TUM SCHOOL OF LIFE SCIENCES

TECHNISCHE UNIVERSITÄT MÜNCHEN

Tailoring bioinformatics methods for studying the challenges in 16S rRNA gene sequencing data analysis

Monica Steffi Matchado

TUM SCHOOL OF LIFE SCIENCES

TECHNISCHE UNIVERSITÄT MÜNCHEN

Tailoring bioinformatics methods for studying the challenges in 16S rRNA gene sequencing data analysis

Monica Steffi Matchado

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung einer Doktorin der Naturwissenschaften (Dr. rer. nat.)

Vorsitz: Prof. Dr. Mathias Wilhelm

Prüfer*innen der Dissertation:

1. Prof. Dr. Jan Baumbach
2. Prof. Dr. Dirk Haller

Die Dissertation wurde am 31.05.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 04.10.2023 angenommen.

To my mom and dad, who have always been my pillars of strength and support. Thank you for your unwavering love, encouragement, and sacrifices that made this accomplishment possible. Your constant guidance, motivation, and belief in me have been invaluable. I am forever grateful for everything you have done for me.

Statement of Originality

Ich versichere, dass ich diese Dissertation selbstständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

I confirm that this dissertation is my own work and I have documented all sources and material used.

14.05.2023

Place and Date

Monica Steffi Matchado

Signature

Acknowledgements

First and foremost, I would like to express my heartfelt gratitude to my supervisor, Prof. Dr. Jan Baumbach, for giving me the invaluable opportunity to join his research group and embark on my doctoral journey under his expert guidance. I am deeply grateful for his unwavering support and mentorship throughout my research, which has been instrumental in shaping my research and career aspirations. Moreover, I would like to extend my sincere thanks to Prof. Dr. Dirk Haller for accepting to be on my supervisory committee and for providing insightful feedback and valuable suggestions during my TAC meeting.

I cannot thank my mentor, Dr. Markus List, enough for his invaluable guidance and constant support throughout my PhD journey. His dedication, expertise, and commitment to excellence have been an inspiration, and his constructive feedback and encouragement have challenged me to grow both personally and professionally. I am honored to have had the opportunity to learn from such an outstanding mentor who has taught me that perseverance, patience, and dedication are essential for success.

I would also like to express my appreciation to PD Dr. Klaus Neuhaus, Dr. Isabel Abellan Schneyder, Prof. Tim Kacprowski, Dr. Malte Rühlemann, and Dr. Fabian Frost for their guidance and expertise, which have been pivotal in identifying critical results in my thesis. I am grateful to Dr. Sandra Reimeiter for her continuous guidance and support throughout my research.

I extend my sincere gratitude to the CRC 1371 for providing the financial support that enabled me to pursue my doctoral program. I am grateful to have the support of my colleagues, including Alexander Dietrich, Markus Hoffman, who have taught me valuable techniques. Also I would like to thank Martina Ruettger, and Dr. Nina Kerstin Wenke who contributed to building a cheerful lab environment.

Finally, I would like to thank my parents, Duncan Matchado and Sophia Kumari Matchado, for their unwavering belief in me and their constant encouragement and support in all my decisions. I am also grateful to my brothers, Sharon Matchado and Rojer Matchado, for their care, trust, and relentless support in pursuing my goals. I would like to express my appreciation to my friends Nandhini and Shenbagam, Sharon and Lavanya, for their unwavering support and for cheering me up during difficult moments. Without the support of these individuals, I would not have been able to reach this far in my academic journey.

Abstract

In the field of microbial ecology, one of the most fundamental questions to ask is "who is there?" There are different ways to find an answer to this question; however, one of the gold standards that has stood the test of time is to sequence 16S ribosomal RNA (rRNA) gene. In the study of bacterial phylogeny and taxonomy, the use of 16S rRNA gene sequences has been by far the most prevalent housekeeping genetic marker employed for a number of reasons. However, usage of the 16S rRNA gene platform has limitations in both technical and computational aspects during different stages of analysis. For instance, during the amplification, primers can induce bias as they may bind to specific hypervariable regions, which are not 100% conservative. Similarly, there are different computational challenges such as the selection of proper pipelines, reference databases and parameters during downstream analysis. In addition, there is an ongoing debate on inferring functional profiles of microbial communities from 16S rRNA gene sequences. It remains an open question if metagenome prediction tools are also suited for more subtle contrasts related to human health. Comprehensive benchmark studies discussing these challenges are scarce. In this thesis, these computational challenges were studied. Moreover, we also focused on providing reliable solutions and recommendations for biomedical researchers. In a comparative study benchmarking different 16S rRNA gene pipelines, we demonstrated that targeting the variable region V3-V4 of the 16S rRNA gene enabled the most precise characterization compared to other primer regions. Amplicon sequence variants or zero-radius operational taxonomic units were found to be the best option for taxonomic characterization, especially when using SILVA or RDP as the reference databases. In a systematic benchmark of 16S-based functional inference tools, I found that these tools had varying levels of accuracy and precision in functional predictions. Among selected functional inference tools, *PICRUSt2* outperformed the other tools but failed to provide accurate results as compared to metagenome functional diversity. I recommended the use of a combination of methods so that a thorough evaluation of the results can bolster the overall performance. In summary, functional prediction tools based on 16S rRNA gene data results in limited sensitivity in detecting differences related to human health and disease. Hence, it can be used in hypothesis generation in conjunction with other methods, such as metagenome profiling. Combining all the best pipelines and methods that we identified from our benchmark analyses, we developed a user-friendly tool called *Namco*, which provides end-to-end analysis for 16S rRNA gene analysis that covers upstream analyses such as raw data processing and taxonomic binning and downstream analyses including basic statistics, machine learning and network analysis, among other features. Overall, the analysis of 16S rRNA gene amplicon data requires careful consideration of experimental design, data processing and downstream analysis methods to ensure accurate and reliable results. Findings from each chapter can help researchers to make informed decisions about which primers to use based on their research goals and which tools to select for 16S rRNA gene-based functional profiling. Finally, integrating my recommendations into a user-friendly tool can help researchers to carry out microbial analysis more efficiently and accurately.

Kurzfassung

Auf dem Gebiet der mikrobiellen Ökologie ist eine der grundlegendsten Fragen, die man sich stellen muss, "Wer ist da?". Es gibt verschiedene Möglichkeiten, eine Antwort auf diese Frage zu finden; einer der Goldstandards, der sich seit langem bewährt hat, ist jedoch die Sequenzierung von 16S ribosomalen RNA (rRNA)-Genen und die Charakterisierung der mikrobiellen Population. Bei der Untersuchung der bakteriellen Phylogenie und Taxonomie ist die Verwendung von 16S rRNA-Gensequenzen aus verschiedenen Gründen der bei weitem am häufigsten verwendete genetische Housekeeping-Marker. Die Verwendung der 16S rRNA-Genplattform unterliegt jedoch in den verschiedenen Phasen der Analyse sowohl technischen als auch bioinformatischen Einschränkungen. So führen die für die Amplifikation verwendeten Primer während der Amplifikation zu Verzerrungen, da sie an spezifische hypervariable Regionen binden, die nicht zu 100 % konserviert sind. In ähnlicher Weise gibt es verschiedene bioinformatische Herausforderungen wie die Auswahl geeigneter Pipelines, Referenzdatenbanken und Parameter während der nachgeschalteten Analyse. Darüber wird über die Ableitung von Funktionsprofilen mikrobieller Gemeinschaften aus 16S rRNA-Gensequenzen debattiert. Es bleibt eine offene Frage, ob Metagenom-Vorhersagewerkzeuge auch für subtilere Kontraste im Zusammenhang mit der menschlichen Gesundheit geeignet sind. Umfassende Benchmark-Studien, die diese Herausforderungen diskutieren, sind rar. In dieser Arbeit wurden diese bioinformatischen Herausforderungen bei der Sequenzierung von 16S rRNA-Genen untersucht. Darüber hinaus konzentrierten wir uns auf die Bereitstellung zuverlässiger Lösungen und Empfehlungen für die biomedizinische Forschung. In einer vergleichenden Studie, in der verschiedene 16S rRNA-Gen-Pipelines verglichen wurden, haben wir gezeigt, dass die variable Region V3-V4 des 16S rRNA-Gens im Vergleich zu anderen Primerregionen die präziseste Charakterisierung ermöglicht. Amplikon-Sequenzvarianten oder operative taxonomische Einheiten mit Nullradius erwiesen sich als die beste Option für die taxonomische Charakterisierung, insbesondere bei Verwendung von SILVA oder RDP als Referenzdatenbanken. In einem systematischen Benchmarking von 16S-basierten Werkzeugen für funktionelle Inferenzen stellte ich fest, dass diese Werkzeuge bei funktionellen Vorhersagen ein unterschiedliches Maß an Genauigkeit und Präzision aufweisen. Unter den ausgewählten Werkzeugen für die funktionelle Inferenz übertraf PICRUSt2 die anderen Werkzeuge, lieferte jedoch im Vergleich zur funktionellen Vielfalt des Metagenoms keine genauen Ergebnisse. Ich empfahl die Verwendung einer Kombination von Methoden, damit eine gründliche Bewertung der Ergebnisse die Gesamtleistung verbessern kann. Zusammenfassend lässt sich sagen, dass funktionelle Vorhersageinstrumente auf der Grundlage von 16S rRNA-Gen-Daten nur eine begrenzte Sensitivität bei der Erkennung von Unterschieden im Zusammenhang mit menschlicher Gesundheit und Krankheit aufweisen. Daher können sie bei der Hypothesenbildung in Verbindung mit anderen Methoden, wie z. B. der Metagenomprofilierung, eingesetzt werden. Durch die Kombination der besten Pipelines und Methoden, die wir bei unseren Benchmark-Analysen identifiziert haben, haben wir ein benutzerfreundliches Tool namens *Namco* entwickelt, das eine umfassende Analyse für die 16S rRNA-Genanalyse bietet, die u. a. vorgelagerte Analysen wie die Verarbeitung von Rohdaten und taxonomisches Binning sowie nachgelagerte Analysen einschließlich grundlegender Statistik, maschinelles Lernen und

Netzwerkanalysen umfasst. Insgesamt erfordert die Analyse von 16S rRNA-Genamplikondaten eine sorgfältige Berücksichtigung der Versuchsplanung, der Datenverarbeitung und der nachgeschalteten Analysemethoden, um genaue und zuverlässige Ergebnisse zu gewährleisten. Die Erkenntnisse aus den einzelnen Kapiteln können Forschenden dabei helfen, fundierte Entscheidungen darüber zu treffen, welche Primer sie je nach ihren Forschungszielen verwenden und welche Tools sie für die 16S rRNA-Gen-basierte funktionelle Profilerstellung auswählen sollten. Schließlich kann die Integration meiner Empfehlungen in ein benutzerfreundliches Werkzeug helfen, mikrobielle Analysen effizienter und genauer durchzuführen.

Graphical abstract

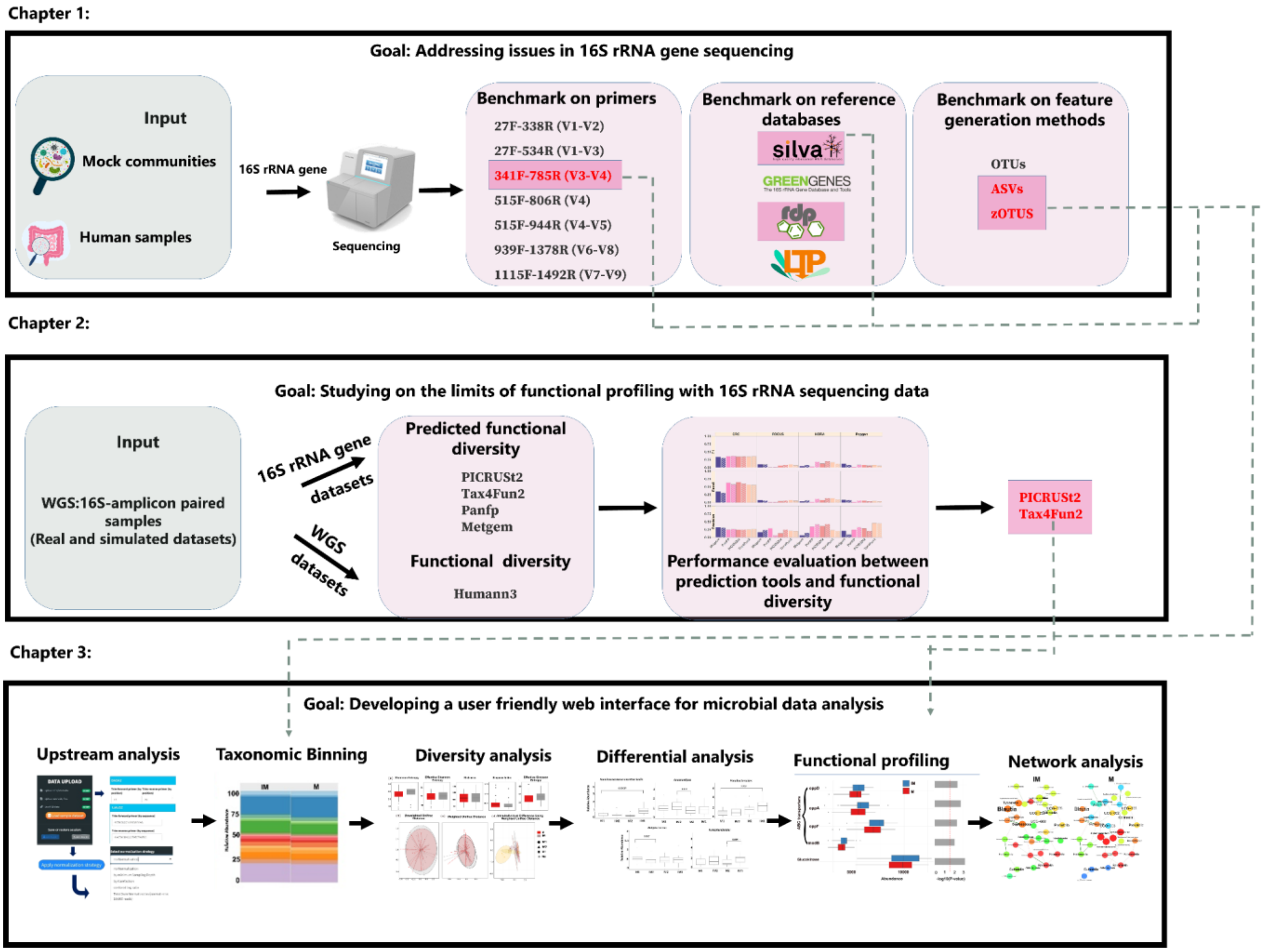


Figure 1: Overview of the thesis
This thesis mainly focuses on different computational challenges in 16S rRNA gene data analysis. In the first chapter, the influence of selection of primer, pipeline and parameters were studied. In the second chapter, the focus was on analysing the accuracy of functional prediction tools using 16S rRNA. The best and recommended outcomes under each benchmark analysis (in chapter one and chapter two) are highlighted in dark pink boxes. Dotted lines explains that best candidates obtained from each analysis are incorporated in the tool called Namco, which is the final chapter of this thesis. (Source: own work)

Contents

Abstract....4

Kurzfassung....5

Graphical abstract....7

1. Chapter one: Introduction....13

1.1. The vocabulary of microbiome research 13

1.2. The human gut microbiome 14

1.2.1. Factors shaping the GI microbiota....14

1.2.2. Human Microbiome Project....15

1.2.3. Second phase of HMP....16

1.2.4. Useful resources from the HMP....16

1.3. The 16S rRNA gene as a taxonomical marker 17

1.3.1. Other marker genes....19

1.4. Workflow of 16S/18S/ITS amplicon sequencing 20

1.4.1. Study design....21

1.4.2. Pre-processing of raw amplicon reads....22

1.4.3. Sequencing and analysis....22

1.5. Challenges in 16S rRNA gene amplicon data analysis 24

1.5.1. Compositionality....24

1.5.2. The challenge of sparsity....27

1.5.3. Functional prediction from 16S rRNA gene sequences....28

1.6. Metagenomics 29

1.6.1. Metagenome assembly....29

1.6.2. Assembly-free metagenomic profiling....31

1.6.3. Genes and metabolic pathways from metagenomes....32

1.6.4. Post-processing analysis....32

1.6.5. Advantages and challenges in metagenomics....33

1.7. Problem statement 34

1.8. Objectives 36

2. Chapter two: Primer, pipeline, parameters: Issues in 16S rRNA gene sequencing....38

2.1. Declaration of contributions 38

2.2. Introduction 39

2.2.1. Issues in 16S rRNA gene sequencing 39

2.3. Material and methods 46

2.3.1. Preparation of human gut samples and mock communities 46

2.3.2. Primer-specific feature classifiers 46

2.3.3. OTU clustering using QIIME1 46

2.3.4. zOTU generation using UNOISE 46

2.3.5. ASV generation using nf-core/ampliseq pipeline 47

2.4. Results 47

2.4.1. Clustering strategies have minor effect on taxonomic profiles 50

2.4.2. Reference databases have an impact on the taxonomic assignment 50

2.4.3. Specific pipeline settings have minor influences on taxonomic classification..... 53

2.5. Discussion 56

2.6. Conclusion and outlook 58

3. Chapter three: On the limits of 16S-based metagenome prediction and functional profiling 61

3.1. Declaration of contributions 61

3.2. Introduction 62

3.2.1. The Human Gut Microbiome: Function Is Crucial 62

3.2.2. Functional omics 62

3.2.3. Metagenome approach for functional characterization 63

3.2.4. Limitations of metagenome approach for functional characterization 65

3.2.5. Functional profile prediction using 16S rRNA gene sequences 65

3.2.6. Functional analysis methods 67

3.2.7. Introduction to metagenome prediction tools 69

3.2.8. Previous benchmarking studies in the literature 76

3.2.9. Importance of 16S copy number normalisation 78

3.3. Problem statement 79

3.4. Materials and methods 80

3.4.1. Population-based cohorts 80

3.4.2. Bioinformatics pipeline for processing of 16S-rRNA Data 82

3.4.3. Generating functional prediction profiles 83

3.4.4. Functional profiling of metagenomics 83

3.4.5. Customised normalisation using rrnDB 83

3.4.6. Validating prediction tools with shotgun metagenomic sequencing data 83

3.4.7. Cohort-wise differential abundance analysis between cases 84
3.4.8. Simulation dataset and processing...84
3.5. Result 85
3.5.1. Correlation is not a suitable performance measure for metagenome prediction tools 85
3.5.2. Metagenome prediction tools except MetGEM show high specificity in predicting KO terms...86
3.5.3. Metagenome prediction tools show low accuracy in predicting differentially abundant KO terms 86
3.5.4. Performance of functional inference tools at the level 1 KEGG functional categories 89
3.5.5. Metagenome prediction performance varies widely across functional categories89
3.5.6. Functional Profiling on simulation datasets 100
3.6. Discussion 101
3.7. Conclusion and outlook 103
4. Chapter four: Namco, a Free microbiome explorer...105
4.1. Declaration of contribution 105
4.2. Introduction 106
4.2.1. Bioinformatics pipelines for 16S rRNA gene analysis......................................107
4.2.2. Statistical analysis of microbiome data with R..107
4.2.4. Limitation of existing tools...109
4.3. Materials and methods 110
4.3.1. Namco: framework 110
4.3.2. Upstream analysis..112
4.3.3. Downstream analysis 114
4.3.4. Use case 120
4.4. Research question 123
4.5. Results 124
4.5.1. Diversity analysis...124
4.5.2. Taxonomic distribution..125
4.5.3. Correlation of gut microbial composition and clinical metadata........................128
4.5.4. Functional analysis 130
4.5.5. Network analysis 132
4.6. Discussion 133

4.7. Conclusion and outlook 134

5. General discussion 135

5.1. Are short-read amplicons suitable for the prediction of microbiome functional potential? 138

5.2. Full-length 16S rRNA gene sequencing 138

5.3. Shallow metagenomics: Maybe a better alternative to 16S rRNA gene sequencing 138

5.4. Employing multi-omics approaches toward microbiome research 139

5.5. Adaptation of machine learning models coupled with other models in microbiome studies 140

5.6. Extensive bioinformatics tools are needed 141

6. Future perspective of the thesis 143

References 146

Publications 185

Manuscript in preparation: 185

List of figures 186

List of tables 189

Abbreviations 190

Chapter one: Introduction

1. Chapter one: Introduction

1.1. The vocabulary of microbiome research

The term "microbial community" refers to a group of coexisting microorganisms. More specifically, multi-species assemblages of (micro)organisms that interact with one another in a continuous environment are referred to as microbial communities. The term "microbiome" was initially defined in 1988 by Whipps and colleagues who studied the ecology of rhizosphere microorganisms [1,2]. They defined the "microbiome" as the combination of the terms “micro” and “biome”, designating as its “theatre of activity a characteristic microbial community” in a “fairly well-defined habitat with specific physio-chemical features”.

Due to the rapid development of this field, there is an uncertainty regarding the terminology used to describe the many aspects of these communities. The incorrect usage of terminology such as “metabolomic”, “metagenome”, “metagenomics”, and “microbiome”, among other words, has led to misinterpretation and confusion by both the scientific and general public about many study findings [3]. Some of the important vocabulary used in microbial research are as follows:

- **Microbiome**: refers to the entire micro-organisms including archaea, bacteria, viruses, lower and higher eukaryotes and their genomes [3] along with their ecosystems
- **Microbiota:** refers to the collection of microorganisms that inhabit a particular environment
- **Metataxonomics:** is defined as a high-throughput method for characterising the complete microbiota and producing a metataxonomic tree that illustrates the relationship between all sequences.
- **Metagenome:** refers to the collection of genes and genomes from microbiota
- **Microbiota diversity**: a metric that measures the number of distinct species present and, depending on the diversity indices used, the evenness of their distribution within the community.

Studies involving metataxonomic analyses that depend on 16S rRNA gene sequencing and analysis frequently make mistakes in terms of terminology. Many times, terms such as "16S survey", 16S sequencing, and analysis are used in the literature. It is not a correct term. The "S", in 16S, stands for the Svedberg uniunit, a non SI unit for sedimentation. The subunits that make up the 30S small ribosomes of bacterial, archaeal, and bacterial ribosomes contain 21 proteins. There is also one structural 16SRNA. This rRNA has approximately 1540 ncleotides.

1.2. The human gut microbiome

The human body is one of the suitable habitats for microbial communities. The human hosts different microbiota including bacteria, protozoa, archaea, and viruses on and within different sites of the body such as the gastrointestinal system (GI)**,** genital organs, respiratory tract, skin and oral cavity [4,5]**.** Among different body sites, the GI tract hosts 10^{14} microbiome, which is nearly 10 times more bacteria cells compared to human cells and 100 times more microbiome as compared to the human genome [6,7]. It is generally believed that the diversity or development of the microbiome originates from the placenta [8–10] and diverges throughout the human lifespan. In the early stages, microbiomes are known to be dominated by the main phyla, which are Proteobacteria and Actinobacteria [11]. The diversity of the gut microbiome becomes comparably stable during adulthood and then decreases in the elderly age [12]. In young adults, *Clostridium cluster XIVa* is found to be dominant, in contrast, in elderly years, the abundance of Bacteroidetes phyla and *Clostridium cluster IV* are found to be the dominant groups [13].

Gut microbiota plays a crucial role in maintaining human health by facilitating several important functions such as including food fermentation, host defence, immune response stimulation, and vitamin generation [14]. For example, the gut microbiome helps to train the immune system to recognize and respond to pathogens, while also helping to prevent overactive immune responses to harmless antigens. The gut microbiome also aids in the development of gut-associated lymphoid tissue (GALT), which is a key component of the immune system. Additionally, certain microorganisms produce anti-inflammatory compounds that can help to regulate the immune response and maintain immune homeostasis [15]. In the meanwhile, dysbiosis or an imbalance in the gut microbiota is also proved to be a major contributor to the pathogenesis of extra-intestinal disorders such as allergies, asthma [16], type 1 diabetes [17], cardiovascular disease [18], metabolic syndrome, and obesity as well as intestinal disorders like inflammatory bowel disease (IBD) [19,20] and irritable bowel syndrome (IBS) [21]. Commensal gut bacteria may also be crucial in reducing inflammation and bacteremia, according to certain theories [22]. However, it should be noted that the composition of the microbiome can vary greatly between individuals and can also change over time within an individual due to a variety of factors such as lifestyle, dietary habits, illness, antibiotic treatment etc [23,24].

1.2.1. Factors shaping the GI microbiota

There is a plethora of internal and external factors that modify gut microbial compositions and ultimately affect human health. Change of microbial intestinal colonisation begins concerning the delivery pattern [25]. There are a couple of studies that proved that the gut microbial composition of infants born by vaginal delivery is different from infants delivered by caesarean [26,27]. Infants born via caesarean are more prone to the risk of developing different diseases such as juvenile arthritis, inflammatory bowel disease, immune deficiencies, leukaemia and asthma [28,29].

Diet is one of the important driving factors in changing the gut microbiome. Several interventional studies had proven that dietary habits play a crucial role in the substantial changes in the composition [30,31]. Major dietary options such as animal or plant-based diets result in alterations of the gut microbiome [30]. For example, fibre-enriched diets help in increasing beneficial bacteria which helps in increasing the production of short-chain fatty acids (SCFA), which in turn helps in repairing dysfunctional metabolism [32]. In contrast to fibre-enriched diets, high-fat diets decrease bacterial diversity and also reduce the production of SCFA [33]. A high fat diet also leads to several diseases including obesity and type 2 diabetes [34].

Other than diet and delivery methods, genetics and medications such as antibiotics are also responsible for changing microbial compositions. Few studies have shown that blood relatives happen to have more similar microbiota than the unrelated and also monozygotic twins have more similar microbial communities compared to dizygotic twins [35]. Nonetheless, so far, there are no genome-wide studies available to prove the relationship between genes and pathways and gut microbiome composition [36].

A large number of small microbiome studies have been published, and the use of high-throughput sequencing technologies has led to a rapid increase in the amount of microbiome data being generated. It is difficult to provide an exact number, but a search on the SRA [37] database using the terms “16S rRNA” AND “human gut metagenome” results in 638 bioprojects with 17,322 biosamples and 104,855 datasets as of February 2023. A search on pubmed using “16S rRNA gene sequencing” yields 55,287 results, indicating that the field is highly active and growing rapidly.

