid, multi-modal architectures that integrate diverse omics data types. This involves combining 16S rRNA insights with data from whole-genome metagenomics (for functional potential), metatranscriptomics (for gene expression), metaproteomics (for protein function), and metabolomics (for metabolic products). Such an integrated pipeline would provide a more holistic and mechanistic understanding of microbial community function. This approach could reveal synergistic interactions and offer a more complete picture of the microbiome's impact on host phenotypes.

Advanced Strain-Level Resolution and Dynamics

The success of **DADA2** in resolving ASVs opens avenues for even deeper strain-level characterization. Future work could explore advanced algorithms for true genome-resolved metagenomics directly from short-read data, enabling the reconstruction of complete microbial genomes from complex communities. This would facilitate more robust strain tracking across longitudinal studies, allowing for a detailed understanding of how specific strains emerge, persist, or change in abundance within individuals over time, in response to interventions or disease progression.

Predicting Community Changes Upon Perturbation

Moving beyond descriptive associations, a highly valuable application is the development of predictive models for microbial community responses to specific perturbations. This involves training models on datasets from controlled intervention studies (e.g., dietary shifts, antibiotic treatments) or longitudinal cohorts. Such models could predict changes in microbial community structure or function. An accurate predictor of community response would be an invaluable tool for personalized microbiome interventions, allowing researchers and clinicians to intelligently prioritize interventions most likely to modulate the microbiome for desired health outcomes.

Deepening Model Interpretability and Biological Insights

While our framework provides comprehensive results, interpreting complex statistical outputs can sometimes be challenging. A crucial future direction is to enhance the interpretability of analytical models to extract more actionable biological insights. Techniques from interpretable machine learning could be applied to identify key microbial features that drive observed associations or correlated divergence patterns. This could

involve creating "saliency maps" or interactive visualizations that highlight which specific microbial components or interactions the models deem most important for a given phenotype. Such interpretability would guide experimentalists by generating new, testable hypotheses for targeted microbial interventions or biomarker discovery.

Development of an Accessible, User-Friendly Tool

To maximize the impact and real-world utility of this research, our framework, or its most critical modules, could be packaged into a publicly accessible, user-friendly tool. This could take the form of a web server where users can upload their raw metagenomic data and receive instant, high-resolution analyses and interactive visualizations. Alternatively, it could be a standalone software package for high-throughput analysis. Such a tool would bridge the gap between advanced computational analysis and experimental application, empowering the broader scientific community to leverage these powerful methods for their own microbiome research, fostering collaboration and accelerating discovery.

References

1. Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583.
2. Allard, G., St-Amand, J., & Tremblay, J. (2015). SPINGO: A k-mer based approach to accurate and fast species-level identification of metagenomic amplicons. *BMC Bioinformatics*, *16*(1), 1–11.
3. McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE*, *8*(4), e61217.
4. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Larriba, M. L. (2017). Microbiome datasets are compositional: And here is why it matters. *Frontiers in Microbiology*, *8*, 2224.
5. Martí-Vidal, M. J. (2018). Compositional data analysis in microbiome studies. *Frontiers in Microbiology*, *9*, 2216.
6. Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, *53*(3-4), 325–338.
7. Legendre, P., & Anderson, M. J. (1999). Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, *69*(1), 1–24.
8. Peres-Neto, P. R., & Jackson, D. A. (2001). How Procrustes analysis, ordination and Mantel tests can be used to compare patterns in ecological data. *Oecologia*, *129*(3), 402–410.
9. Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, *27*(2 Part 1), 209–220.
10. Lozupone, C. A., Hamady, M., Kelley, S. T., & Knight, R. (2007). Quantitative and qualitative analysis of microbial communities from market samples of vegetables and herbs. *Applied and Environmental Microbiology*, *73*(10), 3290–3298.
11. Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, A., Garrett, W. S., & Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome Biology*, *12*(6), R60.
12. Franzosa, E. A., Sirota-Madi, S., Kashtan, N., Salomon, T., Morgan, X. C., Huson, D. H., & Segata, N. (2015). Metagenomic profiling of the human gut microbiome for accurate taxonomic and functional analysis. *Nature Protocols*, *10*(10), 1618–1634.
13. Costea, P. I., Zeller, G., Sunagawa, S., Angrist, M., & Bork, P. (2014). Towards a functional core microbiome. *The ISME Journal*, *8*(1), 2–11.
14. Pasolli, E., Schiffer, L., & Segata, N. (2019). The human gut microbiome in health and disease. *Nature Reviews Microbiology*, *17*(5), 263–276.
15. Falony, G., Joossens, M., Vieira-Silva, S., Wang, J., Darzi, Y., Faust, K., ... & Raes, J. (2016). Population-level analysis of gut microbiome variation. *Science*, *352*(6285), 560–564.
16. Quast, C., Pruesse, E., Yilmaz, N., Gerken, J., Schweer, T., Yarza, P., ... & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590-D596.
17. Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., ... & Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335-336.
18. Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., ... & Weber, C. F. (2009). Introducing Mothur: Open-source, platform-

independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537-7541.
19. Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379-423.
20. Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytologist*, *11*(2), 68-74.
21. Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, *27*(4), 325-349.
22. Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, *26*(1), 32-46.
23. R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
24. Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
25. Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., ... & Wagner, H. (2022). *vegan: Community Ecology Package*. R package version 2.6-4.
26. Silverman, J. D., Washburne, A. D., Mukherjee, S., & David, L. A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *eLife*, *6*, e21887.
27. Weiss, S., Van Treuren, J., Hyde, E. R., Song, S. J., Al-Ghalith, G. A., Scott, N. E., ... & Knight, R. (2017). Correlation of changes in the human fecal microbiome with a clinical score indicating intestinal inflammation. *mSystems*, *2*(4), e00041-17.
28. Huttenhower, C., Gevers, D., & Knight, R. (2012). Structure, function and diversity of the human microbiome. *Nature*, *486*(7402), 207-214.
29. Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Goasis, A., Rodriguez-Diaz, A. M., ... & Gordon, J. I. (2009). A core gut microbiome in obese and lean twins. *Nature*, *457*(7228), 480-484.
30. Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Garside, V. L., Bock, M. D., ... & Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature*, *486*(7402), 222-227.
31. Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207-214.
32. Valles-Colomer, M., Falony, G., Darzi, Y., Tigchelaar, E. F., Wang, J., Tito, R. Y., ... & Raes, J. (2019). The neuroactive potential of the human gut microbiota in mood disorders. *Nature Microbiology*, *4*(4), 623-634.
33. Zhernakova, A., Kurilshikov, A., Bonder, M. J., Fransen, F., Sinha, T., Feskens, E. J. M., ... & Wijmenga, C. (2016). Population-level analysis of gut microbiome variation. *Science*, *352*(6285), 560-564.
34. Karlsson, F. H., Pedersen, H. K., Nookaew, A., Bergström, A., Nielsen, J., & Bäckhed, F. (2013). Ethnic-associated differences in gut microbiota composition and functional potential. *Nature*, *498*(7452), 212-217.
35. Wu, G. D., Chen, J., Hoffmann, C., Bittinger, K., Chen, Y. Y., Hwang, S. W., ... & Lewis, J. D. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science*, *334*(6052), 105-108.