Small microbiome studies are useful for generating hypotheses and identifying associations, however, large projects are necessary to provide a more comprehensive and integrated understanding of the microbiome and its role in human health and disease. Also, with the rapidly growing number of small microbiome studies, standardisation for data collection, storage, analysis, and reporting is necessary to compare and integrate results across studies. Hence, guidelines and tools have been developed to help researchers to design, conduct and share the microbiome data across different studies. Such examples are Human Microbiome Project (HMP) [38], Earth Microbiome Project [39], The Microbiome Quality Control project [40], International Human Microbiome Standards [41] and The American Gut Project [42].

1.2.2. Human Microbiome Project

Several large-scale projects have been created to characterise the microbial composition, its diversity and functional potential thanks to the low cost of sequencing and molecular assays. In 2010, the Metagenomes of the Human Intestinal Tract (MetaHIT) project described the gut microbiome extracted from stool samples collected from a European cohort irrespective of high sequencing cost at that time [43]. As a follow-up study, MetaHIT continued to publish new metagenomes through different European subprojects [44–46]. Followed by the MetaHIT project, in

2012, HMP project was launched and reported the microbial composition from 242 healthy US state adults using 16S rRNA gene profiling and from 139 people using metagenomics. Samples from HMP represented eighteen different body sites including skin, GI, oral and urogenital tract [47]. The main goal of the HMP project was to understand the microbial diversity in humans and also to understand the factors which influence microbial composition. Overall, healthy microbiomes from nearly 2000 individuals have been reported so far.

1.2.3. Second phase of HMP

The second phase of HMP, the Integrative Human Microbiome Project [48] (HMP2 or iHMP) was developed to study mainly the interplay between the host and the microbiome in terms of metabolism and immunity over time. HMP2 encouraged disease-related sub-projects to explore different omics approaches which will help to accelerate future work to study the relationship between the host environment and the microbiome as well as to provide data, specimens and protocols for further research. There were three disease-related sub-projects such as IBD, stressors that affect individuals with prediabetes and pregnancy and preterm birth (PTB) to mainly focus on the role of microbiome interactions in human health and disease state [8,49,50]. These studies elucidated the host response towards microbial activity and also covers intra-relationships. So far, more than 50 articles were published highlighting the outcome of these studies and can be found at https://www.nature.com/collections/fiabfcjbfj and data sources can be found at http://www.ihmpdcc.org.

1.2.4. Useful resources from the HMP

Thanks to the next-generation sequencing technologies, the amount of microbial sequencing projects have hit the multitude. Hence**,** an appropriate and easily accessible storage facility was needed to organise these data. The Sequence Read Archive (SRA) [51], the Database of Genotypes and Phenotypes (dbGaP) [52], the Metabolomics Workbench (https://www.metabolomicsworkbench.org/), and other public and/or controlled-access repositories are among the places where multi-omic data produced by the HMP1 and HMP2 phases have been archived [53]. Once institutional review boards (IRBs) approve a research project, some of the associated metadata may be shared public. However, certain types of data, such as protected metadata and human genome sequences, are subject to controlled access restrictions via dbGaP. This means that researchers must meet specific eligibility requirements and agree to terms and conditions before being granted access. These measures are in place to ensure that the data are used ethically, and to safeguard the privacy and confidentiality of the research participants. All data on the DCC is freely usable (projects PRJNA430482, PRJNA398089, PRJNA326441, PRJNA430481, phs001626, phs001719, phs001523, phs000256, and others). All phases of the HMP created formal data models and related entity connection schemas, which are all freely accessible at https://github.com/ihmpdcc/osdf-schemas. Users can find, query, search, view, and download data from tens of thousands of samples with related metadata on the

DCC website. A user can add a set of conditions, phenotypes, files or subjects to a shopping cart for later use after deciding they are of interest.

1.3. The 16S rRNA gene as a taxonomical marker

A widely used and straightforward approach for studying the taxonomic makeup of microbiota involves amplifying and sequencing a particular marker gene and inferring the composition of the microbiome based on the similarities of the resulting sequences. The 16S rRNA gene is the most established genetic marker choice for microbial analyses [38,54,55]. It is present in all bacteria and is also rarely affected by horizontal gene transfer [56]. The 1542 bp of 16S rRNA gene consists of both highly conserved regions and nine hypervariable regions (V1-V9) **(Fig. 2)**. The conserved regions are used as universal primer binding sites for amplification of the whole gene whereas nine hypervariable regions contain species-specific sequences which can be used to differentiate bacteria and archaea [57].

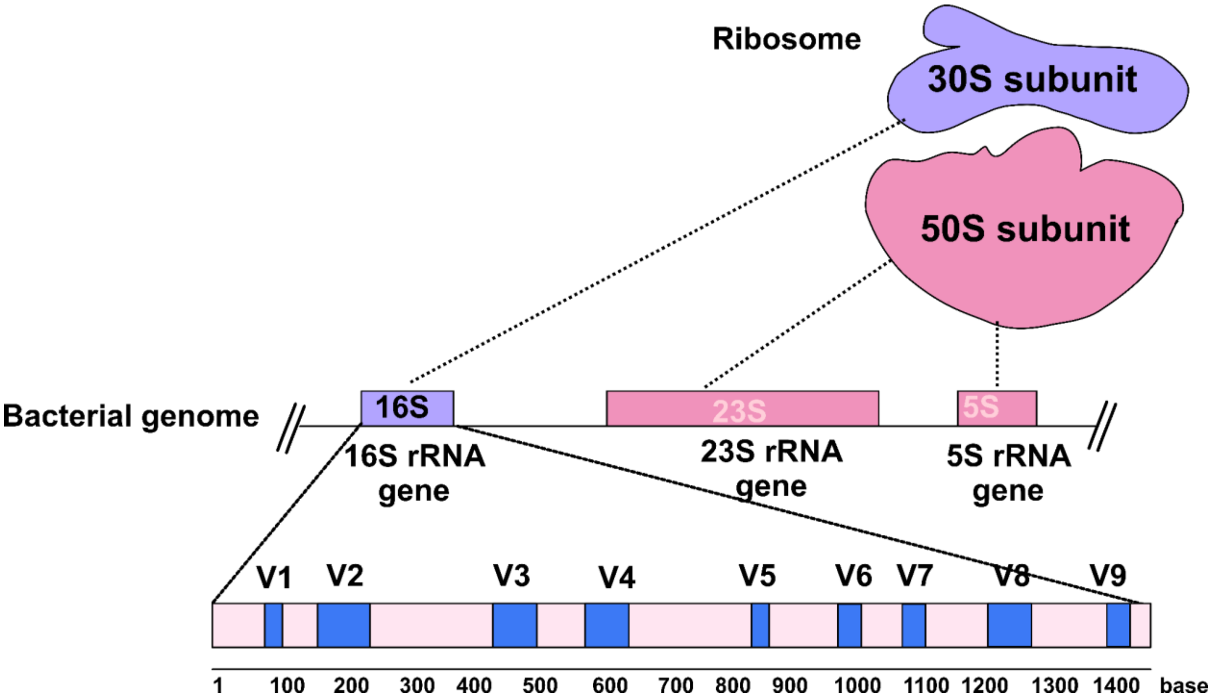


Figure 2: Structural diagram of 16S rRNA.
The 16S rRNA gene molecule is a component of the ribosome. The ribosome is composed of two subunits, the large subunit (50S) and the small subunit (30S) where the 16S rRNA gene molecule is located in the small subunit. The 16S rRNA gene comprises both conserved and variable regions. The conserved regions (represented by pink in 16S rRNA gene structure) of the 16S rRNA molecule are highly conserved across a wide range of bacterial species. 16S rRNA contains nine hypervariable regions (V1-V9) which are highly variable in terms of sequence length, composition, and mutation rate, which makes them useful for bacterial taxonomy and identification. (Source: own work)

n the early days, sequencing entire 16S rRNA gene s was carried out using conventional Sanger sequencing which was laborious and expensive [58]. However, in recent years, researchers started to follow short read sequencing using the Illumina MiSeq 2 × 300 bp platform as they are low cost with less effort. Hence, single variable regions such as V4 or V6 or two combined variable regions such as V1V2, V3V4, and V5V6 regions on the 16S rRNA gene were sequenced to study the microbiome [59–62].

Targeting different sub-regions results in different taxonomic compositions. This may be caused by different biases such as primer bias, the sequence length of the differential hypervariable region, and the uniqueness of hypervariable sequences across different bacteria (Table 1). Analysis using sub-regions may also be limited to the genus level based on the commonly used microbial taxonomy database [63]. Selecting one particular sub-hypervariable region may show bias in the identification of certain phyla. For example, according to Chakravorty et al. [64], V1 showed the best ability in differentiating *Staphylococcus* species whereas V2 and V3 were best in distinguishing all bacterial species to the genus level except for the *Enterobacteriaceae* family. In terms of coming from multiple hypervariable regions, the V1–V2 region was unable to classify phyla belonging to Proteobacteria and V3-V5 towards Actinobacteria. Similarly, V6–V9 was shown best suited to classify the *Clostridium* and *Staphylococcus* genus [57].

An alternative solution would be to sequence a full-length 16S rRNA gene . This can be achieved by utilising PacBio and Oxford Nanopore sequencing platforms capable of producing long-read sequencing [63]. However, long-read sequencing platforms generate reads with lower accuracy than the Illumina platform due to random base-calling errors during repeating sequencing of the same region [65]. Another potential alternative would be to sequence multiple regions of the 16S rRNA gene separately, and then to integrate information from different hypervariable regions as much as possible using bioinformatics approaches [66].

In late 2020, Loop Genomics developed a new approach called synthetic long-read sequencing technology (sFL16S), which transforms short-read sequences into long-read single molecules using an existing Illumina short-read sequencer combined with a unique molecule barcoding technology. It can be adapted to combining various hypervariable regions of the 16S rRNA gene and can be used in further downstream analysis. Additionally, reconstructed synthetic long-read sequences are reported to have high base accuracy [67].

Despite being by far the most often utilised gene for research on microbial community, the 16S rRNA gene has some drawbacks. A significant criticism of the method is that numerous bacteria have multiple copies of the rrn operon, which includes the 16S rRNA gene [68]. Using this approach could result in an overrepresentation of bacteria with multiple copies of the 16S rRNA gene, which poses a problem for abundance studies relying on 16S sequences. The many copies of the 16S rRNA gene present in the genome of some species of bacteria also exhibit high levels of sequence divergence, particularly in extremophiles. These overestimate diversity because their unique rRNA genes provide the impression that they are more than one bacterium.

Several public databases comprising sequences of 16S rRNA gene s are available such as RDP [69], SILVA and Greengenes (GG) [70], and RiboGrove [71]. All these databases retrieve full-length sequences from The International Nucleotide Sequence Database Collaboration (INSDC) which comprises DDBJ, EMBL-EBI and NCBI. These projects also include sequences from PCR amplification to include both culturable and unculturable species. This makes databases prone to include artefacts and also these databases may have incomplete sequences. The selection of a proper database is a crucial step while characterising the microbiome. Proper databases will help in reducing the uncertainty about variations within genes and also help to detect chimeric or artefact sequences.

Table 1: Advantages and disadvantages of using the 16S rRNA gene as a marker gene in microbial analysis

Advantages	**Disadvantages**
All bacteria and archaea contain it.	Present in numerous copies throughout the majority of species which could cause some organisms' abundance to be overestimated.
It has highly conserved regions suitable to design universal primers.	Sometimes, a few organisms do not have a highly conserved region which may result in primer bias.
It consists of high-variability zones that make distinct identifiers	Sometimes variable regions are not able to distinguish species-level resolution.
Several well-maintained reference databases are available for taxonomic identification	Many databases may contain errors and are not regularly updated.
There are well-researched primer pairs available to amplify the majority of organisms with great specificity for bacteria.	May not be specific for some bacterial groups, which could lead to incorrect community composition estimates.

1.3.1. Other marker genes

To overcome the limitations of 16S rRNA gene s as discussed in the previous section, alternative essential housekeeping marker genes, such as amoA, nirS, rpoB, nirK, pufM, pmoA and nosZ and, have been used in studies to decide taxonomic relationships for specific lineages of interest

[72–74]. These genes provide a way to analyse and compare the genetic makeup of different organisms, allowing for more accurate identification and classification.

While, the 16S rRNA gene is the most stable marker gene to identify bacterial species, 18S rRNA has also been a stable marker gene for characterising fungus populations The small subunit rRNA of eukaryotic ribosomes is encoded by the DNA sequence known as 18S rRNA. The 18S rRNA sequence has both conservative and variable sections, much like the 16S rRNA gene (V1-V9, absence of V6) **(Fig. 4(A))**. For 18S rRNA gene analysis, V4 and V9 have been recognized as the most popular and best option among variable regions, since they have the most comprehensive database information and the greatest classification effect [75–78] . The species distinctions among eukaryotic organisms in specific samples are reflected by 18S rRNA sequencing.

The fungal rRNA gene's non-transcriptional region includes the ITS (Internal Transcribed Spacer). ITS1 and ITS2 have often used ITS sequences for identifying fungi. Because ITS may accept more changes during the evolutionary process as a result of less natural selection pressure and exhibits exceptionally wide sequence variation in most eukaryotes, 5.8S, 18S, and 28S rRNA genes are well conserved in fungi. The changes between species (or even strains) are also apparent, despite the conservative type of ITS being very stable within species. Fragments of the ITS sequence are short (350 and 400 bases, respectively) and simple to interpret. They are frequently employed in phylogenetic analyses of various fungi

1.4. Workflow of 16S/18S/ITS amplicon sequencing

Together with technological advancements in DNA sequencing that make it easier to conduct investigations without using culture or cloning, microbiome studies has significantly expanded. The Roche 454 pyrosequencer, the first next-generation sequencer, could only sequence about 120 bases of the bacterial genome in a single run when it was originally introduced in 2005 [79]. Followed by pyrosequencing, Illumina Solexa was introduced in 2006 as a high-throughput DNA sequencing technology that revolutionized the field of genomics. It quickly became one of the most widely used sequencing platforms due to its high accuracy, throughput, and relatively low cost compared to previous technologies. In recent years, Illumina sequencing has been used extensively in microbiome research by covering numerous hypervariable areas of the 16S rRNA gene and span distances of up to 1000 bp [80]. The most used method for microbial community analysis of the human intestine is 16S rRNA gene sequencing because of its high resolution and economical method.

Illumina or PacBio sequencing is used in 16S/18S/ITS amplicon sequencing to read the PCR products that are produced using suitable universal primers of one or more 16S/18S/ITS regions. Taxonomic classification with their corresponding abundance within and between the community, phylogenetic evolution can be obtained by identifying the sequence variation and abundance of the target area. The entire workflow of 16S rRNA gene sequencing is represented in **Fig. 3**

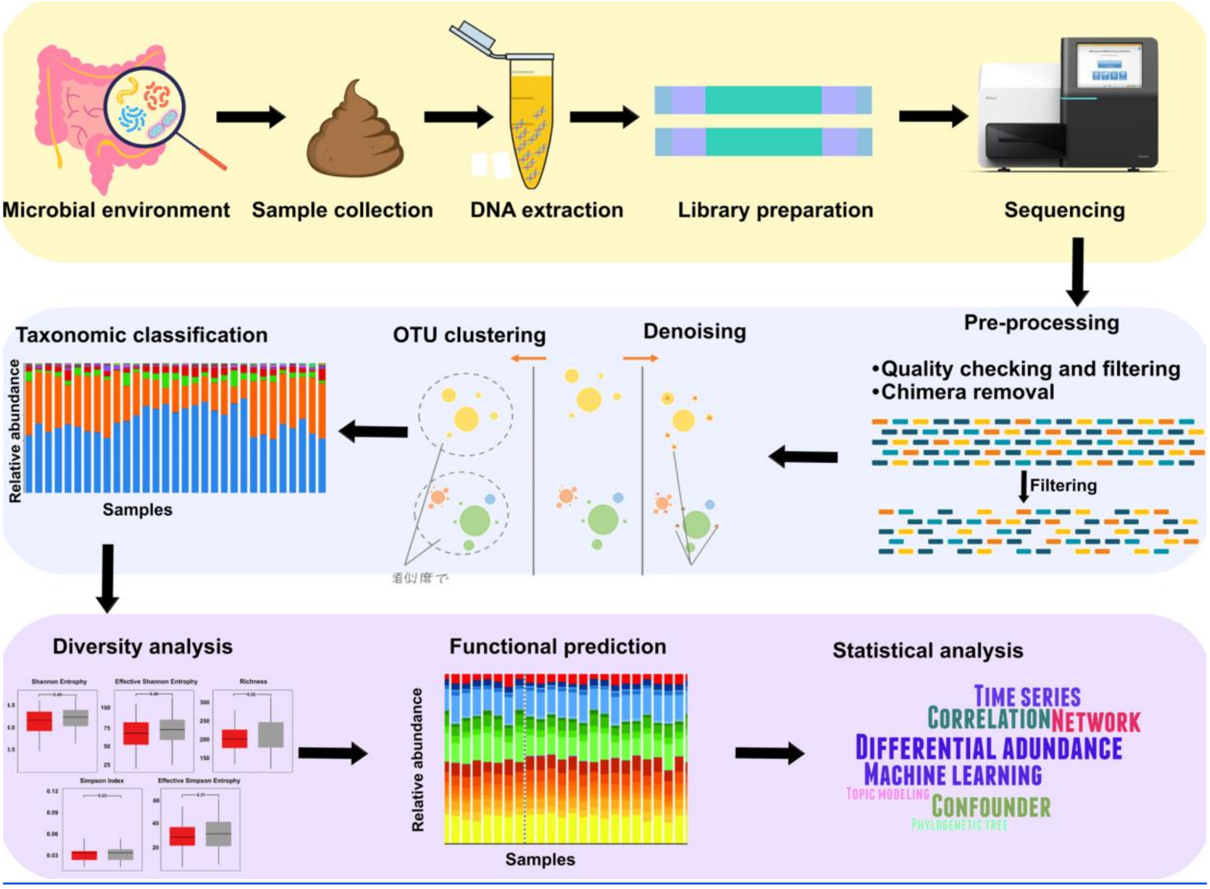


Figure 3: Flowchart of the major steps involved in 16S rRNA gene analysis.
The 16S rRNA gene analysis is a commonly used method in microbial ecology and taxonomy to study the diversity and composition of microbial communities. The major steps involved in a typical 16S rRNA gene analysis are as follows: Sample collection, DNA extraction PCR amplification, library preparation, high-throughput sequencing, Sequence quality control, Operational Taxonomic Unit (OTU) clustering/ denoising, taxonomic assignment and data analysis and interpretation and visualisation. (Source: own work)

1.4.1. Study design

In various microbiome-based investigations, erroneous and obscure trends are frequently seen. A strong study design helps to limit these trends. To prevent uncertainty in biological signals, trials, and failures, any hypothesis should generally be supported by rigorous literature-driven research and preliminary testing, employing pilot investigations. Eliminating confounding factors and improving data processing will both benefit from a rationalised study design [81].

Sample size: Choosing a size that is statistically significant is still an important step, particularly when the final results are employed in clinical contexts and interpretations. The microbial burden differs amongst biological replicates that exist in identical environments [82]. It is difficult to detect weak biological signals because of this variation between similar samples, especially when the effect size is unknown or small. Results from studies with limited sample sizes typically do not accurately reflect findings from population-based studies. It is crucial that sample sizes remain constant and are not changed throughout the course of the investigation [83].

Control: To determine whether a signal is genuine and not merely a stochastic or false result, controls are required. Two or more situations make up a suitably controlled experiment, with one producing observations free of distraction and the others remaining focused on changes [84,85]. Sadly, obtaining appropriate controls is still frequently a challenge, particularly in clinical trials when the microbial composition is influenced by age, gender, ethnicity, food, genetics, and numerous other lifestyle factors. Additional elements, such as animal strains, facilities, housing circumstances, handling, and breeding, may potentially have an impact on the microbial profile in animal research [86].

Metadata: Metadata is a set of information that includes important details about every sample used in an experiment. One of the most important tasks before any downstream analysis can be done is metadata generation. In addition to acting as a sample reference sheet, it aids in avoiding erroneous interpretation of results and emphasises the relative importance of each aspect [87]. Several contemporary statistical comparison methods need the usage of metadata.

1.4.2. Pre-processing of raw amplicon reads

To extract taxonomic information from raw sequences, three key analytical post-processing procedures are essential: (i) concatenating read pairs to generate longer single reads, (ii) examining quality and trimming reads, and (iii) assigning taxonomic classifications. Each step can involve several different tools or methods, and each one might call for programming knowledge and/or extensive computational resources [88]. Every sequencing technique is prone to errors. Amplicon sequencing methodology is especially susceptible to misleading results brought about by improperly sequenced reads as errors in the raw sequence reads might result in inaccurately high estimates of bacterial diversity [89]. Some, but not all of these effects are countered by clustering reads based on sequence similarity [89]. Low-quality reads are typically filtered away during pre-processing, which lowers the amount of erroneous reads [90].

1.4.3. Sequencing and analysis

Once the raw reads are pre-processed, the next step is to identify bacterial composition from 16S rRNA gene amplicon sequences with specialised technological and bioinformatic knowledge. Beginners may find it challenging to process this data due to the wide variety of tools available at each stage of the analytical process and the scripting knowledge they had to be familiar with.

Recently, a number of workflows have been created to get around these restrictions by streamlining the analytical process and enabling unskilled individuals to get familiar with sophisticated programming or computational methods **(Table 2)**. Widely used bioinformatic pipelines for analysing 16S rRNA gene sequencing data are *QIIME2* [91,92], *Bioconductor* [93], *USEARCH* [94], *UPARSE* and *Mothur* [95]. These tools are based on the operational taxonomic unit (OTU) clustering methods. The reads are clustered into OTUs using a similarity cutoff of 97%. However, clustering sequences based on 97% failed to distinguish between actual clustering and sequencing errors. To overcome these, Amplicon sequence variance (ASV)-based denoising methods were introduced. Detailed explanations of clustering and denoising approaches are explained under the Clustering/denoising section in chapter 2. Researchers can infer ASVs either using *DADA2* [96] through the Bioconductor package repository or using several *QIIME2* plugins, such as *DADA2* [96] and *Deblur* [97]. Sequences for ASVs are differentiated at the single nucleotide changes across gene sequences, but sequences for OTUs are binned together if they generally by less than 3% of the total sequences[98].

Table 2: Software tools and packages used for 16S rRNA gene analysis

Purpose	**Description**	**Tools**
Quality check	Software to check the quality of the reads and to trim the low-quality reads	*FASTQC* [99], *Fastx-Toolkit* [100], *Trimmomatic, PRINTSEQ, NGS QC Toolkit, multQC*
16S rRNA gene amplicon pipelines	Tools for analyzing 16S marker gene data	*QIIME2*[91], *Mothur*[95], *SILVA*[101], *DADA2* [96], *MICCA210, FunFrame*[102], *LOTUS2* [103], *IMNGS2* [104], *USEARCH*[105], *UPARSE*[106], *BIOCOM-PIPE* [107], *GenePiper* [108], *Animalcules* [109]
16S rRNA gene amplicon classifiers	Tools for characterizing OTUs from 16S rRNA gene to different taxonomic levels such as phyla, genus and sometimes species	*RDP/SILVAClassifier* [110], *Mothur* [95], *UTAX* [111], *16S Classifier* [112]

16S rRNA gene statistical analysis tools	Tools and software packages for the analysis and statistical comparisons of 16S marker gene datasets	*Mothur*[95], *QIIME2* [91], *Phyloseq* [113], *LEfSe*[114], *STAMP* [115], *MicrobiomeAnalyst*[116], *Rhea* [117], *IMNGS* [104]
OTU picking methods	Methods for obtaining a set of OTUs through read alignment to a database, read clustering against one another, or both	*Closed Reference, Open Reference, De Novo, Mothur*[95], *uclust*[94], *UPARSE83, CD-HIT*[118], *DADA2* [96], *TIC* [119]
ASV generation methods	Methods for obtaining a set of ASVs through denoising approach	*DADA2*[96], *Deblur*[97], *UNOISE3*[120]
Chimaera removal	Tools for removing chimeric sequences	*UCHIME*[121], *DADA2*[96], *QIIME2*[91]
16S rRNA gene databases	Databases with microbial sequences which is used for taxonomy assignment for OTUs	*SILVA*[101], *GG*[70], *RDP*[69], *rrnDB*[122], *16S-ITGDB* [123]

1.5. Challenges in 16S rRNA gene amplicon data analysis

The main objective of the microbial analysis is to gain knowledge about microbial populations and their respective abundance in an environment of interest. Characterising the microbial population is challenging due to many reasons. The main technical challenges include short read length, high throughput sequencing, primer selection and absence of proper reference databases. Also**,** while addressing the genomic diversity in a particular environment, one should be aware of biological challenges such as horizontal gene transfer and different evolution rates between genomes [124].

1.5.1. Compositionality

In general, compositional data can be naturally expressed as probabilities or proportions, or that have a fixed sum. Only information regarding the relationship between the species can be obtained from compositional data [125–128]. Microbial data is often compositional, meaning that it represents the relative abundance of different taxa or species within a sample rather than their absolute numbers. This is a consequence of the limited library size of sequencing-based approaches such as 16S rRNA gene sequencing or shotgun metagenomics.

Since sequencing machines can only sequence reads up to their capacity, the concept of absolute abundance cannot be applied to high-throughput sequencing (HTS) research. These methods provide relative abundance data rather than absolute counts, and this data needs to be normalised to take into account the total sequencing depth and the different sequencing biases. As shown in **Fig. 4B**, analysis of data on a microbiome gathered by HTS frequently assumes that sequencing is, in some manner, calculating the relative abundance of each bacteria with respect to other bacteria in the population. When microbiome datasets are transformed into relative abundance values, normalised counts, or rarefied, this is tacitly accepted [129]. There are different methods for the normalisation of compositional data, such as the use of relative abundance or proportions, or the use of transformed data such as log or square-root transformed data [129,130,131].

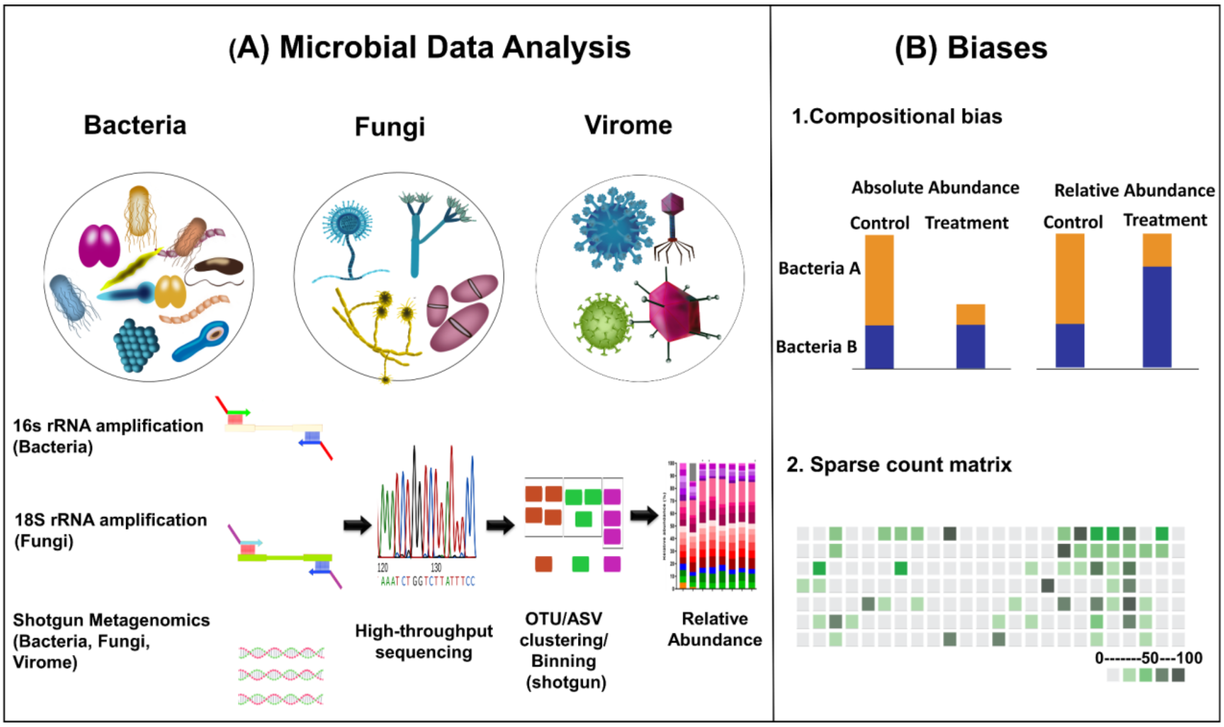

(A) Microbial Data Analysis
Bacteria
Fungi
Virome
16s rRNA amplification (Bacteria)
18S rRNA amplification (Fungi)
Shotgun Metagenomics (Bacteria, Fungi, Virome)
High-throughput sequencing
OTU/ASV clustering/ Binning (shotgun)
Relative Abundance
(B) Biases
1.Compositional bias
Absolute Abundance
Relative Abundance
Control
Treatment
Control
Treatment
Bacteria A
Bacteria B
2. Sparse count matrix
0-------50---100

Figure 4:Overview of microbial data analysis pipeline and biases.
(A) presents a diagrammatic representation of the taxonomic profiling of bacteria, fungi, and the virome, while (B) highlights three significant biases - compositionality, sparsity, and spurious correlations - in the analysis of microbial co-occurrence networks. This figure was originally published as Network analysis methods for studying microbial communities: A mini review[132] in Computational and Structural Biotechnology Journal as an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/) and modified as follows: In the right panel of the figure (Biases) spurious indirect edges part was removed and used for this thesis.

Researchers have developed/adopted methods from other domains, such as geology and chemistry, where compositional data is also common. For dealing with compositional analysis, log-ratio transformations are commonly used, because they allow for the comparison of relative abundances between different features in dataset [127]. The use of log-ratios can detect the relationships between the characteristics in the dataset and make them symmetric and linearly connected. This enables the determination of feature abundances relative to each other. This information is highly relevant to the environment and can provide insights into the microbial community structure, diversity, and interactions. However, it is important to note that log-ratio transformations do not provide information on absolute abundances. This is because sequencing methods only provide relative abundance data and not absolute counts. Therefore, log-ratio transformations are useful for the comparative analysis of the dataset but not for estimating the absolute abundance of features in the environment. In addition to adopting methods from other domains, new tools have been developed specifically for the analysis of compositional datasets. For example, mixOmics [133] is a R package for the statistical analysis of multi-omics datasets that incorporates methods for the analysis of compositional data. This package has been used to analyze microbiome datasets and identify associations between microbiome composition and host phenotypes [134].

One such approach is centered log-ratio (CLR) transformation developed by Aitchison (1986) [125], which transforms the compositional data into a new coordinate system where the spurious correlations are eliminated. To obtain the CLR transformation of a sample, the following steps can be taken using an observation vector of D "counted" features such as genes, OTUs, or taxa, denoted by x = [x1, x2,..., xd] :

$$xclr = [log(x1/G(x)), log(x2/G(x))..log(xd/G(x))]$$

$$G(x) = \sqrt[D]{x1.x2...xd}$$

G(x) stands for x's geometric mean. According to Van den Boogaart and Tolosana-Delgado (2013) [134], the values after the CLR transformation can be utilised as inputs for model construction as well as multivariate hypothesis testing utilising programmes like MANOVA, regression, etc. The CLR-transformed data are scale-invariant, meaning that the ratio should be the same whether the sample has a few reads or many reads; the only thing that changes is the precision of the CLR estimate. The second method is an additive log-ratio transformation (ALR), which Aitchison first proposed, uses one component as the denominator, or reference, and all the other components as numerators [135,136]. There are J 1 log-ratios in the ALR set concerning the chosen reference component, designated by ref, of the following form if there are J components, each having a value of X1, X2,.., XJ [135].

Without removing, swapping out, or replacing the zero count values with pseudo values, it is impossible to calculate the G(x) given sparse data. There are few good alternatives for handling zero count values such as employing the *zCompositionsR* package and [137] the *aldex2*

[131] tool accessible in Bioconductor. The aldex2 tool conducts statistical tests on CLR values from a modelled probability distribution in a dataset, yielding both parametric and non-parametric test outcomes and an estimate of effect size. It is a highly effective technique for decreasing false-positive identification problems in both real and modelled microbiome datasets, while maintaining sensitivity and remaining relatively insensitive to changes in data subsets [138].

While log ratio transformation, specifically the CLR transformation, is a commonly used method for analyzing compositional data, including microbial abundance data, there may be several reasons why some researchers do not apply it. One reason is that the interpretation of transformed data may be challenging. The transformed data may not have direct biological meaning, and the interpretation of statistical results may require additional effort and knowledge of the methods used. Another reason is that it assumes that the abundance of each microbe is independent of the abundance of other microbes in the sample. However, this assumption may not always hold true, and other methods that account for interactions between microbes, such as multivariate analyses, may be more appropriate. Finally, some researchers may not apply log transform simply because it is not the standard or default method in their field or research area. The choice of normalization method often depends on the research question, data structure, and personal preference.

while log ratio transformation is a useful method for analyzing microbial abundance data, there may be other normalization methods such as total sum scaling [139], cumulative-sum scaling [140], that better suit the research question or data structure, or some researchers may prefer to use other methods due to interpretation challenges or other assumptions made by the log ratio transformation.

1.5.2. The challenge of sparsity

The microbiome data can be characterized by sparsity indicating that many taxa are rare, and most of the times, zeros predominate among all other values **(Fig. 4B).** The unique microbial composition of each sample typically results in only a small number of bacterial taxa being commonly observed in the majority of samples, with the remainder being relatively infrequent and found in only a few samples. Consequently, such data are prone to zero inflation [141]. In addition, the sparsity is an indication of the uncertainty of the count of uncommon OTUs/ASVs because these OTUs/ASVs are below the sequencing detection limit, and there are few sequences in each sample. To summarize, some microorganisms and taxa are structurally absent from the majority of samples, while others may be undetectable due to insufficient sequence depth or limitations of the applied techniques [142]. Both of these factors contribute to the inability to detect them. The high percentage of zeros can present challenges for certain statistical tools and modelling approaches that are now available, which may lead to estimations that are biased. According to Tsilimigras and Fodor's findings [143], the analysis of 16S rRNA gene sequencing data faces a significant obstacle in the form of sparsity.

To address sparsity in microbiome datasets, different methods have been developed such as normalization and transformation of the data [129], filtering and feature selection, applying pseudo counts [144], imputation of missing values [145], and modeling approaches to handle sparse data [146]. Normalization and transformation methods can be used to adjust for variation in sequencing depth and abundance, and to standardize the data across samples. Filtering and feature selection methods can be used to remove low-abundance or irrelevant features from the dataset, which can improve the quality of downstream analyses. Imputation methods such as mbImpute [147] can be used to estimate missing values in the dataset, which can help to reduce sparsity and improve the accuracy of downstream analyses. However, it is important to use imputation methods with caution, as they can introduce biases and distortions in the data. Finally, modeling approaches such as sparse regression [148] and network-based methods [149–151] have been developed to handle sparse microbiome datasets. These methods are designed to identify associations between features and outcomes, while also accounting for the sparsity and high-dimensionality of the data.

1.5.3. Functional prediction from 16S rRNA gene sequences

The most important challenge of 16S rRNA gene sequencing analysis is that it can only provide an indirect estimate of microbial function. In other words, 16S rRNA gene sequencing helps in estimating taxonomic composition and their abundance but not the biological function of those taxa [152]. In order to understand what kind of activities microorganisms are engaging in, it is useful to know their functional capabilities.

Several 16S studies attempt to infer the functional contribution of particular community members by mapping 16S gene sequences to their nearest sequenced reference genome [153,154]. Although the precision of such methods is unknown, the relationship between gene content and phylogeny [155,156] raises the possibility that it may be possible to roughly anticipate the functional capacity of microbial communities from phylogeny.

Tools like *PICRUSt2* [157], *Tax4Fun2* [*158*], *PanFP* [159], *Piphilin* [160], *COWPI* [161], and *MetGEM* [162] make an effort to predict the abundances of functional genes based on genomic data recorded in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [163] or based on genomic models [164] To address the absence of functional information in 16S rRNA gene profiles. *PICRUSt2* is the most commonly used prediction tool that uses a hidden state prediction method. Contrarily, *Tax4Fun2* employs sequences that fall within a specified cutoff for similarity to reference sequences [158]. *PanFP* [159] is based on a pan-genome reconstruction that has been functionally annotated and is then weighted with the microbial abundance seen in a particular sample. Each tool follows a different algorithmic approach and relies on different references, leading to a considerable discrepancy in their predictions.

Hence, it remains an open question if metagenome prediction tools are also suited for more subtle contrasts related to human health. Previous benchmark studies [165] could not detect any performance differences between the tested methods, suggesting that a more comprehensive benchmark is needed to recommend guidelines for tool selection and to establish the limits of metagenome prediction tools in human disease research. To study the limitation of 16S functional profiling, metagenomics serves as the most suitable ground truth. By comparing the results of the two approaches, researchers can better understand the strengths and weaknesses of each method and gain a more comprehensive view of microbial community structure and function. The third chapter of the thesis covers an extensive benchmark analysis using metagenome as the ground truth. Hence, I will briefly introduce the concept of metagenomics along with the standardized workflow and advantages over 16S rRNA gene sequencing.

1.6. Metagenomics

As 16S rRNA gene sequencing fails to provide genetic contribution of organisms of the given community in terms of functional genes, shotgun metagenomic sequencing allows for a more comprehensive characterization of the functional genes and metabolic capabilities of a microbial community. In shotgun metagenomics, the entire chromosomes, such as long DNA are randomly cut into small fragments and then sequenced using any one the available sequencing platform [166]. Hence, this approach provides knowledge of the taxonomic composition of the environment under study as well as knowledge of the functional genes in the sample which cannot be provided by 16S rRNA gene sequencing [167,168].

Metagenomics has made it possible to do extensive research on complex microbiomes in the 20 years since it was first used [169–172]. This approach has led to the finding of many crucial bacteria such as environment bacterial phyla with endosymbiotic behaviour [173] and species that can completely nitrify ammonia [174] among other discoveries. Other noteworthy discoveries include tracking of human outbreak pathogens, the strong correlation between the viral [175] and bacterial fractions [176] of the microbiome and inflammatory bowel diseases, and the prevalence of antibiotic genes in commensal gut bacteria [177], which helps in studying the microbial alterations COVID patients [178,179].

After the study design is established, a typical shotgun metagenomics study involves five main steps: (i) collecting, processing, and sequencing the samples; (ii) preprocessing the reads; (iii) analyzing the sequences to profile the taxonomic, functional, and genomic characteristics of the microbiome; (iv) performing statistical and biological post-processing analyses; and (v) validating the results. Each of these stages can be carried out through various experimental and computational methods, presenting researchers with a range of options. Despite its seemingly straightforward nature, shotgun metagenomics has limitations due to potential experimental biases and the complexity of computer analysis and interpretation **(Fig. 5).**

1.6.1. Metagenome assembly

There are many reported methods for metagenomic assembly from sequence data. Selecting the best method is a difficult undertaking that is heavily influenced by the study's objectives. Whole-genome assembly and metagenome de novo assembly share conceptual similarities [180–182]. The de Bruijn graph approach is a well-known algorithm employed in metagenome assembly step [183]. In order to create a de Bruijn graph for single draft genome assemblies, each sequencing read is divided into overlapping subsequences of length k. The edges and vertices of the de Bruijn graph are defined by this collection of overlapping k-mers. Finding a path across the graph that reconstructs the genome is the assembler's duty (s). However, sequencing errors and repetitive sequences can make this task challenging and can lead to misassemblies or assembly fragmentation.

These difficulties are addressed by assemblers that are particular to the genome. *Meta-IDBA* [184] employs a multiple k-mer strategy to overcome the difficulty of selecting an appropriate k-mer length that performs well for both high- and low-abundance species. While *Meta-IDBA* and *MetaVelvet* both provide modifications to partition the de Bruijn graph, the most recent version, *IDBA-UD* [185], optimises the reconstruction for asymmetric sequence-depth distributions. *MetaSPAdes* [186] is a modified version of the *SPAdes* [187] assembler that is designed to handle large metagenomics data sets. It employs several heuristics for graph simplification, filtering, and storage to enhance the assembly process.

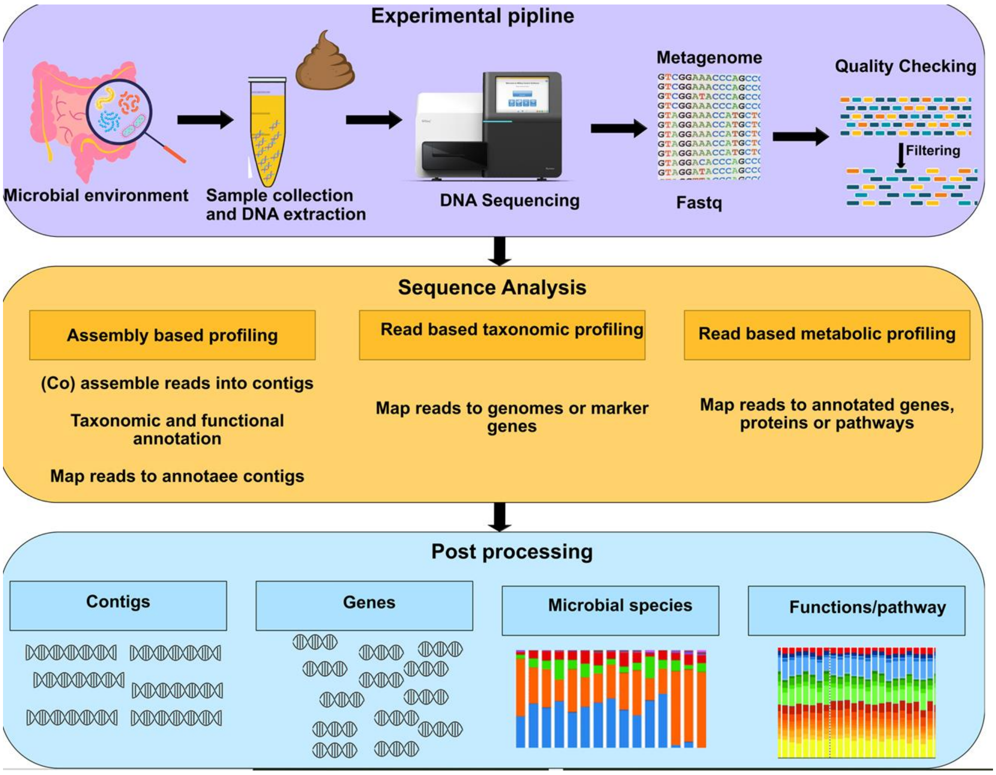

Experimental pipline
Microbial environment
Sample collection and DNA extraction
DNA Sequencing
Metagenome
Fastq
Quality Checking
Filtering
Sequence Analysis
Assembly based profiling
Read based taxonomic profiling
Read based metabolic profiling
(Co) assemble reads into contigs
Taxonomic and functional annotation
Map reads to annotaee contigs
Map reads to genomes or marker genes
Map reads to annotated genes, proteins or pathways
Post processing
Contigs
Genes
Microbial species
Functions/pathway

Figure 5: Summary of a metagenomics workflow. It provides a comprehensive overview of the diversity and functional potential of microbial communities. Similar to 16S rRNA gene sequencing, It also contains sample collection and DNA extraction steps. However, instead of amplifying only marker genes, entire genomic DNA is sequenced and the obtained sequencing reads are preprocessed to remove low-quality reads. Assembly-based profiling, read-based taxonomic profiling, and read-based metabolic profiling are three different approaches to analyse the composition and functionality of microbial communities in a

metagenome sample. Read-based taxonomic profiling, also known as taxonomic classification, uses machine learning algorithms to classify metagenomic reads directly to the taxonomic level of the microorganisms they come from. Read-based metabolic profiling, also known as functional classification, uses machine learning algorithms to predict the metabolic pathways and functions of the microorganisms present in a metagenome sample. (Source: own work)

1.6.2. Assembly-free metagenomic profiling

Metagenomes' taxonomic profiling reveals which microbial species are present and provides an estimation of how abundant they are. Through external sequencing data sources, such as publicly available reference genomes, this can be done without assembly. This method can reduce assembly issues, accelerate computing, and allow for the profiling of low-abundance organisms that cannot be created from scratch. Its key drawback is that it is challenging to profile previously uncharacterized microorganisms. With the advancement of sequencing technologies, more and more genomes are being sequenced each year, including those from previously unculturable species [188]. This is made possible by new techniques such as single-cell sequencing, metagenomic assembly, and new cultivation methods. As a result, the number of reference genomes is rapidly increasing [189]. Based on the sample type, such as the human gut [46], the diversity of reference genomes is now widespread enough to enable assembly-free taxonomic profiling, even for very low-abundance microorganisms that lack adequate sequencing coverage and depth for genome assembly. The absence of representative reference genomes makes it difficult to analyse increasingly varied habitats, such as soil and oceans. As a result, when examining metagenomes from these contexts, it is typically wise to employ *de nov*o based assembly.

Human-associated metagenomics studies [47,190–192] have utilized assembly-free taxonomic profilers that rely on reference genomes [193] and environment-specific assemblies to achieve species-level resolution[192]. Despite the potential for false positives in read mapping to genomes, these methods have been successful with post-processing techniques such as lowest common ancestor strategies [194] or compositional interpolated Markov models. [195]. However, their computational run times are not faster than assembly-based methods. *Kraken* [196] is a taxonomic profiler that employs the lowest common ancestor approach but utilizes k-mer matching instead of sequence mapping to enhance speed of computation.

Taxonomic profiling is a fast and precise method that does not require assembly and has been used in various applications. This method involves selecting representative or distinctive genes (markers) from the available reference sequences. *mOTU* [197] concentrates on universally conserved yet phylogenetically informative markers, whereas *MetaPhlAn* [198] utilizes several thousand clade-specific markers with high discriminatory power and was successful to quantitatively profiling the microbiome from multiple body areas for the HMP..

1.6.3. Genes and metabolic pathways from metagenomes

Utilising specialised single-genome characterisation technologies, it is possible to determine a microbial community's genetic composition from a fragmented but high-quality metagenome assembly. These start with a gene identification stage, typically with a setting for a metagenomic-specific parameter [199], and are followed by pipelines that employ homology to annotate data, which are frequently used to characterise pure isolated genome assemblies. The microbial gene catalogue of the human [46] gut was assembled using metagenomic assemblies, even though this method is frequently constrained by the significant number of uncharacterized genes in reference database catalogues.

By translating sequence searches against protein families [200], metagenomic data sets [201] were understood. Databases like KEGG [202] or UniProt [203], which combine manually annotated and computationally predicted protein families, can be employed to characterise the microbiome's functional potential. In the *HUMAnN* pipeline [200], aggregates individual protein families into higher-level metabolic pathways and functional modules, which are then visualized through graphical reports using *MEGAN4* [194] or as comprehensive tables of metabolic presence, absence, and abundance. However, a significant hurdle in characterizing the metabolic potential of a community, regardless of whether assembly-free or assembly-based approaches are used, is the lack of annotations for accessory genes in most microbial species.

A complementary approach to metagenome metabolic function profiling is to perform a detailed characterization of specific activities of interest. The spread of antibiotic resistance, for instance, can be determined by identifying the genes responsible for resistance to antibiotics in a microbial community [204]. This strategy has relied heavily on ad hoc methods and manually curated databases of antibiotic-resistance genes (ARDB) [205] was the first extensively used resistance database and is now supplemented by new resources, like Resfams [206]. Significant resources are often put into describing a metagenome's repositories, both through cultivation-based isolation investigations and through focused metagenomic analyses for specific gene families of interest. These methods can complement each other, as cultivation-based isolation can provide information about the growth conditions and behaviour of individual species, while metagenomic analyses can provide a more comprehensive view of the diversity and abundance of species in a sample. Additionally, metagenomic analysis can validate results from cultivation-based studies by identifying the presence of specific gene families in a sample.

1.6.4. Post-processing analysis

No matter what techniques are employed for initial metagenomic sequencing studies, the results will include data matrices of samples as opposed to microbial characteristics (i.e., species, taxa, genes and pathways). Statistical tools are used in post-processing analysis to analyse these matrices and determine how the results relate to the sample metadata. Many well-known R packages, like *DESeq2* [207], *Vegan* [208], and *MetagenomeSeq* [209], which were initially created for amplicon sequencing or other purposes like transcriptome analysis, can be utilised for metagenomic analysis.

1.6.5. Advantages and challenges in metagenomics

Metagenomics offers several advantages, including (i) the ability to screen for target genes or active products without the need for culture conditions, (ii) the discovery of novel medications from marine and extreme environmental microbial resources, (iii) the identification of new members of existing enzyme families or enzymes that exhibit activity in specific physicochemical contexts, which can be beneficial for industrial applications, and (iv) the investigation of carbon, nitrogen, and sulfur cycle metabolism in environmental microbes through metagenomics and metabolomics.

Similarly, metagenomics also has several challenges including (i) The complexity of the microbial community's living conditions can make it difficult to harvest genes from species with low abundance. (ii) The gene cloning process can result in the loss of DNA fragments, which can make it difficult to identify and study specific genes. (iii) Heterologous expression of foreign genes, which is the expression of a gene in an organism that is not its native host, occurs infrequently, making it difficult to study the function of specific genes. (iv) The currently available screening techniques for enzymes are limited and may not be able to satisfy all needs. (v) The low efficiency of screening techniques means that only a small number of positive clones can be selected from a large number of clones. (vi) Only a limited number of the enzymes that have been recently discovered may be suitable for industrial applications due to various restrictions, such as temperature, pH, and other related factors. To address these challenges, future research in functional metagenomics will need to focus on developing new screening techniques that are more efficient, sensitive, and easy to use, as well as improving methods for cloning and expressing foreign genes to study their function

There are several computational challenges in metagenomics, including (1) data size and complexity: Metagenomic data can be very large and complex, making it difficult to analyze and interpret; (2) assembly and binning: assembling and binning metagenomic data can be challenging due to the high diversity and complexity of the data; (3) taxonomic and functional annotation: assigning taxonomic and functional information to metagenomic sequences can be difficult due to the lack of reference genomes and the high degree of novelty in the data; (4) comparative metagenomics: comparing metagenomic data across different samples and environments can be challenging due to the high degree of variation and complexity in the data; (5) data integration: integrating metagenomic data with other types of data, such as transcriptomic, proteomic, and metabolomic data, can be challenging and requires advanced computational methods.

Main applications of metagenomics in microbiome research are identification and characterization of the microorganisms, investigation of functional capabilities of microbial communities and study of the relationships between microbial communities and their host organisms. In conclusion, metagenomics has revolutionized the field of microbiome research by enhancing our understanding of the human microbiome and its impact on human health. Its applications in microbiome research have the potential to advance our understanding of human health, environmental biology, and biotechnology.

1.7. Problem statement

The gut microbiota plays an important role in maintaining human health by facilitating several important functions such as host defence and immune response stimulation. Research has shown that the gut microbiome can have an impact on several diseases such as obesity, diabetes, inflammatory bowel disease, and even mental health disorders. Moreover, the gut microbiome can be affected by several factors such as diet, antibiotics, and other medications which can be used to modulate the gut microbiome in order to improve health. Exploring the gut microbiome can offer a more profound insight into the intricate interplay between microorganisms and their host, presenting opportunities for novel therapeutic approaches to tackle a range of diseases.

One of the most common approaches to study microbial population is shotgun metagenomic sequencing which allows for a more comprehensive characterization of the functional genes and metabolic capabilities of a microbial community than 16S rRNA gene sequencing. There are several computational challenges in metagenomics: Data size and complexity, assembly and binning due to the high diversity and complexity of the data, taxonomic and functional annotation due to the lack of reference genomes and comparative metagenomics due to the high degree of variation and complexity in the data.

Due to these challenges in metagenomics, 16S rRNA gene sequencing is still a popular choice for many studies. The main advantages of 16S rRNA gene sequencing are that it is cost-effective, has high resolution at the genus level, availability of large reference databases for 16S rRNA gene sequences, can also be used in comparative studies and high reproducibility. That being said, many studies are now starting to combine both metagenomic and 16S rRNA gene sequencing for a more comprehensive understanding of the microbial community.

The 16S rRNA gene is considered as a gold standard marker gene used for the identification and classification of bacteria and archaea. The conserved regions of the 16S rRNA gene are used to confirm the presence of bacterial or archaeal DNA, while the hypervariable regions are used to differentiate between different species and strains.

The analysis of 16S rRNA gene amplicon data can be challenging due to a number of factors. One of the main challenges is the short read length, which can make it difficult to accurately assign reads to specific taxa, especially at lower taxonomic levels. This is because the short read length doesn't capture the whole gene sequences, and it could lead to different sequences being assigned to the same species.

The second challenge is the high-throughput sequencing, which can produce a large amount of data, making it difficult to manage and analyse which in turn results in the need for computational resources and bioinformatics expertise to process and analyse the data. The third challenge is primer selection, as the choice of primer can affect the diversity of the resulting amplicon library and bias the results. Different primer sets can target different regions of the 16S rRNA gene and may result in different species or genera being amplified. Therefore, it is important to choose the primer set that is most appropriate for the research question and the sample being studied.

The fourth challenge is that 16S rRNA gene sequencing only provides an estimate of the microbial taxonomic composition, which can provide information about the types of microorganisms present in a sample. Nonetheless, it doesn't offer immediate insights into the functional potential of these microorganisms. Mapping 16S rRNA gene sequences to their nearest reference genome can provide some information about the functional capacity of microbial communities. This is because the presence of certain genes or pathways can be inferred based on the presence of homologous genes in the reference genome. However, it is important to note that this approach is only a rough estimate of functional capacity and can be affected by the quality and completeness of the reference genome. Another approach is to use functional gene markers such as genes involved in specific pathways or proteins that are known to perform specific functions. However, this approach also has limitations, as the presence of a gene or protein does not always indicate that it is actively being expressed or that it is functional in the microbe.

There is a great demand for new tools to address this growing need. Even though R packages and command line tools are powerful for microbial data analysis they can be difficult to use for those who don't have the necessary bioinformatics and scripting skills. There is a demand for easy-to-use tools that will assist novices with end-to–end microbial analysis. Many of the web-based tools do not cover downstream analysis, often omitting raw data processing, or only offer standard analysis that may not work for complicated data sets. They are often limited to functional profiling and use obsolete methods like *Tax4Fun* or *PICRSUt1*. These tools were outperformed *PICRUSt2,* which do not provide confounder analysis.

This thesis has studied these challenges and provided a reliable solution/recommendation for the users. **Fig. 1** gives an overall outline of the thesis explaining different computational challenges focused on each chapter.

1.8. Objectives

This work aims to address computational challenges in microbial data analysis and recommend suitable solutions. Three main objectives of the studies are as follows:

- Addressing issues in 16S rRNA gene sequencing in terms of microbiome characterization
- Studying the limits of functional profiling with 16S rRNA gene sequencing data
- Developing a user-friendly web interface for microbial data analysis

The primary aim of this thesis was to investigate the impact of various factors such as primer sequence selection, reference databases, clustering techniques, and pipeline parameters on the outcomes of microbial data analysis. In order to achive this, To achieve this objective, a significant benchmark dataset consisting of 16S rRNA gene amplicon sequences targeting diverse V-regions was employed. Various software tools with distinct parameter sets were employed in the benchmark pipeline. In addition to mock communities, the evaluation was conducted using complex human fecal samples for comparison.

The second objective elucidated the limitations of functional profiling derived from 16S rRNA gene amplicon sequences were addressed through benchmark analysis. Public datasets and simulation datasets were used to test the accuracy of the most widely used functional prediction tools such as *PICRUSt2*, and *Tax4Fun2* against metagenomics as the ground truth.

The third and final objective was to develop a user-friendly web interface called *Namco* with suitable recommendations obtained from the first two objectives. *Namco*, a multifunctional R shiny interface, acts as an all-in-one solution for microbiome analysis. It covers the entire spectrum of microbial analysis including upstream analyses such as raw data processing and taxonomic binning and downstream analyses such as basic statistics, machine learning and network analysis, among other features. Each objective is explained in detail in each chapter separately.

Chapter two: Primer, pipeline, parameters: Issues in 16S rRNA gene sequencing

2. Chapter two: Primer, pipeline, parameters: Issues in 16S rRNA gene sequencing

2.1. Declaration of contributions

This chapter is the result of a comparative analysis of different pipelines and parameters in 16S rRNA gene sequencing data under the guidance of Dr. Markus List, Head of Big Data in Biomedicine Group and Prof. Klaus Neuhaus, Head of Core Facility Microbiome and Dr. Sandra Reitmeier, Core Facility Microbiome, Technical University of Munich. Later, it was published under the open access journal *msphere* in February 2021 [55]. The experiment design and performing of experiments were done by Dr. Isabel Abellan-Schneyder, who was a former PhD student at the Technical University of Munich. The bioinformatics/computational analysis described here were performed by Monica Matchado (myself).

2.2. Introduction

For many years, the field of microbiology depended on traditional cultivation techniques to determine which bacteria are harmful or beneficial to human health. The fact that this technology can only identify a subset of the myriad of microbial communities is one of the most significant drawbacks associated with it. Game-changing molecular diagnostic tools such as PCR [210], DNA fingerprinting [211,212], and NGS [57,80] have become significantly faster, more sensitive, and cost-effective in recent years.

To perform Next-Generation Sequencing (NGS) on the 16S rRNA gene, the initial step involves extracting DNA from the sample. Following that, a specific region of the 16S rRNA gene is amplified and sequenced. Finally, the generated sequences are identified based on their similarity to reference 16S rRNA gene sequences that are present in public databases. The primary advantages of using the 16S rRNA gene as a marker gene include the independence of the method on the culturability of bacteria present in the sample, the ability to determine the relative abundance of all bacteria in the sample, the feasibility of parallel sequencing hundreds of samples simultaneously, and obtaining the results on the same day as sample collection. Moreover, the method allows the relative abundance of all bacteria present in the sample to be determined [213].

However, there are various drawbacks to 16S rRNA gene sequencing. For example, The binding of primers to regions that are not fully conserved across all bacterial species during the amplification of target DNA often leads to bias [60], The high degree of similarity in the 16S rRNA gene among closely related bacterial species limits their identification to the genus level. [214]; In addition, While 16S rRNA sequencing can indirectly detect changes in microbial community structure that may indicate the occurrence of horizontal gene transfer (HGT) events, it cannot confirm the presence of HGT or identify the specific genes that have been horizontally transferred. Therefore, while 16S rRNA sequencing can provide valuable insights into microbial community composition, it is not the most appropriate method for studying the details of HGT events.

2.2.1. Issues in 16S rRNA gene sequencing

2.2.1.1. The choice of 16S rRNA gene primers affects the microbiome analysis

All prokaryotes have a copy of the 16S rRNA gene, which has both rapidly evolving areas that help us to categorise them into distinct taxonomic groups and slowly evolving regions that are highly conserved among different species. Species-specific fast-evolving regions can be isolated by using primers designed from the slowly-evolving regions to perform PCR amplification. The forward primer in a primer pair is designed to bind to the sense sequence of the bacterial 16S rRNA gene, and the reverse primer is optimised to bind to the antisense sequence [215]. The selection of the primer pairs plays a significant role in determining how accurate the 16S rRNA gene sequencing will be. Despite the fact that environmental microbiologists estimate that less than 2% of bacteria can be cultured in the laboratory[216], many of the bacterial 16S primers currently in use were designed using sequence data obtained from species that were cultured *in vitro*. This was done in order to improve the accuracy of the primers.

The selection of an appropriate primer and/or V-region is dependent on a number of factors including, but not limited to, technical properties such as the read-length that can be sequenced using a defined methodology, the environment that is being targeted, and whether or not it is desirable to compare the results of the current study to those of previous ones. The most targeted V-regions are typically V1-V2/V3, V3-V4/V5, or V4 [55,217,218]. Both the resulting taxonomic profiles and the resolution of the taxonomy are affected by the primer or V-region which is used [219,220]. In addition, the capacity of various V-regions to detect particular bacteria is inconsistent, and the affinities of various primer pairs for DNA binding are not identical; this results in the introduction of biases throughout the PCR process [221,222]. During the amplification step, other sources of bias can come from the amount of genomic DNA that is given as input, the number of cycles that are employed in the PCR process, the procedure itself, or the type of DNA polymerase that is being used [223–225] .

2.2.1.2. **Clustering/denoising**

The first step in preprocessing is to identify and remove low quality reads. After being quality-trimmed, raw sequences are assigned to corresponding taxonomic labels (for example, assigning a sequence with the genus *Bacteroides*) to provide more insights for subsequent studies. Direct assignment of sequences to phylotypes and OTU-based processing are the two basic methods to identify taxonomic nomenclature [226]. When sequences are directly allocated to a taxonomy based on the similarity of the sequences with reference databases, the accuracy of the sequencing platform and reference databases heavily influence the assignment. When reference databases are lacking, assignment issues could occur because newly discovered sequences from an experiment might not fit into established taxonomic lineages [226].

A popular alternative strategy is to organise sequences into OTUs based on sequence similarity and perform analyses on these groups. OTUs are clusters made by combining sequences that share a particular degree of similarity. Typically, sequences that are 97% identical are regarded as belonging to the same species. A representative or average sequence from each OTU is typically then compared with a reference database to get more useful information about that OTU's taxonomy.

OTUs may not closely resemble true biological species, as they are created without the aid of reference data, and sequences within a particular group may be related to several taxa. OTU generation may affect sample diversity estimates, too [226,227]. When using OTU-based approaches, abundance tables (also known as OTU tables) can be generated using both phylotype and OTU-based methods. Numerous publicly accessible toolkits [91,95,228] have been developed to streamline sequence processing and analysis. These suites provide iterations of numerous standard algorithms for sequence processing, taxonomic characterisation and analysis.

There have been numerous methods proposed for defining OTUs such as closed-reference based methods, open-reference based methods and *de novo* methods.

Closed-reference based OTU clustering, on the other hand, uses a pre-existing reference database to compare and group sequences. In this method, sequences are compared to the reference database, and any sequences that match a reference sequence above a certain similarity threshold are grouped together as an OTU [229]. Closed-reference is more accurate than open-reference but can be more computationally intensive. If sequencing reads are sufficiently identical to the related reference sequence, they are allocated to a closed-reference OTU. Consistent labelling is the ability to compare closed-reference OTU assignments between independently processed data sets when the same reference database is used. The use of a reference database in closed-reference OTU picking allows for faster and more accurate classification of reads, as the reference sequences have already been annotated and validated. However, it can also lead to a loss of information, as reads that do not match the reference sequences are not considered in the analysis.

The third method, *de novo* clustering is a method of grouping similar sequences together without using any pre-existing reference database. The sequences are compared to each other, and groups are formed based on a defined similarity threshold. The advantage of this method is that it can discover new OTUs that may not be present in a reference database, but it is less accurate than open and closed reference based. A variety of methods for creating these clusters have been devised. The definition of *de novo* OTU is reliant on the relative abundances of the community being sampled, even under ideal conditions of zero sequencing errors and unlimited sequencing depth. This data set dependence implies that it is not merely a practical concern. Consequently, *de novo* OTUs identified in different datasets cannot be compared.

Open-reference OTU clustering [230] can be defined as a hybrid method combining both closed reference and denovo clustering. It starts with a reference database of known microbial sequences. The sequencing reads are first mapped to this reference database, and reads that do not match any known sequences are then clustered de novo, as described above. The result is a set of OTUs that includes both previously known and novel taxonomic groups. This method is generally more conservative than de novo clustering and is thought to be more accurate, especially when the reference database is of high quality. One advantage of open-reference OTU clustering is that it can detect rare or low-abundance taxa that may not be present in the *de novo* clustering results. However, the choice between de novo and open-reference clustering will depend on the research question, the quality of the reference database, and the desired level of taxonomic resolution.

Unlike OTU clustering, the ASV or zOTU approach aims to identify specific sequences within a pool of reads, rather than grouping similar sequences together as in OTU clustering [98]. This is achieved by determining which specific sequences were read and how frequently each sequence was read. Afterward, an error model for the sequencing run is integrated with this data, enabling the comparison of similar reads and the determination of the likelihood that a particular read with a given frequency is not due to a sequencing error. This creates a p-value for each specific sequence, where the null-hypothesis is equivalent to that specific sequence being a consequence of sequencing error. In other words, sequence-specific p-values can be used to evaluate the statistical significance of each identified ASV and can be used to filter out low-quality ASVs and to prioritise the analysis of ASVs that are statistically significant. This approach can improve the accuracy of downstream analyses and help to identify biologically meaningful differences between different groups of samples. Once the sequences have been denoised and corrected, the abundance of each unique sequence is calculated based on the number of times that sequence is observed in the data. This abundance is typically reported as a count or a relative abundance, which is the count of each unique sequence divided by the total number of high-quality reads in the sample. This approach allows for more accurate detection of true genetic variations within a sample **(Fig. 6**).

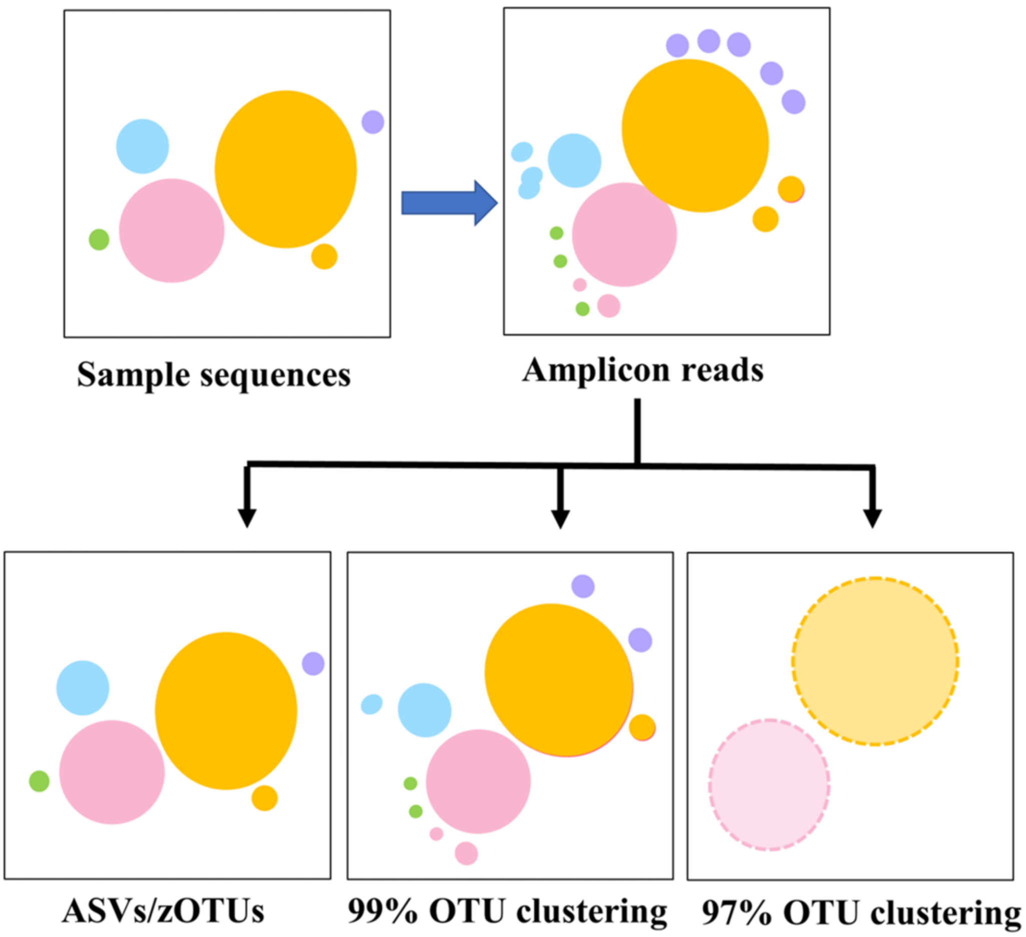


Figure 6: Different types of clustering methods utilized in 16S rRNA gene analysis. Errors are induced during the amplicon sequencing step as shown at the top part of the figure. These errors can represent presence of chimera, misrepresentation by sequencing depth or primer bias. Traditional OTU clustering with 97 % sequence similarity considers them to be a microbial species and cluster together without separating between true species

and erroneous sequences. OTU clustering with 99% performs relatively better than the 97 % clustering. However, ASV denoising algorithms are able to identify biases and help in obtaining much greater resolution of the original diversity. The source of sequencing error due to high-throughput sequencing and the various clustering methods such as OTU clustering with 95 and 97% sequence similarity and denoising algorithm (ASV or zOTUs) are depicted in the diagram. Size of the circles represent abundance and colour different sequence identities. (Source: own work)

ASV techniques can provide more accurate microorganism identification and give a clearer understanding of the diversity present in a sample. In contrast, OTU clustering methods group similar sequences together into an abstracted consensus sequence, which can lead to the inclusion of sequencing errors and closely related species of microbes within a single OTU. This can result in an overestimation of the diversity present in a sample. ASV methods, on the other hand, identify exact sequences and only slight variations between them, providing a more accurate representation of the variety present in a given sample and allowing for a more fine-grained analysis of the data [98,231,96,232]. The authors of the Prodan *et al.* [233] study evaluated and contrasted the six different bioinformatic pipelines for the processing of amplicon data. They examined six different pipelines and found that the sensitivity and specificity varied greatly between each of them. Therefore, the selection of pipelines should be done very carefully and only after testing.

2.2.1.3. Selection of reference databases

After the denoising step, a key aspect of microbiome investigation is taxonomic classification. For conducting sequence-to-taxon matching, bioinformatic pipelines like *QIIME 2* [91] and *Mothur* [95] rely on 16S rRNA gene sequence databases. There are almost nine million rRNA sequences of bacteria, archaea, and some eukarya in SILVA [101], one of the most popular rRNA gene sequence databases. Sequence duplications and inconsistent coverage of clades in these data repositories had been suggested due to the complexity of the data sources [234]. Additionally, updating the database requires a lot of work. Other popular databases with extensive taxonomic annotations are RDP [235] and GG [70], EzBioCloud [236], 16S-UDb[234] and also tissue specific databases such as HOMD [237].

Ribosomal Database Project (RDP)

The phylogenetic categorization for the prokaryotic organisms in the INSDC sequencing databases is provided by the RDP [235]. In 1989, the University of Illinois at Urbana-Champaign began working on the project. When the first release was made available in 1992, there were only 471 16S rRNA gene sequences. The project was transferred to Michigan State University in 1998 after a few releases. Sequences have been screened with Pintail [238] to find chimaeras and faults in assembly of sequences, comparing each sequence to a number of high-quality related sequences using the 16S rRNA gene model intragene variability. If every comparison is unsuccessful, the sequence is tagged as questionable but is left in place. RDP now has 2 110 258 sequences, 2 017 562 of which are classified as bacterial sequences (292 001 of which come from cultivated organisms), and 92 463 of which are classified as archaeal sequences (3442 of them coming from cultured organisms). The 9319 bacterial and 380 archaeal type strain 16S rRNA gene sequences, that form the core of the RDP taxonomy, serve as a link between taxonomy and phylogeny [239,240]. You can browse and download the RDP taxonomy as well as all of its related services at http://rdp.cme.msu.edu.

SILVA

The small and large rRNA components for bacteria, archaea, and eukarya are phylogenetically categorised by SILVA [101] and stored in the European Nucleotide Archive. For the 16S rRNA gene, SILVA offers two alternative alignments: SSU Parc, which has 2 492 653 sequences, is designed for biodiversity assessments; it has rudimentary quality filtering and no guidance tree. SSU Ref, which has 618 442 sequences, is designed for phylogenetic analysis and probe design. It only contains high quality, virtually full-length sequences. SINA (SILVA Incremental Alignment) is used to compute these alignments incrementally, beginning with a carefully selected seed. The same method used in RDP, Pintail [238], is employed to screen out aberrant sequences. The ARB software package serves as the foundation for SILVA [241]. The latest version of SILVA is 138.1 and can be accessed at https://www.arb-SILVA.de/.

GreenGenes

The 16S sequences in GenBank [242] are phylogenetically categorised by Greengenes (GG) [243]. One of its objectives is to incorporate ideas and manual edits from its user base, from the replacement of some database fields to the suggestion of a new name for a taxonomic group. A donor taxonomy (RDP, GG, GG normalised, NCBI or SILVA) can be transferred from a donor taxonomy to a recipient tree using the tax2tree tool, which is included in the most recent version of the database. A database of known non-chimeric sequences can be used with both of the programmes now in use, *UCHIME* [121] and *ChimeraSlayer* [51], although *UCHIME* also operates in a de novo mode. The ARB software suite is compatible with GG [70], which offers an import filter to maintain a local ARB database's synchronisation with the GG database. There are 1 049 116 aligned sequences with a length of more than 1250 nucleotides in the GG taxonomy. GG offers full-length and chimera-checked 16S rRNA gene sequences that were assembled from sequences contributed by various curators from a single study. However, there may be several discrepancies in taxonomic nomenclatures at the phylum level across many curators [244]. You can browse and download it at http://greengenes.lbl.gov.

The All-Species Living Tree Project

The All-Species Living Tree Project (LTP) [245] is a global effort to build and maintain meticulously maintained databases of 16S rRNA gene and 23S rRNA gene sequences, alignments, and phylogenetic trees for each type strain of bacteria and archaea. The SSU and

LSU databases, which are both tiny but taxonomically representative, can be utilised as a guide for categorization and identification in a variety of fields of application. The entire tool is available for free at www.arb-SILVA.de/projects/living-tree and is continually updated with newly identified species and their matching nucleotide sequence entries. It includes the full database, alignment, metadata, and tree. The additional value is used in two ways: first, LTP serves as a reference tool for microbial systematics, and second, it provides a source of curated data that other microbial information resources can easily use.

EzBioCloud

The EzBioCloud [236] 16S database contains data on bacteria, archaea, and eukarya and is primarily intended for species-level identification. It encompasses validly published names from LPSN, Candidatus, prospective species, and uncultured microorganisms and includes the full taxonomic hierarchy from phylum to species. Aside from PCR amplicon sequencing, the database also includes 16S sequences obtained from genome assemblies, which are of higher quality [246].

The findings of Sierra et al. (2020) [247] demonstrated that the utilisation of several reference databases could result in changes and disparities in the taxonomic makeup of given samples, particularly at the genus-level classification. Recently, a few studies have suggested that higher taxonomic resolution might be achievable through the utilisation of environment-specific database systems [237,244,248]. For the samples that were evaluated, for instance, Meola et al. [249] could demonstrate that the level of taxonomic accuracy could be significantly improved by making use of their database that was tailored to the environment. To summarise, a number of different factors can have an effect on the taxonomic classification of specific samples; as a result, documentation and consistency are required in order to reduce the impact of these factors when comparing different research.

While many comparative studies have been published on issues such as the impact of different procedures, methods and bioinformatics pipelines on 16S rRNA gene sequencing, there are still areas where more research is needed. For example, It is always unclear which primer set is most appropriate for a particular research question or sample type (in our case human gut) and how choosing specific databases affect the taxonomic characterizations. Comparative studies can also help to identify potential sources of bias or variability introduced by different primer sets, which can be used to improve the accuracy and sensitivity of 16S rRNA gene sequencing. Such studies can help to identify the most appropriate primer sets, reference databases and pipelines for different research questions and sample types, and can help to optimize the design of future studies. Hence, in this chapter, we carried out a benchmark analysis to evaluate the influence of different primer regions, clustering methods (OTUs, zOTUs, and ASVs), databases (GG, the RDP, the genomic-based 16S rRNA Database, SILVA, and The All-Species Living Tree), and bioinformatic settings on microbial composition outcomes. Overall, this chapter is essential for advancing our understanding of the microbiome and for developing best practices for 16S rRNA gene sequencing.

2.3. Material and methods

2.3.1. Preparation of human gut samples and mock communities

Stools samples were collected from a group of 33 healthy volunteers and were directly resuspended via shaking and vortexing. All samples were aliquoted (in 600-μl portions) and stored at −80°C until DNA extraction. A mock community is a specific mixture of microbial cells generated in vitro. Three mock communities with an increased number of known bacterial species were developed.Three communities are as follows: ZymoBIOMICS microbial community DNA standard (Zymo Research; catalogue no. D6306) with 8 bacterial species, (ii) a more complex in-house mock community (ZIEL-I) with 13 different bacterial species, and (iii) an even more complex in-house mock community (ZIEL-II) with 19 different bacterial species. The detailed explanation on extraction of gDNA, primer selection and *in silico* testing, library preparation of different variable regions of the 16S rRNA gene can be found in our published research [55].

2.3.2. Primer-specific feature classifiers

In general, customized feature classifiers that consider the distinct characteristics introduced by sample preparation, sequencing primer, and read length outperform naive classifiers trained on full-length sequences [250]. The RDP [69], SILVA [101], GG [70], GRD, and LTP [245] databases were utilised to develop primer-specific feature classifiers to improve taxonomy classification. These classifiers were created for each V-region or primer pair utilizing the q2-feature-classifier [110], which is a naive Bayes taxonomic classifier implemented in QIIME2-2019.10 [91].

2.3.3. OTU clustering using QIIME1

QIIME1 [91] was used for the OTU-generating approach. The sequence readings are clustered by *QIIME1* at a sequence identity of at least 97%. By using cutadapt 2.10, forward and reverse primer sequences as well as the low-quality reads were removed from the demultiplexed paired-end reads [251]. The trimmed reads were linked together by multiple join paired ends.py, and then multiple split libraries fastq.py was used to create a single fasta file containing all of the samples. OTU abundance tables were produced by employing the *UCLUST* clustering approach and the pick de novo otus.py script located under *QIIME1.* Throughout the de novo clustering stages, mapping files for OTUs were produced, along with representative sequences, sequence alignments, and taxonomic alignment files. The RDP database was utilized as a reference database to identify OTUs with a minimum sequence similarity of 97%.

2.3.4. zOTU generation using UNOISE

The objective of the *USEARCH-UNOISE3* project is to reconstruct the exact biological sequences into zOTUs. The paired-end raw reads were combined using the fastq_mergepairs script from USEARCH version 11 [94], and the primer sequences were removed by using the *fastx_truncate script.* Merging and primer removal steps were carried out prior to quality filtering because primer removal helps to lower the number of errors that are expected, and merging prior to quality filtering improves the base call error estimates that are captured in the overlapping regions, as the author of *USEARCH/UPARSE* suggested [106]. The processed readings were deduplicated before being grouped into zOTUs using de novo analysis. The RDP database [69], version 11 (project release), was consulted in order to assign taxonomies to the representative zOTU sequences.

2.3.5. ASV generation using nf-core/ampliseq pipeline

The nfcore/ampliseq nextflow pipeline was utilised in order to conduct the analyses on the human datasets as well as the three mock communities [252,253]. Processing 16S rRNA gene amplicon sequencing data is made easier with nfcore/ampliseq, which is an end-to-end solution based on the *QIIME2* library. *FastQC* was used to evaluate the sequence reads' quality in their raw form [99]. The cutadapt program was used to trim primer sequences as well as bases that had low quality scores [251]. For the purposes of denoising and synthesising ASVs, The nf-core/ampliseq pipeline utilized the *DADA2* [96] package to denoise and generate ASVs. Truncated lengths for forward reads (250 to 280 bp) and reverse reads (180 to 260 bp) were determined based on the quality profile and amplicon length, and used during the DADA2 denoising process. The correlation between the truncated lengths and the number of ASVs generated was examined, and these specific lengths were applied to the forward and reverse reads.

2.4. Results

We used a systematic approach to evaluate the global influence of a number of characteristics in mock communities with known compositions as well as in human samples **Fig. 7**. First, an analysis was done to determine the best primers to use in order to target the many variable regions of the 16S rRNA gene. We show that the choice of primers has an effect on the taxonomic makeup, which can be seen in a multidimensional scaling (MDS) plot of samples that came from the same donor **(Fig. 8)**. Second, we explored the ways in which the employment of various clustering algorithms and taxonomy assignment methods affects the outcomes of the categorization of bacterial taxonomies, as well as the extent to which these influences are present. The detailed explanation on primer choice influences the estimated microbial composition can be found elsewhere [55].

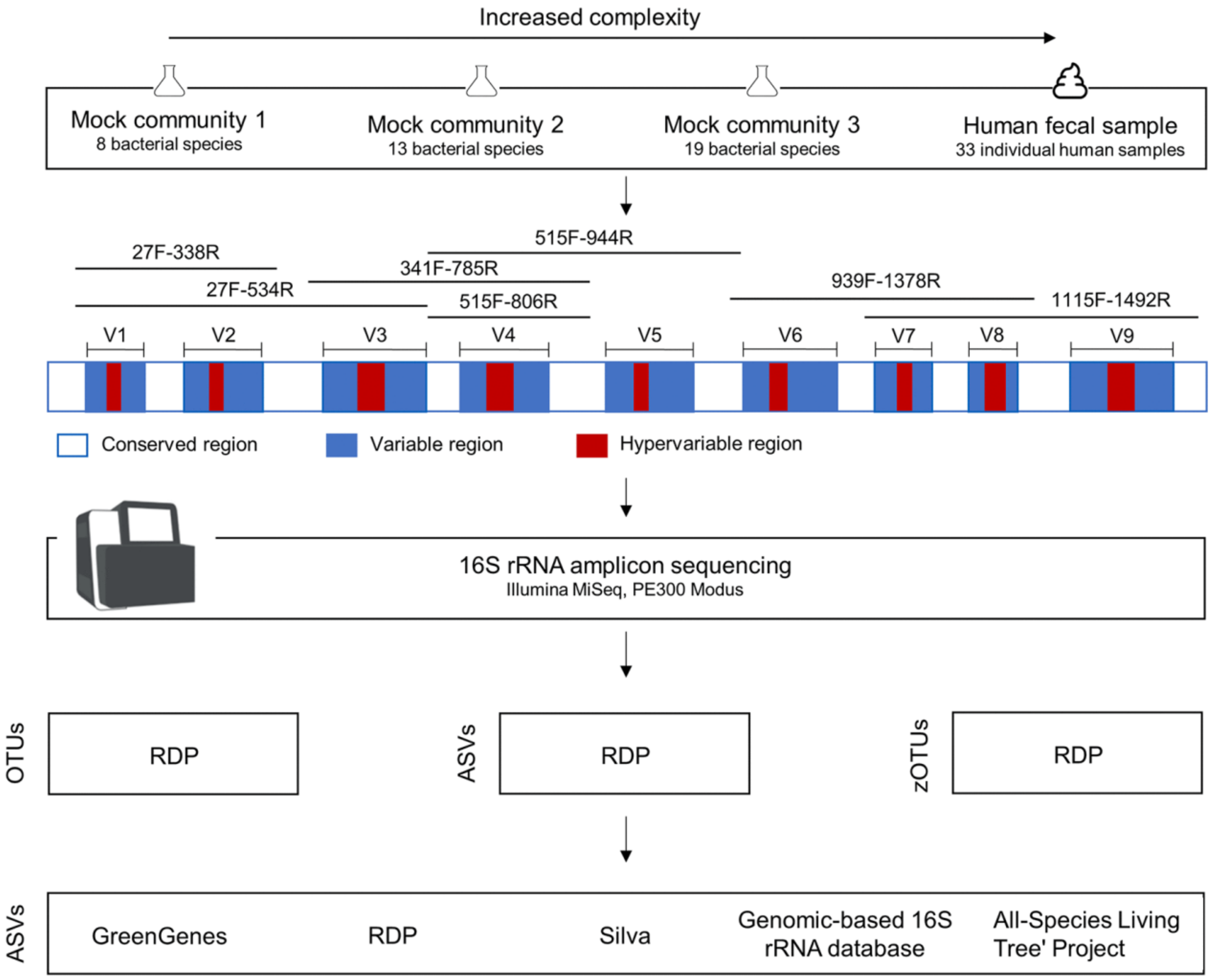


Figure 7: Overview of the analysis strategies used in this study.

A total of 33 stool samples from healthy volunteers and three mock communities with varying bacterial species were analyzed using Illumina MiSeq sequencing. Amplicons were generated using different primer pairs targeting different V-regions of the 16S rRNA gene. The resulting sequencing data were processed and analyzed to investigate the microbial composition of the samples. The study investigated the impacts of different primers, feature generation approaches such as OTU, ASV and zOTU approaches and reference databases on the microbial profiles. This figure was originally published in Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing (open access) under Creative Commons Attribution 4.0 International license

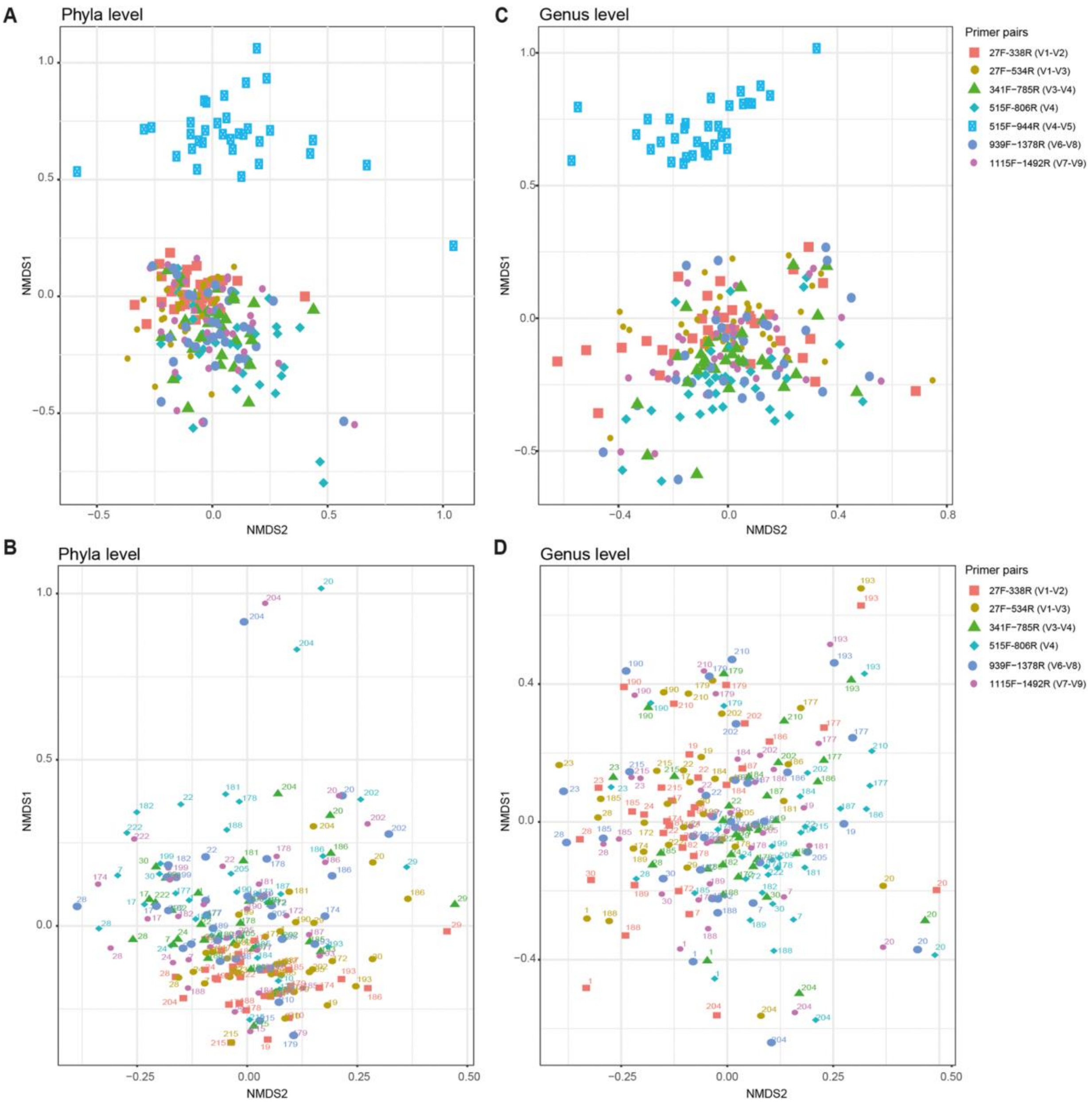


Figure 8: MDS plots illustrating microbiome's distinct communities in human samples across different primer pairs. The top panels (A and C) show the clustering results when the V4-V5 region was included, while the bottom panels (B and D) show the results when this region was omitted. The blue squares in panels A and C represents the 515F-944R primer pair, which appear to be different from all other clusters. The samples in the bottom panels (B and D) are labeled based on donor number, indicating that the study may have investigated the microbial diversity in different individuals. This figure was originally published in Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing (open access) under Creative Commons Attribution 4.0 International license.

2.4.1. Clustering strategies have minor effect on taxonomic profiles

ASV clustering, as an alternative to the OTU technique based on 97% identity, has garnered a lot of attention in the most recent studies [98,247]. In this study, we evaluated a number of alternative ways to clustering to see if any of them had an effect on the taxonomic profiles that were assigned to the ZIEL-I mock community. When compared to the influence of primer choice, the clustering process appeared to have only a marginal impact on the taxonomic assignment (**Fig. 9(A)**). The differences that were seen for each strategy were mostly visible by identification issues at the genus level. *Bacillus* could not be categorised down to the genus level when utilising the ASV technique for clustering the data. On the other hand, this was feasible when the zOTU and OTU techniques were utilised. In general, we discovered that ASVs performed the best for the majority of the other genera. Results obtained from the further study of a human sample subset were found to be equivalent to those obtained from the ZIEL-I mock community (example of one representative sample is shown in **Fig. 9B).** Nevertheless, in the mock communities, neither the clustering of OTUs nor zOTUs created a bigger bias; hence, the influence of clustering is restricted. However, it is important to consider the choice of different clustering approaches when analysing complex environmental samples based on the research question and the samples' characteristics

2.4.2. Reference databases have an impact on the taxonomic assignment

In an ideal scenario, 16S rRNA gene sequences should accurately represent their origin organism. Nevertheless, this relies not only on the primer pairs used or the method of sequence data extraction from raw data, but also on the quality of the reference database and, therefore, the accuracy of the taxonomic classification. To assess this, we thoroughly evaluated five commonly used databases: RDP, GG, GRD, SILVA and LTP.

When evaluating the Zymo mock community with only eight distinct bacteria, we detected only a few slight changes in the taxonomy attributed to the various V-regions employed. Moreover, differences were negligible when other reference databases were included in the investigation **(Fig. 10 (A))**. *Bacillus* could not be categorised at the genus level when using RDP for primer pair 515F-806R (V4), although it was allocated to the *Bacillales* family. Based on our analysis, the taxonomic classification of *Escherichia/Shigella* showed the highest accuracy when using SILVA or RDP as the reference database. This resulted in the smallest deviation from the expected composition of the mock community. The performance of GG database was found to be inadequate in identifying *Escherichia/Shigella* and *Listeria* at the genus level, leading to poor results. Although GG database showed substandard performance in identifying bacterial species in the Zymo mock community, other parameters did not seem to have a significant effect. To gain more insights, two additional mock communities with increased complexity were employed, as a mock community comprising only eight bacterial species provides limited information.

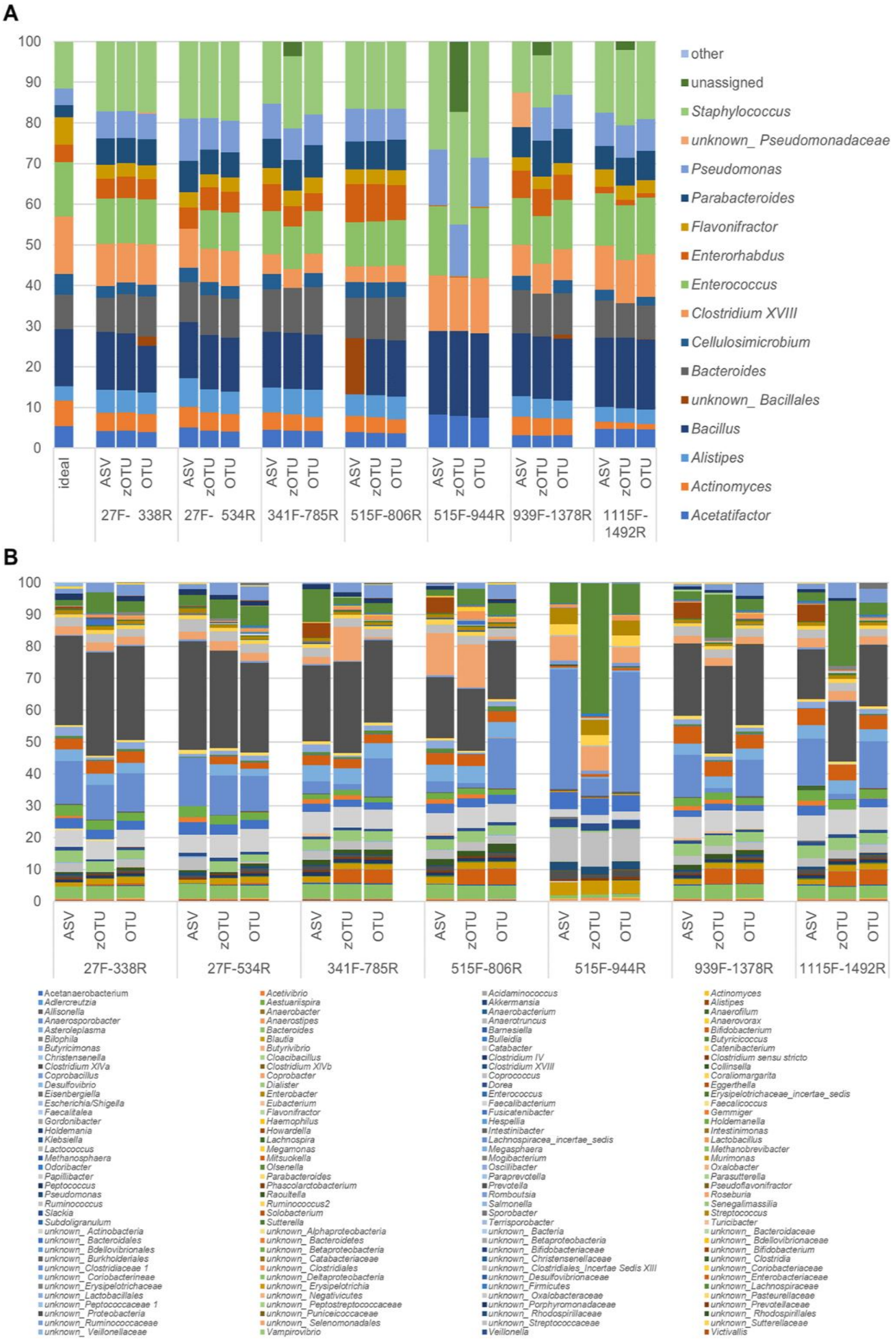


Figure 9: Comparison of the impact of the clustering method on taxonomic categorization for the ZIEL-I mock community. Section (A) represents the taxonomic categorization for the ZIEL-I mock community (A) and Section B represents a representative sample of human DNA (T1). This figure was originally published in Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing (open access) [55] under Creative Commons Attribution 4.0 International license.

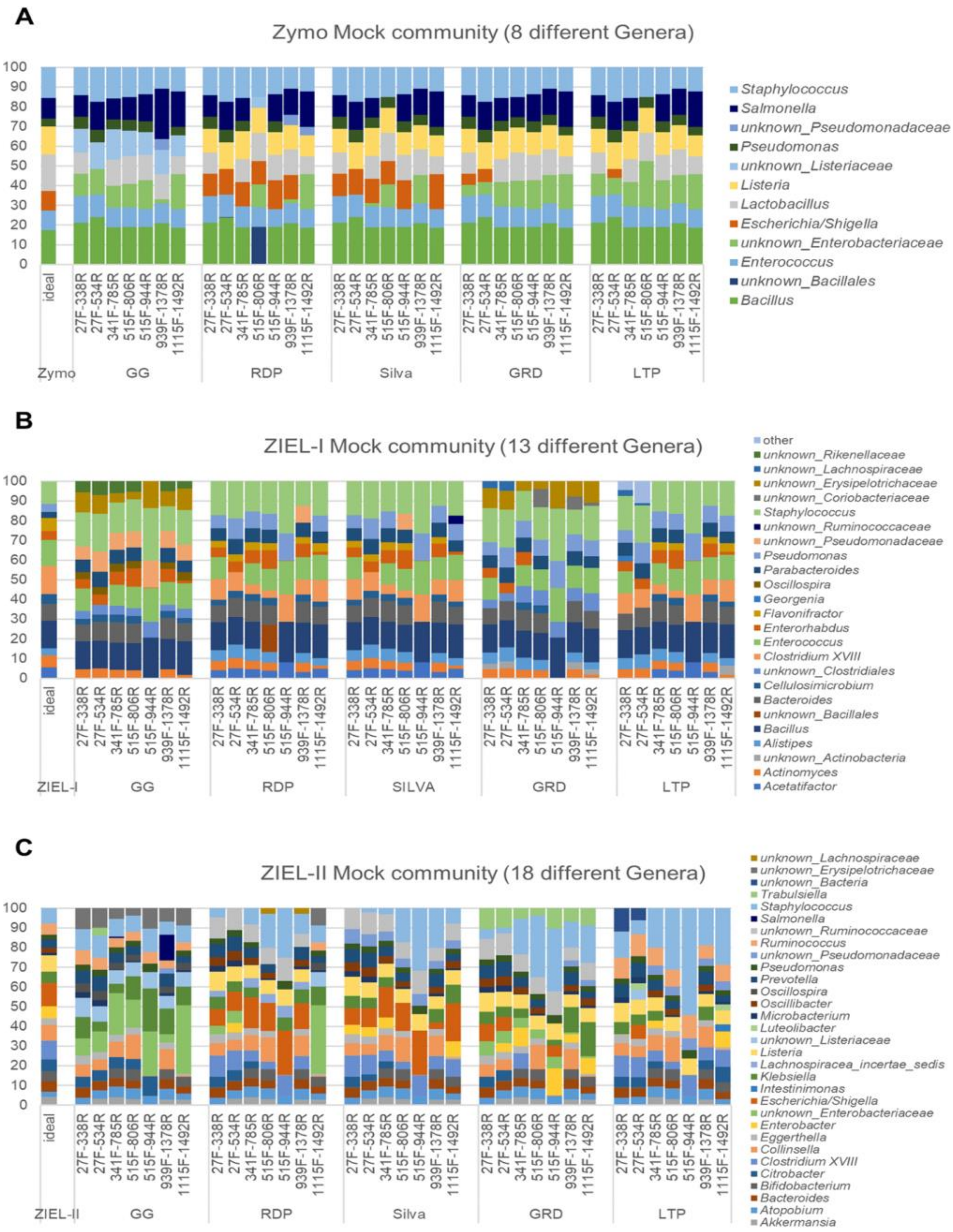


Figure 10: Genus level comparison of mock communities using different primer regions and different databases as references (GG, GreenGenes; RDP, Ribosomal Database Project; GRD, genomic-based 16S rRNA database; LTP, The All-Species Living Tree Project). This figure was originally published in Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing (open access) [55] under Creative Commons Attribution 4.0 International license

The ZIEL-I mock community contains bacteria that are typical of those found in the gut, and it has 13 species spread across 13 different genera **(Fig. 10B).** Taking this into consideration, GG once again had the worst performance. Utilising GG as a point of reference made it impossible to assign genus-level classifications to *Acetatifactor*, *Bacillus*, *Clostridium*, or *Pseudomonas.*

The ZIEL-II mock community including 19 microorganisms in 18 taxa increased the complexity of the comparison. Furthermore, we deliberately included species that posed problems in previous research (data not shown). Again, regardless of the database, the 515F-944R (V4-V5) primer pair exhibited poor performance. Nearly, 14 to 18 taxa were classified at the genus level for primer pair 341F-785R (V3-V4), however only 7 to 9 taxa were identified for primer pair 515F-944R (V4-V5) with SILVA database classification. Using the 27F-338R (V1-V2) primers, *Akkermansia* could not be identified. The taxonomic classification of *Microbacterium* was found to be insufficient when using 341F-785R (V3-V4) primers. In terms of accuracy, Enterobacter and *Ruminococcus* were classified most precisely by SILVA. Overall, SILVA and RDP performed best as reference databases, providing the most precise taxonomic classifications. At the genus level, SILVA had the fewest number of uncertain classifications, followed by RDP, LTP, GRD, and GG.

2.4.3. Specific pipeline settings have minor influences on taxonomic classification

After studying the impact of clustering/denoising and selection of reference databases on taxonomic profiles using mock and human samples, we also investigated the potential significance of certain pipeline parameters on ASVs, as its performance was marginally better to that of zOTUs and OTUs. A number of pre-processing steps such as elimination of primers and adapters, the elimination of chimeras, the trimming of low-quality reads, and the merging of paired-end reads were performed as these help in avoiding false positive feature generation. In the step of merging and removing chimeras, losing sequences can occur if the removal steps are performed incorrectly. Deciding on truncated length should be made with caution as for merging demand a minimum overlap length of 20 bp as well as identical sequences in forward and reverse reads. Yet, we anticipated that the truncation phase would have the most significant influence on the outcomes. In general, truncation is essential because it helps to remove poor quality bases. It is possible to decide the truncated length based on the quality of the bases on both forward and reverse reads and also based on the length of the amplicon which was chosen for the study of interest. In the present investigation, Different truncation lengths were evaluated for forward and reverse reads of the V4 region (primer pair 515F-806R) using the ZIEL-I mock community in this study. On the basis of the quality score (q) of less than twenty and the length of the amplicon, various ranges of forward read lengths (250 to 280 bp) and reverse read lengths (180 to 250 bp) were chosen. According to the results, variations in the lengths of the forward and reverse truncated sequences have a direct impact on the percentages of sequence counts that are retained after the filtering steps (**Fig. 11 (A)).** For example, retaining 90% of the input reads was possible when the forward read length was set to 250 base pairs and the reverse read length was set to 180 base pairs. When the length of the reverse read was increased, the percentage of retained reads went from 90% to 68% during the course of the experiment. The similar pattern of behaviour was seen for forward truncated length combinations of 260 bp and reverse length combinations ranging from 180 to 250 bp. Nevertheless, retaining a lower percentage of reads was achieved when the forward read length was either 270 or 280 base pairs and the reverse read length was between 180 and 250 base pairs. This resulted in a range from 85% to 65% of reads being retained. The number of reads that made it through the filter has significantly decreased, which is the primary cause of the overall drop in the percentage of retained readings. After that, during the processes of denoising and merging, this lower total number of reads was the only one that was processed.

The findings indicate that the retained percentage of reads differed slightly among the various truncated length combinations, but these discrepancies did not have a significant impact on the number of ASVs generated. The total number of ASVs ranged from 10 to 20, depending on the truncated length combinations. The combination of truncated lengths of 250 bp for forward reads and 180 bp for reverse reads yielded the greatest number of ASVs (20), while other length combinations produced only 10 to 15 ASVs **(Fig 11 (B))**.

We conducted a local BLAST search to investigate if the differences in the number of detected ASVs were due to contaminated reads that did not correspond to bacteria in the ZIEL-I mock community. The analysis compared the reads generated using different forward and reverse read combinations with a reference sequence using a ≥97% identity cutoff, ≥90% coverage, and an E value of ≤0.00001. The results showed that 91% to 100% of the ASVs were aligned to the reference sequence of the mock community for each forward and reverse read combination. Only one mismatch was found in the highest number of mismatches. Furthermore, a small number of non-hits were obtained that did not meet the BLAST cutoffs described above. **(Fig. 11 (C)).**

We evaluated the effect of truncation on a diverse microbial community by examining 33 human stool samples. The variation in retained reads after truncation was lower in the stool samples compared to the mock community. The truncation combination of 250 bp for forward reads and 180 bp for reverse reads retained the highest number of reads **(Fig. 11 (D))**. Interestingly, the percentage of retained reads decreased from 89% to 67% when increasing the reverse read length from 180 to 250 bp for 250 bp forward reads, suggesting inadequate removal of low-quality sections obstructs merging. The number of ASVs ranged from 1,219 to 2,363 for different truncated length combinations, leading to an investigation of the impact of the number of ASVs on taxonomic classification at the genus level **(Fig. 11 (E)).** We analyzed the number of generated ASVs using 280-bp forward reads in combination with various reverse read lengths. The number of ASVs ranged from 2,057 to 2,231, and the number of different genera (including unknown and unclassified entries) ranged from 131 to 143.

Overall, the study suggests that the choice of truncated length combinations during data processing can influence the number of ASVs obtained, but the differences may not be drastic. However, it is important to note that this is a specific finding from the study on the ZIEL-I mock community, and the results may differ for other microbial communities or sequencing platforms. However, the study recommends that truncation for each amplicon length should still be tested because low-quality bases can impair read clustering. This means that the quality of the sequence data could affect the clustering of reads into ASVs, even if the reads do not match contaminants. Therefore, it is important to optimize the truncation parameters for the specific sequencing platform and dataset being analyzed to ensure accurate results.

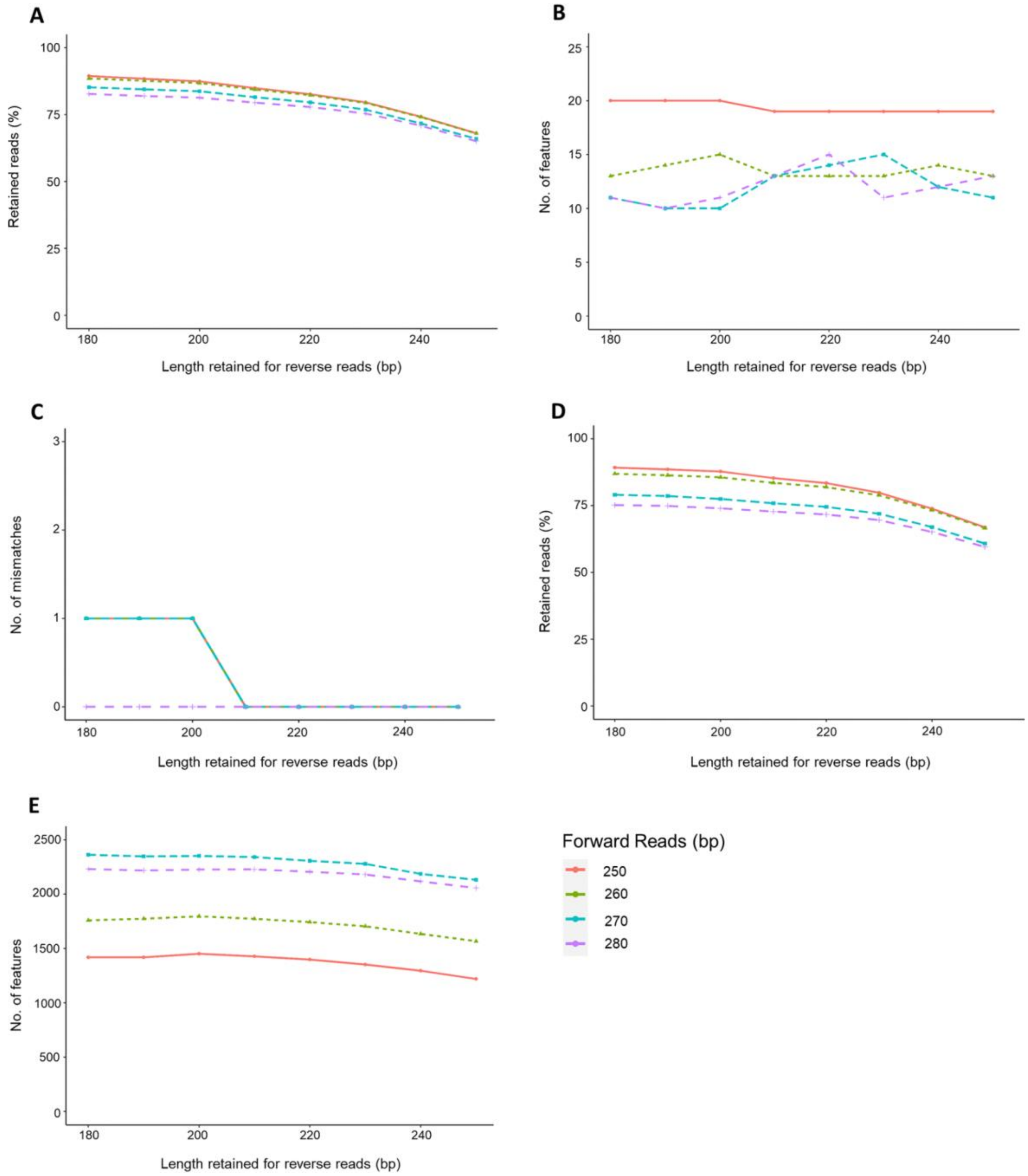


Figure 11: Studying the impacts specific pipeline settings have on taxonomic classification. The impacts of varied lengths of forward and reverse reads after truncation are illustrated in figures (A and B) for the ZIEL-I mock community. The effects are shown on the number of features obtained (A) and the percentage of sequences maintained after denoising (B) for the ZIEL-I mock community. (C) During a local BLAST search against reference sets, the numbers of mismatches that were discovered are displayed; these mismatches served as a manner of measuring how accurate the ASV predictions. The human data set was analyzed with the primary emphasis being placed on retained reads after denoising and truncation (D) and the number of features obtained (E) for each read-length combination. This figure was originally published in Primer, Pipelines, Parameters: Issues in 16S rRNA

Gene Sequencing (open access) [55] under Creative Commons Attribution 4.0 International license

2.5. Discussion

When sequencing the short-amplicon 16S rRNA gene, it is usual practice to utilize primers that span more than one V-region. This allows for more precision in the identification of bacteria when compared to reading a single region. Primers with the sequences V1-V3, V3-V4, and V3-V5 [254,255] are some of the most popular, and have been used in large population-based cohorts like the HMP [48,49,255] and others. Nevertheless, there will be a distinct bias in the results if a different primer combination or V-region is utilised. In addition, DNA extraction and sample processing, sampling, storage, sequencing analysis, and data processing all contribute to the introduction of additional biases. In the past ten years, many of these factors have been investigated for a wide range of ecosystems, including the human gut [256–260], oral and skin microbiomes [261–263] and food-related ecosystems [264]. Despite this, very little research has been done looking at how various factors that cause prejudice interact with one another. In this chapter, we investigated the impact of selecting a particular primer, reference databases, clustering algorithm, and certain pipeline parameters in conjunction on human stool samples and mock communities of varying degrees of complexity utilising contemporary research methods. We intended to provide the scientific community with up-to-date instructions for experimental design and data analysis, thus one of our goals was to highlight the contribution that each of these parameters makes to the precision with which taxonomic assignments are made. In order to draw conclusions beforehand, it is necessary to evaluate the optimal performance of each experimental setting by employing a variety of experimental procedures and settings.

The impact of different primer pairings on the resulting microbial profiles was examined, revealing that the 341F-785R (V3-V4) primer pair exhibited slightly better performance than the other combinations, irrespective of the reference database used. Thus, it is a suitable option for analysing microbial communities in human gut samples. This is also consistent with Thijs et al. [265], who suggested using the primer pair 341F-785R for soil and plant-associated bacterial microbiome research which was also consistent with Rausch et al. [266], who recommended using the V3-V4 region instead of the V1-V2 region. The primer combination 515F-944R (V4-V5) performed well when analysing the microbiota profile of the Zymo mock community, but did not perform well on more complex mock communities like ZIEL-I and ZIEL-II. This indicates that the primer combination may not be suitable for complex microbial ecosystems. Therefore, it is important to include mock communities in routine 16S rRNA gene analysis. The V4-V5 region was considered to be a good match based on theoretical sequence analysis by Yang et al (2016)[267]. but real sample analysis showed poor performance. To overcome the hypervariable region specific issues, It is possible to sequence the entire 16S rRNA gene by utilizing third-generation sequencing technologies [57,268] or by generating short reads and assembling them into a synthetic full-length sequence [269]. These methods have shown potential for providing taxonomic identification at the species or strain level [57]. However, they are not yet widely adopted for high-throughput sequencing due to their lack of cost-effectiveness, repeatability, and user-friendliness. More research is required to make these methods competitive. Furthermore, the error rates for long-read sequencing are still relatively high [270].

Clearly, the limitations of using mock communities to represent the complexity of microbial ecosystems, such as those found in human feces samples, are acknowledged. Therefore, we included 33 human fecal samples in our analysis. Our results showed that the phylum-level classification for Bacteroidetes (excluding 515F-944R), Proteobacteria, and Firmicutes is consistent, regardless of the primer pair used to target the V-region of interest. However, the detection of Actinobacteria, Tenericutes, Lentisphaerae, and Verrucomicrobia varies with different primer pairings, highlighting the importance of selecting appropriate primers. The targeted locations also exhibited considerable variation within individuals of the same genus due to the large number of unknown or unclassified taxa at the genus level and the abundance of taxa in general. To address these issues, ecosystem-specific reference databases and novel bioinformatics methods that integrate data across V-regions while accounting for region-specific bias are necessary. Therefore, large-scale research encompassing various V-regions is essential to train taxonomic classifiers that dynamically account for any region-specific bias. Although sequencing the full-length 16S rRNA gene may render this unnecessary, the choice of primers, such as 27F and 1492R, for virtually full-length sequencing would still influence the results.

It is a well-known fact that the utilization of various bioinformatic methods might affect the microbiota composition that is determined [247,271,272]. To the best of our knowledge, there has been little investigation into the impact of reference databases on taxonomic prediction. To address this gap, we assessed the performance of five different databases using three mock communities. We evaluated each database's ability to accurately identify the correct taxonomy and capture the known diversity of the mock samples. Our results showed that the SILVA and RDP databases were the most reliable 16S rRNA gene databases in terms of true positives at the genus level, with similar performances consistently superior to those of GRD, LTP, and GG. However, our findings were consistent with those of Park and Won [246], who showed that GG was unable to classify certain bacteria, including *Escherichia/Shigella, Listeria, Acetatifactor, Bacillus, Clostridium*, and *Pseudomonas*. As GG was last updated in 2013, its continued use is questionable.

In addition to the aforementioned, we discovered that each database's quality could only be evaluated utilising a range of V-regions and a sufficiently complicated mock community. Using common bacteria to simulate communities with a low level of complexity did not disclose database difficulties. Consequently, low-complexity mock communities may be used as positive controls in existing pipelines for general quality monitoring, but they are not advised for discovering fundamental flaws when establishing a new study, pipeline, or laboratory. In addition, for various body sites (or settings), unique, sufficiently complicated mock communities should be utilised. Adding ubiquitous bacteria, such as the human skin commensal Cutibacterium acnes and other similar bacteria, should be investigated.

In microbial data analysis, taxonomic assignment can be impacted by the denoising and OTU clustering processes, which represent a third element. To investigate this aspect, we compared conventional OTUs, created using 97% clustering with QIIME1, with ASVs generated by DADA2 denoising 96 and zOTUs generated by the USEARCH denoising algorithm [120,273] All three clustering approaches yielded a similar number of identified traits for the simulated community. Interestingly, ASV clustering proved to be highly effective in the human datasets, despite their greater complexity. These results support previous studies [98,103], which have shown that ASVs are the most reliable option currently available. ASVs demonstrated the greatest degree of agreement with the expected composition of the mock community that was tested. Nevertheless, zOTUs demonstrated comparable performance and have the added benefits of being more durable and straightforward to use.

As illustrated, specific variables, such as the truncation length, affect the number of reads preserved for subsequent analysis processes. It is essential to choose an appropriate truncation length, as too-short reads result in insufficient or nonexistent overlaps that cause issues when merging. In contrast, it can be challenging to integrate excessively long reads due to their lower sequence quality. There is a trade-off between integrating readings of lesser quality and the sensitivity for recognizing low-abundance genera, which influences the amount of discovered ASVs for various truncation lengths. By systematically shortening the length of reverse reads, the number of infrequently detected sequences increased while sequencing errors decreased. This demonstrates the significance of this parameter to the reliability of analytical results. To find appropriate truncation lengths in order to evaluate this potential bias, we propose utilizing sufficiently complex, compositionally-known dummy communities. In addition, it is crucial to report this parameter (and all others) in terms of the reproducibility of analysis results.

2.6. Conclusion and outlook

Overall, based on our analysis of 3 mock communities and 33 human samples, we recommend the use of primers targeting the V3-V4 region for human gut samples due to their strong overall performance. We suggest utilising either SILVA or RDP as a reference database. We presently advise adopting ASVs or zOTUs, since only slight changes were identified between clustering algorithms. Regarding pipeline configuration, we recommend testing shortened length combinations for each study's primer pairs. For instance, we would advise to check the quality of the raw reads and decide the truncated length mainly based on the quality score. Also, while deciding the truncated length for forward and reverse reads, we recommend to leave a minimum overlap length of 20 base pairs (bp) to properly merge the paired-end reads. The length of the overlap needed may vary depending on the length and specificity of the primers used in the PCR amplification, as well as the length and variability of the target region being sequenced.

Therefore, it's important to calculate the expected length of the overlap based on the specific primers used in the experiment and to ensure that the length of the overlap is at least 20bp plus any natural variation in the target region. To establish reliable and comparable outcomes, we recommend tailoring the final parameters based on the V-region amplicon lengths. We propose developing a mock community that is specific and complex enough to reflect the targeted microbial environment. This approach can ensure that the study design and analysis pipelines are appropriate for the desired sample type or bacterial community of interest **(Fig. 12).**

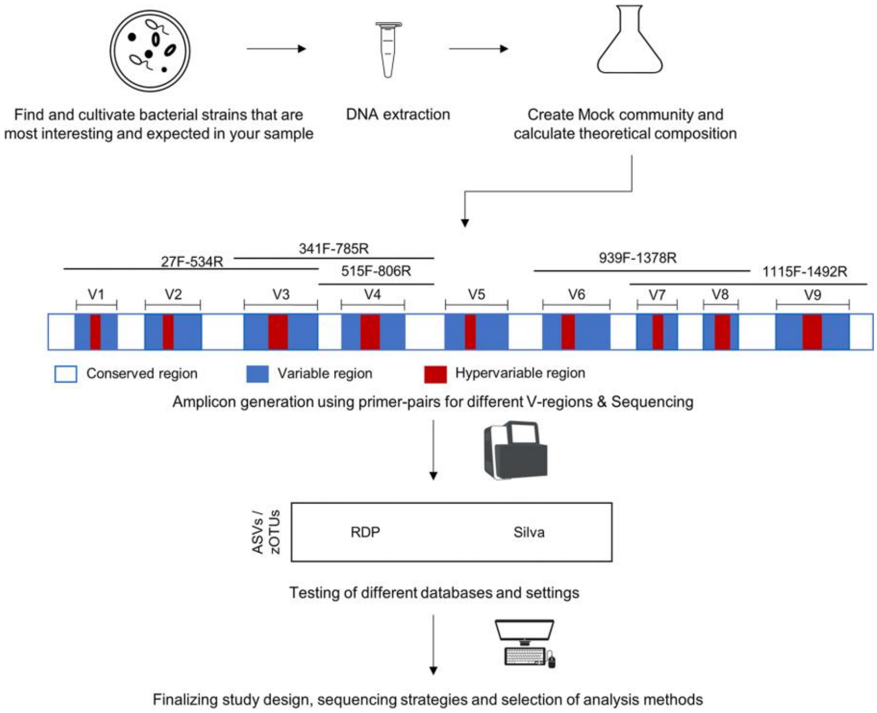

Find and cultivate bacterial strains that are most interesting and expected in your sample
DNA extraction
Create Mock community and calculate theoretical composition
27F-534R
341F-785R
515F-806R
939F-1378R
1115F-1492R
V1
V2
V3
V4
V5
V6
V7
V8
V9
Conserved region
Variable region
Hypervariable region
Amplicon generation using primer-pairs for different V-regions & Sequencing
ASVs / zOTUs
RDP
Silva
Testing of different databases and settings
Finalizing study design, sequencing strategies and selection of analysis methods

Figure 12: Recommendation for a validation process prior to initiating new microbiome investigations, particularly in rare habitats. Prior to beginning new microbiome investigations, the recommended validation procedure should be implemented, particularly for unusual situations. Even pre-existing parameter combinations that are frequently used might be subject to revaluation. As a result, complicated mock communities have to be utilized and sequenced, with a wide range of alternative primer pairs being put through their paces to determine which ones deliver the finest results within the environment of interest. This figure was originally published in Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing (open access) [55] under Creative Commons Attribution 4.0 International license

Chapter three: On the limits of 16S-based metagenome prediction and functional profiling

3. Chapter three: On the limits of 16S-based metagenome prediction and functional profiling

3.1. Declaration of contributions

This chapter is the result of a benchmark analysis of different functional prediction pipelines from 16S rRNA gene sequencing data under the guidance of Dr. Markus List, Head of Big Data in Biomedicine Group, Technical University of Munich, Prof. Tim Kacprowski, Technische Universität Braunschweig, Dr. Malte Rühlemann, Institute of Clinical Molecular Biology, Kiel University, Dr. Fabian Frost, Department of Medicine, University Medicine Greifswald and Prof. Jan Baumbach, Institute for Computational Systems Biology, University of Hamburg.

3.2. Introduction

3.2.1. The Human Gut Microbiome: Function Is Crucial

The diverse communities of microbes that live in the human gastrointestinal system are important regulators of human health and diseases. The development of culture-independent sequencing continues to improve studies of microbial community biology. Shotgun metagenomic and metatranscriptomic (also known as "meta-omic") measurements can be used to answer a growing number of questions ranging from the epidemiology of the human microbiome with respect to biomarkers and therapy, to the transmission and evolution of strains in situ [274–277].

The fact that the gut microbiota helps human in a number of crucial functions attests to its significance. These include the conversion of indigestible dietary components into absorbable metabolites [278], vital vitamins synthesis [279,280], the elimination of harmful substances, providing protection against infections and intestinal barrier [281], and the stimulation and control of the immune system [282]. The majority of these functions are tied to one another and closely related to human physiology. Short-chain fatty acids, for instance, are byproducts of microbial fermentation and are crucial for intestinal cells. They also play a significant role in immunomodulatory processes like T cell development, which can have an impact on the gut microbiota. Overall, it is widely recognised that microbial function rather than its taxonomic composition is considerably more useful [283,284].

The combination of high-resolution, high-fidelity, and high-throughput omics techniques, along with comparative analyses, hypothesis testing in appropriate experimental systems, and intervention studies in human subjects, constitutes a commonly employed strategy for unraveling the intricate web of microbial interactions and identifying potential avenues for enhancing human health. The first kind of research should ideally produce testable ideas on the nature of the roles unique microbiota confer on human physiology, how and why these functions differ between individuals, and their influence on human health. Through the use of 16S rRNA gene amplicon sequencing and metagenomics, the structural properties of the gut microbiota have been extensively discussed in this context [285]. Nevertheless, in order to develop precise hypotheses for mechanistic investigations with the goal of understanding relationships between host and microorganisms, observational studies should link specific activities of the microbiome to specific microbial populations that bestow these functions. In addition, these investigations should uncover biologically relevant and meaningful health status indicators. Functional omics are crucial in this regard.

3.2.2. Functional omics

A comprehensive functional evaluation of the human gut microbiome is now possible with the help of functional omics read-outs generated and integrated from metagenomics, metatranscriptomic, metaproteomic, and metabolomic investigations. Functional omics are more sensitive than the metagenome, which contains a lot of information, according to previous studies

[44,286]. As a result, it is anticipated that functional omics will provide a more realistic depiction of microbes in host's health and disease states [44,287]. For instance, despite only minor changes in observable microbial community structure, changes in gene expression have been discovered in response to dietary treatments such as fermented milk products and the oral consumption of medication . These results appear to run counter to the generally accepted interpretation of metagenomic data, according to which metagenomic functional profiles are less varied than taxonomic profiles [288]. The latter idea, however, may not accurately reflect reality for a number of reasons. In terms of methodology, it has been established that normalization procedures widely used in practice that ignore the taxonomic patterns significantly understate functional variability [289]. Another factor is the clustering of genes into large functional categories, such as complete metabolic modules, based mostly on homology rather than the direction of metabolic flux. Finally, the possible variability of the vast majority of the functioning genes in a metagenome is not considered. Additionally, researchers [290,291] have found that functional profiles in metatranscriptomes are more varied than metagenomic profiles. One possible explanation for this observation is that metatranscriptomes capture the active microbial community at a specific point in time, whereas metagenomes capture the genetic potential of the community [292,293], which includes inactive and dormant organisms. Additionally, functional profiles in metatranscriptomes are often more specific and precise than metagenomic profiles, as they can provide information on the expression of specific genes and their metabolic pathways [294]. This level of resolution allows researchers to better understand the metabolic processes and interactions that are occurring within a microbial community [295,296].

3.2.3. Metagenome approach for functional characterization

To ascertain whether functional omics has the potential to unravel crucial functional aspects of the microbiome, there are fundamental queries that demand resolution such as (i) To ascertain whether functional omics has the potential to unravel crucial functional aspects of the microbiome, there are fundamental queries that demand resolution, and (ii) Can a microbial functional state measured at a single time point offer insights beyond a mere momentary representation? [297]. At the metagenomic, metatranscriptomic, and metaproteomic levels, it is notable that inter-individual variation is observed to be greater than intra-individual variation for functional profiles. This is something that should be taken into consideration [290]. The differences in functional profiles can serve as direct cues to the functions involved in the interactions between the microbiome and the host.

The metagenome approach is a valuable tool for functional characterization of microbiomes and can provide important insights into the role of microbial communities in host health and disease [298] . Metagenomic data can be analysed using tools such as functional annotation pipelines, which can identify the presence of specific genes and pathways involved in various functions. These pipelines typically involve several steps, including sequence quality control, assembly, gene prediction, and functional annotation. For example, after quality filtering, the high-quality reads are assembled into contigs or scaffolds. This can be done using software such as SPAdes [186,187] , IDBA-UD [185], or MEGAHIT [299]. The resulting contigs or scaffolds can then be used for gene prediction and functional annotation. Once the good quality assembly of reads are obtained, the next step is to predict genes from the assembled contigs or scaffolds using several methods such as MetaGeneMark [300], Prodigal [301], or FragGeneScan [302]. The predicted genes can then be used for functional annotation using sequence homology searches, HMMs [303] and machine learning algorithms. Once the genes and pathways are annotated, the data can be analysed to identify functional pathways that are enriched in specific microbial communities or samples. This can provide insights into the functional potential of the microbiome and its role in various biological processes, such as nutrient cycling, host-microbe interactions, and disease states.

There are several pipelines which automate these steps. One such bioinformatics tool is known as HUMAnN 3 [304] (HUMAnN stands for HMP Unified Metabolic Analysis Network), is utilized for the purpose of analyzing metagenomic data, more specifically for the purpose of researching the functional potential of microbial communities found in various environments such as the human gut. Raw data from metagenomic sequencing is input into HUMAnN 3 where it is subjected to a number of processing steps before ultimately producing a comprehensive functional analysis of the microbial community. This analysis includes the identification and quantification of functional pathways and enzymes. In order to reliably identify functional gene families and assign them to the relevant organisms in the community, HUMAnN 3 consults a reference library consisting of the genomes and pathways of microorganisms. Moreover, it normalizes gene family abundances to account for changes in genome size and copy number, making it possible for samples to be compared more accurately. In general, HUMAnN 3 is a useful tool for studying the possible functional capabilities of microbial communities and the ways in which these communities interact with the surrounding environment.

The metagenome sequencing offers a more comprehensive perspective of the genetic diversity and potential functional capacity of the microbial community; nevertheless, it is unable to detect the actual gene expression and metabolic activity of the community at a particular point in time [291]. Metatranscriptomics can provide a more dynamic and detailed view of the functional activity of the microbial community by revealing which genes are actively expressed and how their expression levels change in response to environmental conditions or other factors [169]. This is because metatranscriptomics can reveal which genes are actively expressed and how their expression levels change. In light of the variations that can be found between metagenomic profiles and metatranscriptomic profiles, it is necessary to evaluate the discriminatory power of metatranscriptomics. Metatranscriptomic functional profiles are at least as efficient in addressing differences as metagenomic profiles, according to different studies [290]. As a result, functional omics are able to shed light on microbial activity and highlight crucial microbiome-conferred characteristics. However, metatranscriptomics is still considered rare due to several factors such as technical challenges such as isolation and sequencing of RNA, which is a more technically challenging process than DNA-based sequencing methods as RNA is also more unstable and prone to degradation. The second challenge is the presence of low abundance of RNA present in a microbial community. And finally , the analysis of metatranscriptomic data can be complex and requires sophisticated bioinformatics tools, which may be a barrier to entry for some researchers. Overall, functional omics are able to shed light on microbial activity and highlight crucial microbiome-conferred characteristics.

3.2.4. Limitations of metagenome approach for functional characterization

An alternative approach to metagenome sequencing is 16S rRNA gene sequencing. Despite the fact that 16S rRNA gene sequencing is less expensive than the metagenomics method, it is still considered inferior as it can only identify taxa that can be amplified by the chosen set of "universal" primers. There is a tendency towards preferentially detecting certain groups of bacteria and archaea while overlooking microbial eukaryotes and viruses, which results in a bias in the identification process. Furthermore, metagenomics offers the ability to study the functional potential of the microbiome by examining the prevalence of genes present within the microbial community, whereas the 16S approach is primarily restricted to observing alterations in the taxonomic structure of microorganisms. Even while functional omics approaches, such as metagenome, appear promising, researchers should be mindful of a few drawbacks. For instance, the increased cost of metagenome sequencing impedes its use in research involving a large number of samples, which are often required to establish sufficient statistical power for finding real differences. This is due to the fact that high sample sizes are required to ensure sufficient statistical power for detecting real differences. In addition, metagenome sequencing can be extraordinarily challenging when working with samples that have low biomass or are dominated by DNA from non-microbial organisms [305–307]. In addition, because host contamination drowns out the bulk of the microbial signal generated by metagenomic approaches in many host-associated microbiome situations, the only profiling method that is practically viable is the 16S method [308]. As a result, 16S is expected to be a technology utilised frequently, even as costs associated with sequencing continue to fall.

3.2.5. Functional profile prediction using 16S rRNA gene sequences

The goal of functional analysis is to answer two important questions such as “What kinds of functions are the bacteria able to perform and which metabolic pathways are highly active in a given environment”. It is important to find out what the different metabolic functions of the organisms in the sample are, as well as how many of them have similar functions. The basic idea is to compare OTUs/ ASVs to a reference database that has the functional profiles of microbes and find the best match. For the OTUs/ASVs that are not paired with a known organism, algorithms can look for the organisms which are similar and use them as references to infer their functional profiles.

There are several reference databases commonly used in microbial functional profiling, including: (1) The Kegg Orthology database (KEGG)database [163] which provides a hierarchical classification of genes and proteins based on their orthologous relationship. It includes annotations for metabolic pathways, genetic information processing, environmental information processing, cellular processes, and human diseases across various organisms, including bacteria, archaea, and eukaryotes. (2) The SEED Subsystems database (http://pubseed.theseed.org/) [309] is another database that provides a functional annotation system based on a hierarchical classification of subsystems, which are groups of functionally related genes involved in specific cellular processes or metabolic pathways. (3) The UniProt database is a comprehensive resource for protein sequence and functional information, and includes annotations for biological processes, molecular functions, and cellular components. (4) the EggNOG database [310] provides orthologous groups of proteins, and includes annotations for functional categories based on the Gene Ontology (GO) and other classification systems and (5) the COG database [311,312]provides clusters of orthologous genes across prokaryotes, and includes functional annotations based on a functional classification system. Different databases may be more suitable for different types of microbial communities or research questions, and it is important to carefully evaluate the quality and completeness of the annotation in each database. Additionally, some studies may use multiple databases or combine functional profiling with other types of omics data to gain a more comprehensive understanding of microbial communities.

Among these databases, KEGG is the most widely used reference system for functional annotation. Each KO term is assigned a unique identifier and includes information on the corresponding genes and proteins, as well as links to related pathways and functional categories. To perform microbial functional profiling using KEGG, the metagenomic sequencing data is first analyzed to identify the genes and proteins present in the sample. These genes and proteins are then annotated with KO terms based on their functional annotation, and the abundance of each KO term is quantified. The resulting KO abundance profile can be used to identify the functional pathways and biological functions present in the microbial community. By comparing the KO profiles of different microbial communities, researchers can gain insights into the functional differences between these communities and how they may be influenced by various factors.

While KEGG Orthology (KO) terms are a widely used reference system for microbial functional profiling, there are some limitations to their use. One of the limitations is that KO terms are organized into hierarchical categories, which may not always reflect the complexity and diversity of microbial metabolic pathways. Some pathways may be split into multiple categories, while others may be grouped together in a single category, leading to potential inaccuracies in functional profiling. Furthermore, KO terms are based on the reference database and annotation methods used, which may vary across studies and lead to inconsistencies in functional profiling results. Different databases may also use different terminology and organization, making it challenging to compare results across studies. Lastly, functional profiling using KO terms does not provide direct information on gene expression or protein activity, which may be important for understanding the functional dynamics of microbial communities.

Even though the KEGG database is continually updated and curated, it may not include all microbial genes and functions, and some annotations may be incomplete or inaccurate. This can lead to false positives, where genes are incorrectly annotated with specific functions, or false negatives, where genes with important functions are not annotated. Another limitation is the potential for redundancy and overlap between KO terms, which can complicate the interpretation of the results. Some KO terms may also be highly specific to certain organisms or metabolic pathways, which can limit their usefulness in broader comparisons between microbial communities. To address these limitations, researchers may need to carefully evaluate the quality and completeness of the functional annotation in the KEGG database, and consider using alternative databases or annotation methods to complement or validate their results.

3.2.6. Functional analysis methods

One common problem in predicting functional profiles from 16S rRNA gene analysis is the lack of direct functional information [157,213,313]. The taxonomic assignment of 16S rRNA gene sequences can provide insight into the composition of a microbial community but it does not necessarily provide information about the functional capabilities of the community or functional genes or pathways present. For example, two microbial communities that are taxonomically similar may have vastly different functional capabilities due to differences in gene expression or environmental conditions [314,315]. Therefore, prediction of functional profiles from 16S rRNA gene data relies on statistical inference and machine learning algorithms that use the taxonomic information as a proxy for functional potential. Two commonly used methods for predicting functional profiles from 16S rRNA gene sequences are distance-based methods and phylogenetic methods [157,316]. While these approaches have shown promise, they are still subject to the same limitations as taxonomic assignment based on 16S rRNA gene sequences, and caution should be exercised when interpreting the results.

3.2.6.1. Distance-based algorithms

Distance-based algorithms rely on computing the similarity between OTUs based on their representative sequences and a precomputed distance matrix. The distance matrix stores the pairwise distances between all the OTUs under consideration, where a shorter distance indicates a greater degree of similarity. Following this, the most functionally similar reference OTUs are used to derive the resulting functional profile. The similarity is typically pre-computed and then stored in a distance matrix so that the search can be completed more quickly. The OTUs are represented in the rows and columns of the distance matrix, and the numbers within the matrix reflect the distance between OTUs in the rows and columns that correspond to those numbers. The shorter the distance, the greater the degree of similarity between the OTUs. There are a variety of approaches that can be taken in order to compute the degree of similarity between OTUs. Pairwise alignment is a traditional method which compares two sequences by aligning them and counting the number of matches, mismatches, gaps, and other variations between them. The second method is multiple sequence alignment which compares more than two sequences and computes a consensus sequence that represents the similarities and differences between them. The third method is clustering which groups sequences based on their similaritie using different algorithms such as hierarchical clustering, k-means clustering, and fuzzy clustering.

Phylogenetic analysis: This method involves constructing a tree-like representation of the evolutionary relationships between sequences. Phylogenetic analysis can provide insights into the evolutionary history and divergence of different sequences.

Machine learning: This method involves using algorithms to learn patterns and relationships between sequences and their functional profiles. Machine learning can be used to classify sequences based on their functions or predict functional properties from sequence data.

3.2.6.2. Phylogenetic tree-based algorithms

A phylogenetic tree, also known as a graph in which nodes represent different species, depicts the evolutionary relationships between different organisms and serves as the foundation for this particular group of methods.

These can be used to predict the functional profiling of microbiomes by inferring the functional capabilities of bacterial taxa based on their phylogenetic relationships with other taxa that have known functions. Phylogenetic trees are constructed based on the genetic relations between different bacterial taxa, and the topology of the tree reflects the evolutionary relationships between those taxa. By comparing the phylogenetic tree of a microbial community to a reference database of bacterial genomes with known functions, researchers can infer the functional capabilities of the bacterial taxa in the community.

One common method for inferring functional capabilities from phylogenetic trees is the use of ancestral state reconstruction [317] algorithms, which use the topology of the phylogenetic tree and the known functions of closely related bacterial taxa to predict the most likely functions of the ancestral taxa at each node in the tree. The evolutionary ancestor that is most likely to have been shared by two nodes is their common parent [188,318]. This allows researchers to infer the functional capabilities of bacterial taxa that may not have been directly observed in the microbial community. While phylogenetic tree-based algorithms can provide valuable insights into the functional capabilities of microbial communities, it's important to note that they are based on inference and may not accurately reflect the true functional capabilities of individual bacterial taxa. Additionally, other methods such as metagenomic sequencing and metaproteomic analysis can provide more direct information about the functional capabilities of microbial communities

There are two categories of approaches to choose from when building a phylogenetic tree. Methods based on the distance matrix first precompute the distance matrix that exists between all sequences, and then cluster the sequences in order to compute the tree in such a way that the distance that exists between clustered nodes is as short as it can possibly be.

3.2.7. Introduction to metagenome prediction tools

3.2.7.1. PICRUSt

PICRUSt stands fro Phylogenetic examination of communities through the reconstruction of unobserved states and is a software package written in Python and R. It is freely available under the GNU General Public License.

Fig. 13 illustrates the complete workflow of *PICRUSt.* The entire workflow can be divided into two main parts: Gene content inference and metagenome inference**.** The GG database is utilised by *PICRUSt,* more specifically versions 13.5. The use of GG is the tool's most significant drawback. GG is an outdated database that is no longer being maintained, which means that any results obtained through the use of *PICRUSt* will also be out of date. Despite this, *PICRUSt* continues to find widespread application across a variety of studies [319–322] .

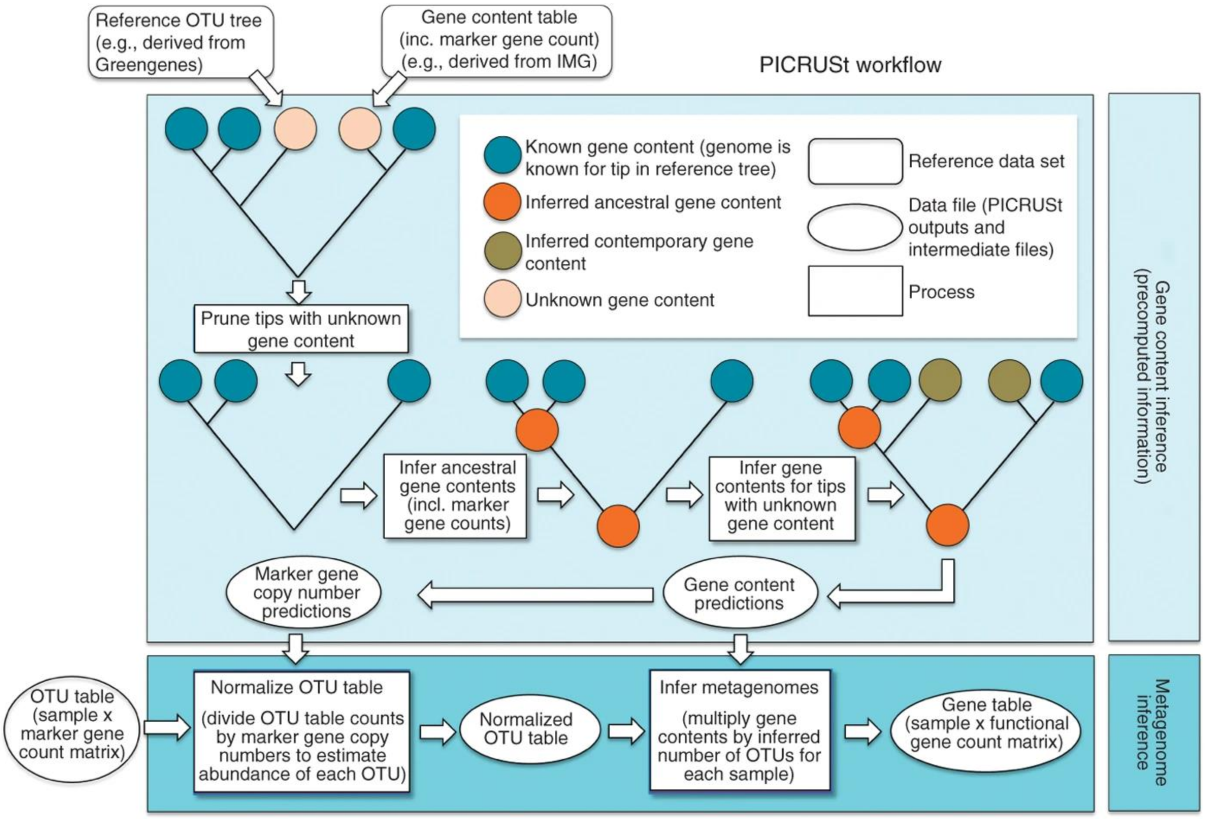

Reference OTU tree (e.g., derived from Greengenes)
Gene content table (inc. marker gene count) (e.g., derived from IMG)
PICRUSt workflow
Known gene content (genome is known for tip in reference tree)
Inferred ancestral gene content
Inferred contemporary gene content
Unknown gene content
Reference data set
Data file (PICRUSt outputs and intermediate files)
Process
Prune tips with unknown gene content
Infer ancestral gene contents (incl. marker gene counts)
Infer gene contents for tips with unknown gene content
Marker gene copy number predictions
Gene content predictions
Gene content inference (precomputed information)
OTU table (sample x marker gene count matrix)
Normalize OTU table (divide OTU table counts by marker gene copy numbers to estimate abundance of each OTU)
Normalized OTU table
Infer metagenomes (multiply gene contents by inferred number of OTUs for each sample)
Gene table (sample x functional gene count matrix)
Metagenome inference

Figure 13: Workflow of two main methods in PICRUSt. The gene content inference workflow as well as the metagenome-inference workflow. The gene content inference workflow uses a reference OTU tree (operational taxonomic units) and a table of gene content. The gene content table lists the genes that are included in reference OTUs. It also includes information about known gene content. The gene content inference workflow uses this information to predict gene content for OTUs with unknown gene content. It also includes predictions of marker gene number. This workflow produces a table of the predicted gene content for all OTUs within the reference tree. The metagenome-inference workflow uses an OTU table. This contains the counts of OTUs per-sample, the copy number of each marker gene, and the gene contents of each OTU created by the gene

content analysis workflow. The metagenome inference workflow uses this information to generate a metagenome table, which includes counts of gene families per sample. This figure is originally from Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. License was requested and obtained under No: 5517000514162 from Springer Nature.

Gene content inference

PICRUSt gets the entire reference tree from GG and precomputes the KO terms profiles in this stage. The outcome is a KO terms profile for each and every bacterium in GG. This step is independent of the sample size and only performed once. The designers of *PICRUSt* precalculated the data for GG versions 13.5 before publishing the results on the *PICRUSt* website for download.

To forecast unknown functional profiles, the gene content table from IMG, which contains functional profiles for known genomes, is employed. The reference OTU tree is compared to the gene content table in order to identify sequences with an uncertain functional profile. Then, a technique for ancestral state reconstruction is employed to generate a phylogenetic tree containing all OTUs from the reference tree. For OTUs without a known functional profile, an estimated profile is derived using the provided OTU's position in the phylogenetic tree and the functional profiles of the OTUs nearest to it.

Although the *PICRUSt* website provides instructions for gene content inference with data from any database of the user's choosing, in practice it is difficult and time-consuming to execute all of the processes. The designers of *PICRUSt* thus developed *PICRUSt2* [157], which differs from *PICRUSt* primarily in its support for multiple reference databases.

Metagenome inference

This phase takes an OTU table that was provided by the user and, with the help of the gene content table from the step before it, predicts the metagenomic content of the sample that was provided. The prediction is made by adding up all of the functional profiles that correspond to the OTUs that are found in the input table and taking into account how abundant those OTUs are. This was accomplished in the previous phase. Since *PICRUSt* cannot cope with OTUs on this level that have unknown functional profiles, the provided OTUs have to be closed-reference chosen against the desired version of GG. *PICRUSt* provides a script that, in the event that the input table was not close-reference chosen, will correct the input table by eliminating all OTUs that are not featured in the precomputed table.

PICRUSt's strengths lie in its transparency and its well-documented design. Disadvantages include dependence on the GG database, challenges in switching to a new reference database, and the mandate that all input data be closed-reference chosen against GG.

Developers of PICRUSt introduced index called nearest sequenced taxon index (NSTI) which is to evaluate the novelty of organisms within an OTU table with respect to previously sequenced genomes. The NSTI provides an estimate of how closely related the organisms in a microbial community are to those that have been sequenced and are therefore available in reference databases. A lower NSTI score indicates that the microbial community is closely related to the sequenced organisms, while a higher NSTI score indicates that the community is more distantly related.

3.2.7.2. PICRUSt2

PICRUSt2 is the most recent version of the *PICRUSt* program. It has the same fundamental capabilities as *PICRUSt*, but the user-provided reference data makes it much simpler to put those capabilities to use.

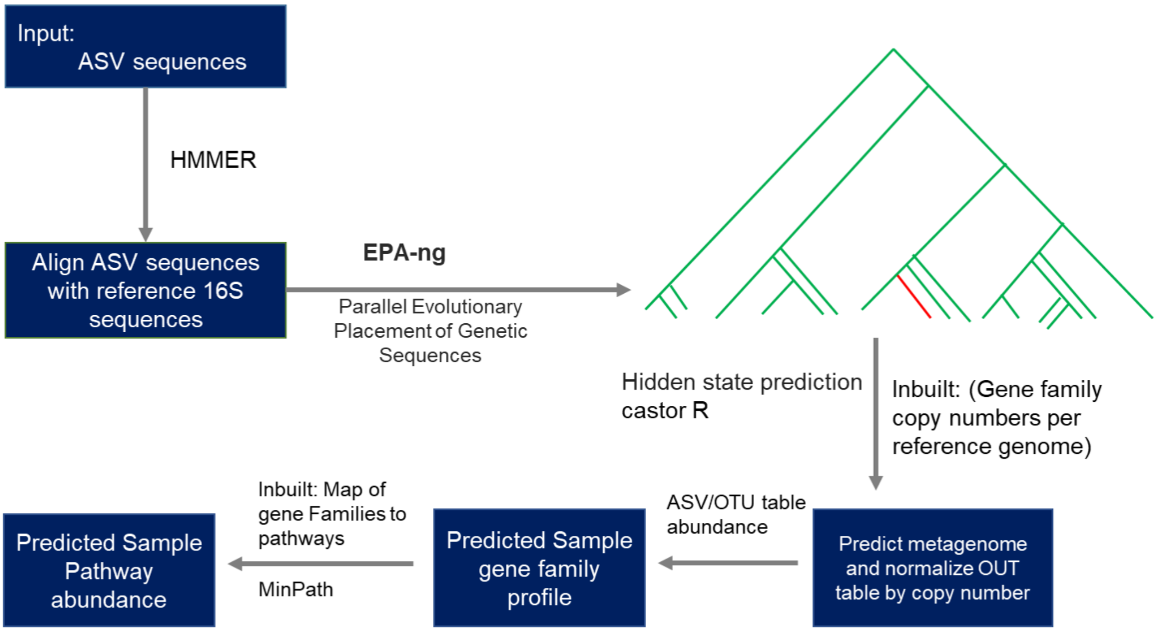

Input:
ASV sequences
HMMER
Align ASV sequences with reference 16S sequences
EPA-ng
Parallel Evolutionary Placement of Genetic Sequences
Hidden state prediction castor R
Inbuilt: (Gene family copy numbers per reference genome)
Predict metagenome and normalize OUT table by copy number
ASV/OTU table abundance
Predicted Sample gene family profile
Inbuilt: Map of gene Families to pathways
MinPath
Predicted Sample Pathway abundance

Figure 14: Simplified workflow of PICRUSt2. The PICRUSt2 method involves several steps, starting with the phylogenetic placement of marker gene sequences onto a reference tree. This is followed by hidden-state prediction, which uses machine learning algorithms to infer the presence or absence of specific genes and pathways based on the observed marker gene abundances. Finally, sample-wise gene and pathway abundance tabulation is performed to generate a table of predicted gene and pathway abundances for each sample in the dataset. (Source: own work)

When compared to *PICRUSt1*'s documentation, *PICRUSt2*'s documentation is significantly more comprehensive and informative. **Fig. 14** provides a visual representation of the flow of data. The gene content inference of *PICRUSt2* occurs in aligning sequence to EPA method while the inference of the metagenome occurs in the following steps. Users have the option of carrying out each stage separately or the pipeline in its entirety. *PICRUSt2* gives users the option to either offer their own reference data or select the computing method and specify its parameters through the use of command line arguments.

PICRUSt2 requires more resources than *PICRUSt* did when it was first released. A minimum of 16 GB of RAM is required to run the first phase of the *PICRUSt2* pipeline, which is the alignment and tree building stage; however, depending on the input data, even that minimum may not be sufficient.

3.2.7.3. Tax4Fun

Tax4Fun [323] is an open source R tool that utilises the SILVA database as reference. It can estimate the functional capabilities and metabolic characteristics of a metagenome.

Tax4Fun employs a different method than *PICRUSt* for OTUs with uncertain profiles. In contrast to *PICRUSt,* which constructs the ancestral tree using the nearest neighbour method, *Tax4Fun* includes a sequence similarity check. Since there is always a nearest neighbour in a tree, *PICRUSt* connects all OTUs, even if their distances are great. *Tax4Fun* joins the nearest neighbours and then applies a linear transformation if the sequences share a minimum degree of similarity. *Tax4Fun* should therefore be more effective for metagenomes including a high number of poorly described bacteria.

The comparison between *PICRUSt* and *Tax4Fun* [30] demonstrates that *Tax4Fun* is more accurate. Since the two programs use different reference databases, this could be due to the superior quality of the SILVA database data compared to the GG database data. To conclusively demonstrate that the *Tax4Fun* method is more efficient, a comparison using the same database would be required.

The implementation in R is an advantage of *Tax4Fun* over *PICRUSt*. *Tax4Fun* is a R package, thus it may be used on any operating system that has R installed. In contrast, *PICRUSt* must be installed and used on a Linux-based machine. RStudio, R's simple and intuitive user interface, is more popular among non-informatics users than Python. *Tax4Fun*'s reference data come from the SILVA database, which is more current than GG in *PICRUSt*.

3.2.7.4. Tax4Fun2

Tax4Fun2 [158] is a package for R that is used for predicting functional profiles and functional gene redundancies in prokaryotic communities using 16S rRNA gene sequences. The package has been shown to be highly accurate and robust, with higher accuracy compared to other tools such as *PICRUSt* and *Tax4Fun*. The functional predictions in *Tax4Fun2* **(Fig. 15)** summarizes the user-supplied OTU table based on the results of the nearest neighbour search. A summary table is used to generate a specific association matrix that only contains the functional reference profiles of the nearest neighbors. By combining the abundance information from the OTU table and functional information from the association matrix, a sample-specific functional profile is created. The predicted profiles are then summarized based on KEGG pathways, with only OTUs meeting a similarity threshold (usually 97%) included in the functional prediction. The unused taxonomic units, which are those OTUs with no close match in the reference data, are recorded as the fraction of taxonomic units unused (FTU), along with the number of sequences assigned to them (fraction of sequences unused = FSU). High values of FTU and FSU may indicate lower quality of predicted metagenomes, as they suggest that the predictions were only made for a small proportion of the total microbial community.

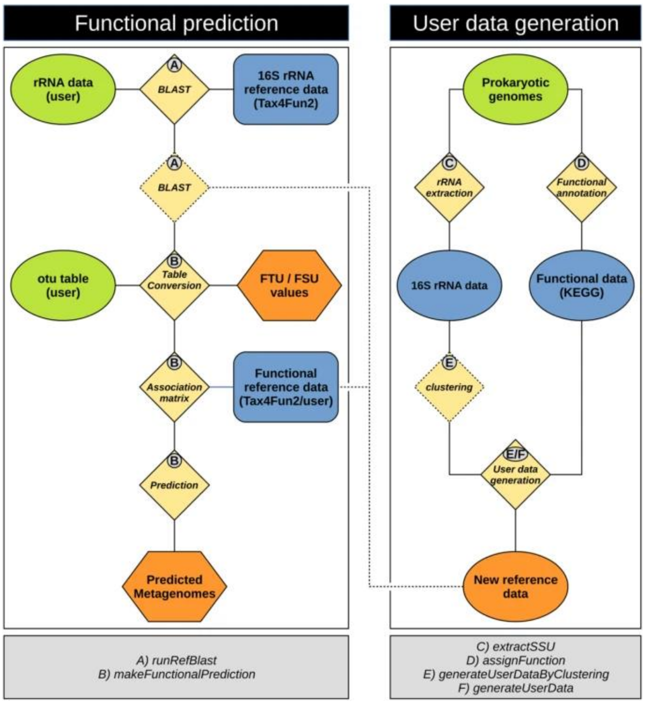


Figure 15: Entire workflow of Tax4Fun2. The Tax4Fun2 workflow commences with aligning 16S rRNA gene sequences against a reference database, which includes the user's supplied reference data. Following this, nearest neighbours are identified and OTU abundances are summarized based on the search outcomes for each sample. An association matrix (AM) is produced, comprising functional profiles of the references identified in the 16S rRNA search. The summarized abundances and functional profiles stored in the AM are merged to predict a metagenome for each sample. Tax4Fun2 offers FTU (Functional Taxonomic Units) and FSU (Functional Sequence Units) values in the log file to provide insights into the community's functional diversity. To generate a habitat-specific dataset, Tax4Fun2 provides functions to functionally annotate prokaryotic genomes and to extract 16S rRNA gene sequences. User-defined reference data sets can be generated and incorporated into the prediction. If many genomes are given, extracted 16S rRNA gene sequences can be clustered using the uclust algorithm as an optional step. This figure was originally published as Tax4Fun2: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences under Creative Commons Attribution 4.0 International license.

3.2.7.5. PanFP

PanFP [159] is a computational method that reconstructs a pangenome from the 16S rRNA gene OTU of known genes and genomes pooled from the OTU's taxonomic lineage The tool employs prokaryotic complete genomes acquired from the National Center for Biotechnology Information (NCBI) and maps KEGG orthologs (KO) terms to proteins via cross-reference ID mapping. This is accomplished using the UniProt KnowledgeBase (UniProtKB) database, which provides cross-reference ID mapping between NCBI Refseq and UniProt. The taxonomic lineage of a taxon is the set of nodes that need to be traversed from the root to the taxon [324]. The method constructs a pangenome for each OTU by generating a superset of all genes that are present in organisms taken from the dataset of prokaryote genomes at the taxonomic lineage corresponding to the OTU in question. As a consequence of this, the OTUs that belong to the same taxonomic lineage share the same pangenome. After that, it accumulates functional compositions in the superset in order to construct a functional profile of the pangenome. In the last step of the process, the OTU-sample table is transformed into a lineage-sample table by consolidating the frequencies of OTUs in a sample with the same lineage. This results in the creation of a function-sample table, where the functional profiles of various lineages are combined using weights corresponding to the relative abundance of each lineage in the sample **(Fig. 16).** By accounting for the range of genomes present within a single taxonomic lineage, this approach enables a more accurate assessment of the functional capabilities of microbial communities.

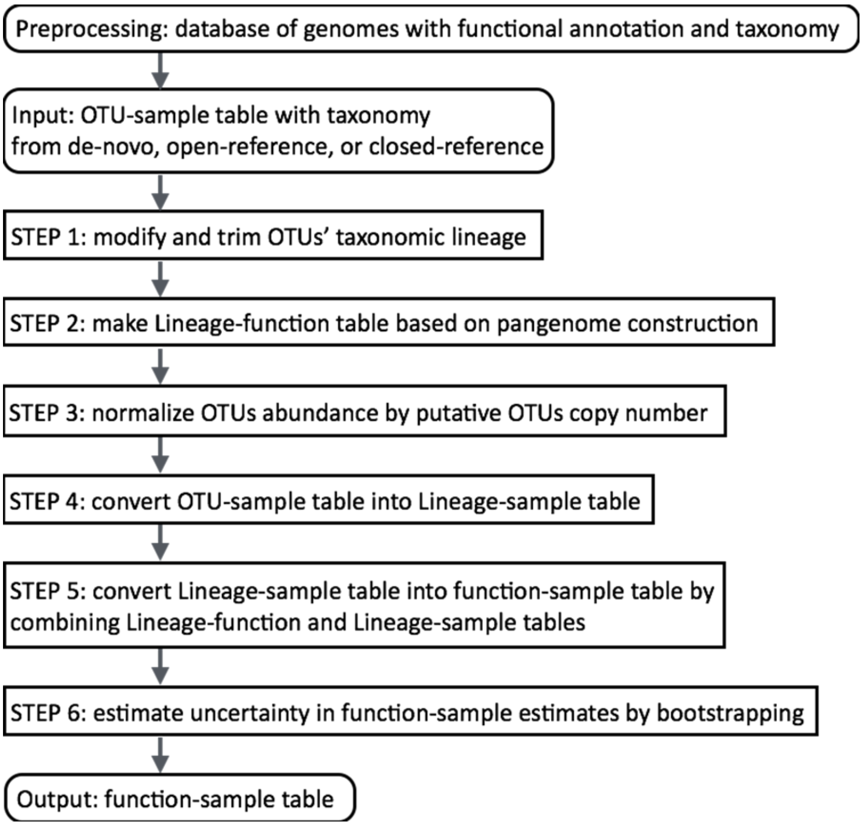
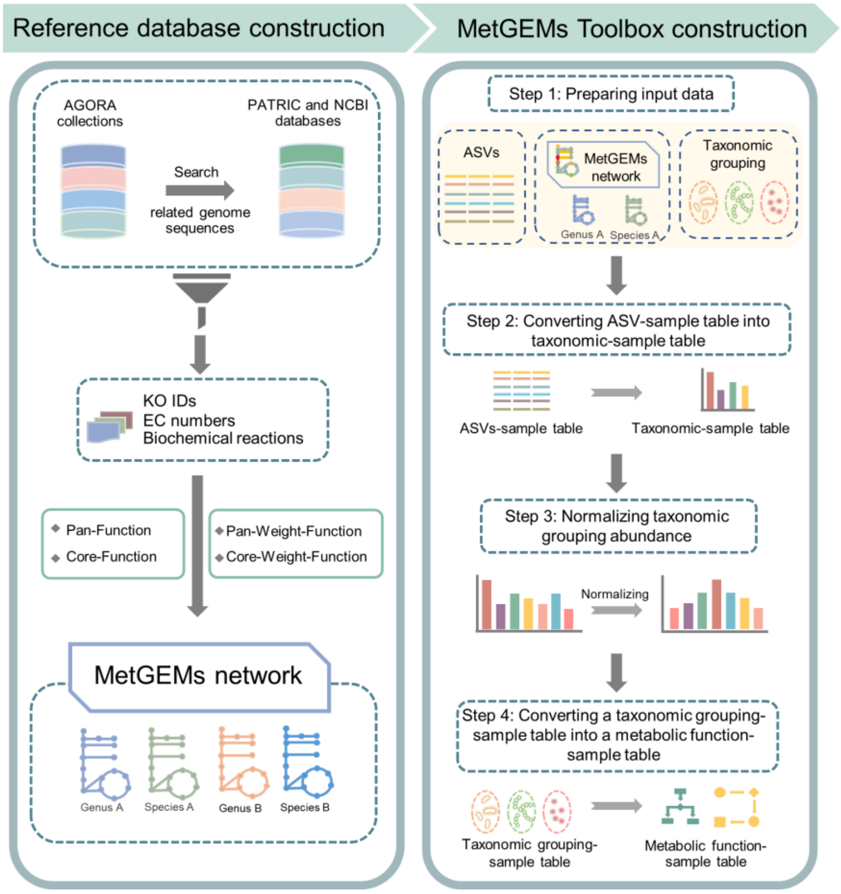

Preprocessing: database of genomes with functional annotation and taxonomy
Input: OTU-sample table with taxonomy from de-novo, open-reference, or closed-reference
STEP 1: modify and trim OTUs' taxonomic lineage
STEP 2: make Lineage-function table based on pangenome construction
STEP 3: normalize OTUs abundance by putative OTUs copy number
STEP 4: convert OTU-sample table into Lineage-sample table
STEP 5: convert Lineage-sample table into function-sample table by combining Lineage-function and Lineage-sample tables
STEP 6: estimate uncertainty in function-sample estimates by bootstrapping
Output: function-sample table

Figure 16: Flow diagram of PanFP. PanFP is a computational tool used for predicting the functional potential of microbial communities using whole-genome shotgun sequencing data. It involves six steps as shown in the above workflow. In short, it starts with the input taxonomic table, trim the taxonomic lineage to a certain level and create a lineage-function table is created by mapping the predicted genes from the OTUs to functional categories using a reference database. This figure was originally published as PanFP: pangenome-based functional profiles for microbial communities (open access) in BMC Research Notes [159] under Creative Commons Attribution 4.0 International license.

3.2.7.6. MetGEM

The *MetGEMs* [162] toolbox utilizes genome-scale models to infer the metagenomic content from 16S rRNA gene sequences, with a specific emphasis on annotating the metabolic functions of the human gut microbiome **(Fig. 17)**. ASV abundance tables along with corresponding taxonomic groups were given as input. Default models such as k_core and e_core were chosen to predict KO terms and Enzyme Nomenclature (EC) abundances, respectively, which was previously shown to provide good estimation in typical situations [162]. Since *MetGEM* does not provide pathway abundances, the output from *MetGEM* was subjected to the *PICRUSt2* pathway prediction step.



Reference database construction
MetGEMs Toolbox construction
AGORA collections
PATRIC and NCBI databases
Search
related genome sequences
KO IDs
EC numbers
Biochemical reactions
Pan-Function
Core-Function
Pan-Weight-Function
Core-Weight-Function
MetGEMs network
Genus A
Species A
Genus B
Species B
Step 1: Preparing input data
ASVs
MetGEMs network
Genus A
Species A
Taxonomic grouping
Step 2: Converting ASV-sample table into taxonomic-sample table
ASVs-sample table
Taxonomic-sample table
Step 3: Normalizing taxonomic grouping abundance
Normalizing
Step 4: Converting a taxonomic grouping-sample table into a metabolic function-sample table
Taxonomic grouping-sample table
Metabolic function-sample table

Figure 17: A schematic workflow of a development MetGEMs toolbox. MetGEMs toolbox is divided into five distinct sections. Firstly, it involves the evaluation of the genome-scale models (GEMs) that are used in the framework. The second section pertains to the implementation of the MetGEMs network within the computational framework. The third section involves the validation capabilities of the MetGEMs toolbox using shotgun sequence data. The fourth section focuses on the MetGEMs toolbox's ability to assign enzyme function and related functional categories specifically in the context of the human gut microbiome. Finally, the fifth section concerns the annotation of putative enzyme functions in allergic disease, which is carried out using the MetGEMs toolbox. This figure is originally from MetGEMs Toolbox: Metagenome-scale models as integrative toolbox for uncovering metabolic functions and routes of human gut microbiome [162] (open access) in plos computational biology under Creative Commons Attribution 4.0 international license.

3.2.8. Previous benchmarking studies in the literature

To this day, there have only been a handful of studies that have been conducted to assess the accuracy of functional prediction tools based on 16S rRNA gene data [165,325–327]. For instance, Djemiel *et al.* [326] conducted a study of one hundred previously published publications on functional predictions using a text mining approach. They then noted the drawbacks, one of which was the absence of reference genomes, particularly for soil ecosystems.

There are also comparisons between tools and their predecessors. Although Aßhauer et al [323,328] compared these tools against one another and to shotgun-sequenced data, they did so only using the four datasets listed in the *PICRUSt1* release publication. Very few studies have employed both programs on their data, and the main aim was only to confirm the final results are agreed with each other. They failed to compare the accuracy of both the tools performed or even to note any underlying biases connected with either program [329,330]. For instance, in their study, they obtained the same results using *PICRUSt1* and *Tax4Fun1*, but *PICRUSt1* also revealed a substantial difference in the abundance of genes encoding cellobiohydrolase between the treatment strategies [329].

In addition to the authors of the *Tax4Fun1* release publication, two other research groups also conducted benchmarking experiments to try to elucidate the differences between *PICRUSt1* and *Tax4Fun1*. In one study contrasting *PICRUSt1* and *Tax4Fun1* on a single dataset, researchers discovered that while *Tax4Fun1* made around 15% more functional assignments overall, *PICRUSt1* performed better on assigning functions that *Tax4Fun1* had overlooked [331]. After considering the pros and cons of each method, they concluded that for the most accurate functional forecast, it was best to employ them both [331]. In the other study, authors only compared their tool Piphillin with *Tax4Fun1* and *PICRUSt1* but not in comparison to one another [332].

Sun *et al.* [165] examined the performance of three widely used metagenome prediction methods *(PICRUSt*, *PICRUSt2*, and *Tax4Fun*) across seven datasets containing paired 16S rRNA gene data and MGS data. Even though MGS data is not a genuine gold standard for functional activity in the microbiome as explained in earlier, it can be used as a ground truth to benchmark 16S rRNA gene functional prediction algorithms. The most common approach to compare the performance between functional prediction from 16S rRNA gene and MGS has been Spearman correlation. However, Sun *et al*. [165] discovered that inferred abundances exhibited a good Spearman correlation between 16S-predicted and MGS-derived gene abundances even when sample labels were shuffled. This demonstrated that functional profiles do not differ as much as changes in taxonomic composition would suggest and that correlation is not an appropriate performance metric for evaluating the effectiveness of functional prediction systems. In addition, the authors demonstrated that evaluating a specific contrast, i.e. the difference between two groups, with a Wilcoxon test resulted in p-values indicating a moderate Spearman correlation between 16S-predicted and MGS-derived genes and a very low correlation after sample permutation. In this study, the authors concluded that functional prediction methodologies were successful for humans but unsuccessful for other animals and environmental samples, and that core functions are more correctly represented than niche-specific activities. In this investigation, however, they found that differences in microbial composition and function are strongly linked to geographic location[333].

3.2.9. Importance of 16S copy number normalisation

While using the 16S rRNA gene as a microbial community barcode gene comes with several benefits, it also has some limitations, such as the presence of biases during amplification and sequencing [334,335], the challenge of accurately identifying and categorizing short sequences taxonomically, and the lack of established benchmark studies to help with the quality control, filtering, and analysis of 16S sequence datasets obtained from new sequencing technologies [89,336]. Another important bias in amplicon sequencing is the number of 16S rRNA gene copies which varies considerably, confounding the abundance prediction of abundance [337,338].

There is a wide range of 16S copy numbers among bacteria with completely sequenced genomes **(Fig. 18);** for instance, *Erythrobacter litoralis* has only one copy, whereas *Photobacterium profundum* has fifteen [339,340]. Because of this copy number variation, differences in the relative abundance of 16S gene sequences in an environmental sample depends on both differences in the abundance of various species as well as differences in genomic 16S copy number [341].

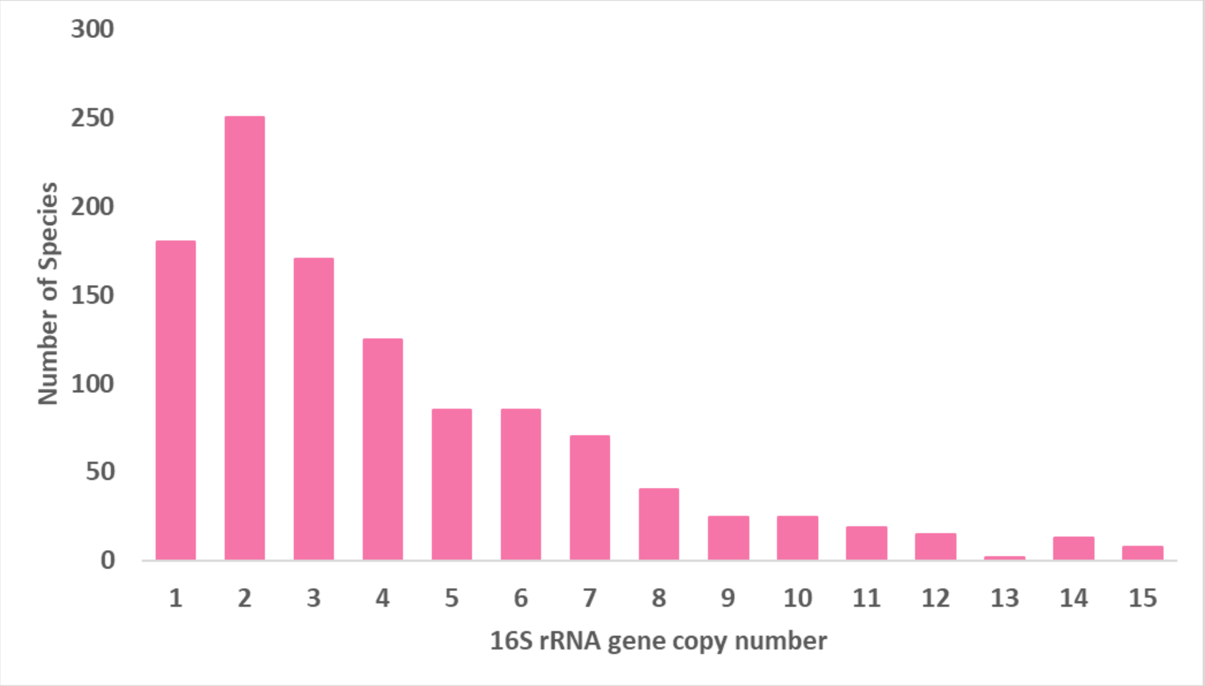


Figure 18: A breakdown of the various 16S rRNA gene copy levels found in bacterial population. Complete genome sequences were gathered from the NCBI and TIGR genome databases, the rrndb database, and the scientific literature (Source own work).

This issue could be overcome by employing a single-copy protein-coding gene as a microbial barcode [342], such as rpoB, although such genes are not as extensively employed as the 16S rRNA gene , and all barcode genes and sequencing technologies are subject to bias. Despite its widespread use in environmental surveys, PCR amplification of 16S rRNA genes may not always be ideal. In these cases, metagenomic data can provide an alternative by allowing the use of genes with more stable copy numbers [343]. Most evaluations of community diversity and composition rely on the assumption that the abundance of 16S rRNA gene sequences accurately reflects the abundance of the organisms possessing those sequences. The impact on estimates of microbial community structure of using 16S rRNA gene abundance as a proxy measure of organismal abundance, and the extent to which this assumption is justified, remain open questions.

Several tools have recently been developed for predicting genomic copy numbers using phylogenetic methods [157] and based on sequenced genomes [344]. rrnDB [122] provides precise and well-documented data on the copy numbers of rRNA operons across prokaryotes. Each entry for an organism in the rrnDB database contains comprehensive information that is directly linked to external databases such as the RDP [69], GenBank [242], PubMed and several culture collections. There are now 27,655 records for Bacteria (representing 7,203 species) and 448 records for Archaea (representing 348 species) in the most recent version (5.8) of rrnDB [122].

3.3. Problem statement

It thus remains an open question if metagenome prediction tools are also suited for more subtle contrasts related to human health. Moreover, Sun *et al.* [165] could not detect any performance differences between the tested methods, suggesting that a more comprehensive benchmark is needed to recommend guidelines for tool selection and to establish the limits of metagenome prediction tools in human disease research. Hence, we considered the most widely used metagenome prediction tools *PICRUSt2* [157], *PanFP* [159], *Tax4Fun2* [158] *and MetGEM* [162] in a systematic benchmark. We did not include Piphillin [160] as it is only available as a web server (the command line version is in the testing stage). We tested the reliability of the prediction tools using matched 16S rRNA gene and MGS human datasets obtained from different cohorts including KORA (type 2 diabetes) [345], FoCuS, PopGEN (obesity) [346], colorectal cancer (CRC) [277] as well as simulation datasets for different functional categories. Our contribution is three-fold: (i) considering human cohorts for type 2 diabetes, colorectal cancer and obesity, we tested if health-related differential abundance measures of functional categories are concordant between 16S-predicted and metagenome-derived profiles; (ii) using simulated data, we investigated if technical biases could explain the discordance between predicted and expected results; (iii) since 16S copy number is an important confounder in functional prediction, we investigated if a customised copy number normalisation with the rrnDB database could improve the results [122]. According to our findings, the current metagenome prediction tools lack sensitivity to accurately identify health-related functional alterations in the microbiome, and therefore, their usage should be approached with caution. Additionally, we have observed notable variations in performance among the individual tools tested, and provide suggestions for selecting appropriate tools.

3.4. Materials and methods

3.4.1. Population-based cohorts

In total, we selected five cohorts with paired 16S rRNA gene and MGS data for functional prediction analysis. The CRC dataset was downloaded from SRA project number PRJEB6070. Datasets of FoCuS, KO termsRA[345] and PopGen [346] are under controlled access due to the informed consent given by the cohort study participants. KORA data are available upon request from KORA (https://epi.helmholtz-muenchen.de/). Dataset for FoCus and PopGen are available upon request (https://portal.popgen.de). For the prospective KORA cohort (2018), sample preparation and sequencing of the V3V4 region in paired-end mode on an Illumina MiSeq was performed by the ZIEL – Core Facility Microbiome [347]. For the FoCus cohort and PopGen cohort, a detailed overview of the cohorts is given in Table 3 and the workflow of the benchmark analysis is described in **Fig. 19**

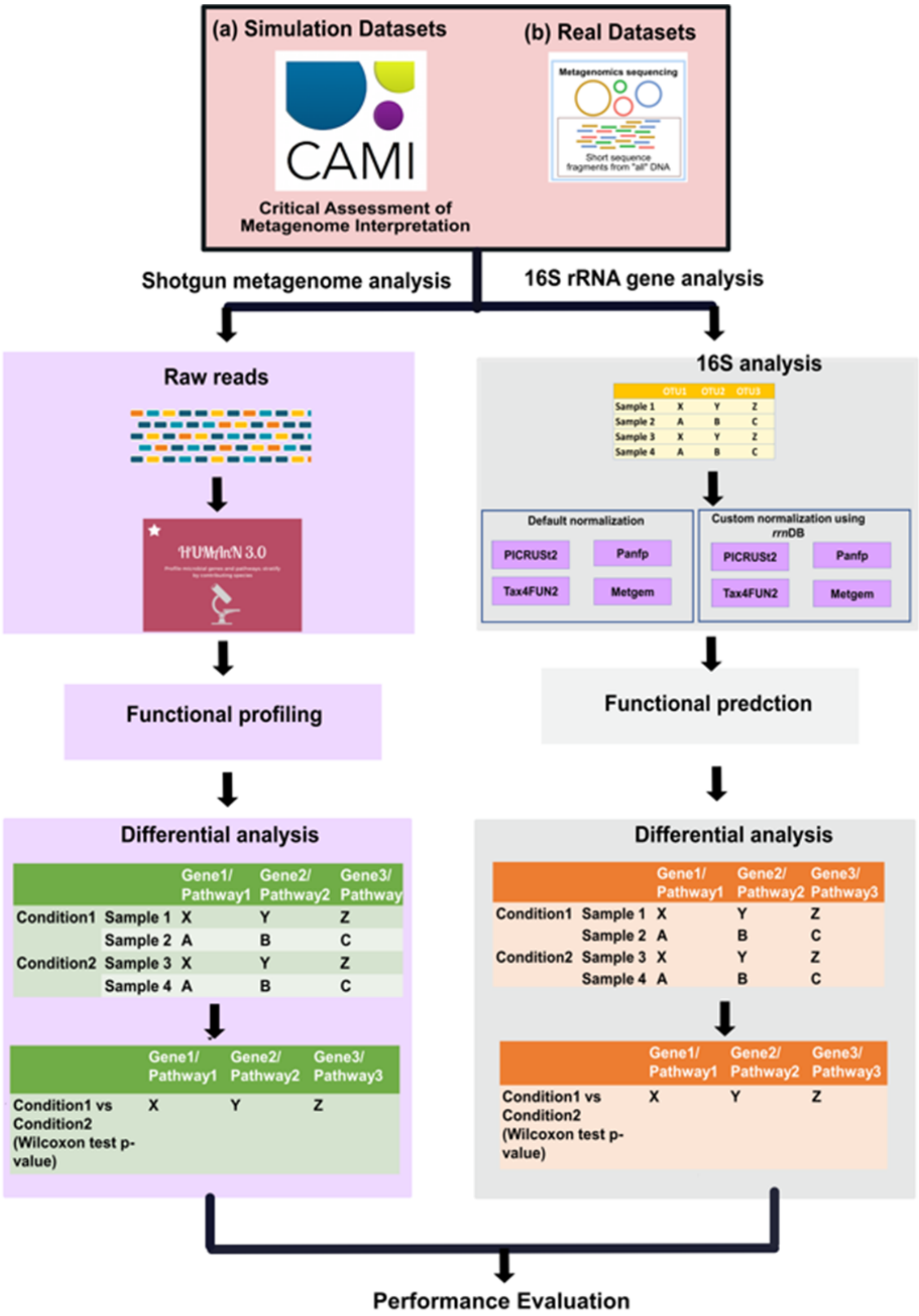

(a) Simulation Datasets
(b) Real Datasets
CAMI
Critical Assessment of Metagenome Interpretation
Metagenomics sequencing
Short sequence fragments from "all" DNA
Shotgun metagenome analysis
16S rRNA gene analysis
Raw reads
HUMAnN 3.0
16S analysis
OTU1 OTU2 OTU3
Sample 1 X Y Z
Sample 2 A B C
Sample 3 X Y Z
Sample 4 A B C
Default normalization
Custom normalization using rrnDB
PICRUSt2
Panfp
Tax4FUN2
Metgem
Functional profiling
Functional predction
Differential analysis
Gene1/ Pathway1 Gene2/ Pathway2 Gene3/ Pathway
Condition1 Sample 1 X Y Z
Sample 2 A B C
Condition2 Sample 3 X Y Z
Sample 4 A B C
Gene1/ Pathway1 Gene2/ Pathway2 Gene3/ Pathway3
Condition1 vs Condition2 (Wilcoxon test p-value) X Y Z
Differential analysis
Gene1/ Pathway1 Gene2/ Pathway2 Gene3/ Pathway3
Condition1 Sample 1 X Y Z
Sample 2 A B C
Condition2 Sample 3 X Y Z
Sample 4 A B C
Gene1/ Pathway1 Gene2/ Pathway2 Gene3/ Pathway3
Condition1 vs Condition2 (Wilcoxon test p-value) X Y Z
Performance Evaluation

Figure 19: Overall benchmarking workflow to compare and evaluate the performance of functional inference tools from 16S rRNA gene sequences to metagenomics.. In this approach, both public and simulated paired MGS-16S rRNA gene datasets were chosen. For MGS, the HUMAnN 3 pipeline was used to retrieve the functional profiles and PICRUts2, Tax4Fun2, PanFP and MetGEM for 16S rRNA gene datasets. Wilcoxon rank-sum test was performed evaluating the null hypothesis for each dataset and comparing the significant KO terms between the two techniques (Source own work).

3.4.2. Bioinformatics pipeline for processing of 16S-rRNA Data

The raw sequencing reads were processed using *QIIME2* [91] First, the paired-end reads with sufficient quality scores were imported into *QIIME2*, and the primer sequences were removed by trimming the first 17 and 20 bases from the forward and reverse reads, respectively. Then, the denoising algorithm *DADA2* [96], was used to infer Amplicon Sequence Variants (ASVs) from the reads while filtering out potential chimeras. A sample metadata file was used to add information about the experimental conditions associated with each sample. Taxonomy was assigned to the resulting ASVs using the SILVA database [101] and the ASV was converted into a biom file format using the biom-convert command.

Table 3: Overview of disease cohorts

Cohort name	geographic location	reference, sample size (paired WGS and 16S)	disease focus	References	Publication
KO RA	Augsburg, Bavaria, Germany	60	Diabetes vs Healthy	https://epi.helmholtz-muenchen.de/	[345]
FoCus	Kiel, Schleswig-Holstein, Germany	101	Healthy vs Obese	https://portal.popgen.de	[348]
Popgen	Kiel, Schleswig-Holstein, Germany	86	Healthy vs Obese	https://portal.popgen.de	[346]
CRC	Germany	182	CRC and Healthy	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB6070	[277]

3.4.3. Generating functional prediction profiles

The ASV tables along with corresponding fasta sequences obtained from the denoising step were given as input into *PICRUSt2, Tax4Fun2* whereas only ASV tables along with taxonomic lineage were given to *PanFP* and *MetGEM. PICRUSt2* was used with a default cutoff NSTI of 2.0 and converted the obtained prediction table functions to relative abundances. Each method has its own method to perform 16S copy number normalisation. The BLASTn command in *Tax4Fun2* was performed with the default setting of 97% similarity and the precalculated files with the default settings. After generating the prediction profile, the relative abundances were directly output by *Tax4Fun2,* so there was no need to perform any transformation. The output of the make FunctionalPredictions() function in the *Tax4Fun2* script was modified so that it now provides the ratio of used sequences rather than the default ratio of unused sequences. *PanFP* and *MetGEM* were also followed as per the documentation.

3.4.4. Functional profiling of metagenomics

KneadData 0.7.4 was employed for sequence quality control and removal of human reads, while bacterial gene abundances were computed using *HUMAnN 3* [304] through nucleotide-based alignment against the Chocophlan database [349]. Gene abundance tables were then grouped by the uniref90_ko command and relative abundance was calculated using humann_renorm_table command in the *HUMAnN 3* pipeline, respectively.

3.4.5. Customised normalisation using rrnDB

To test the effect of copy number normalisation on the functional prediction, we repeated the entire workflow as mentioned above, replacing the gene copy number normalisation step by obtaining the copy numbers from rrnDB. We processed rrnDB with the *PICRUSt2* place_seq.py command (see https://github.com/mruehlemann/16s_cnv_correction_databases for details) and used the resulting abundance tables as input for other tools by skipping their built-in copy number normalisation steps.

3.4.6. Validating prediction tools with shotgun metagenomic sequencing data

We used two methods to evaluate the consistency and accuracy of functional prediction tools. It is important to mention that a direct comparison of the functional profiles produced by all four tools is nearly impossible due to various modifications made to the KEGG Orthology subsequent to the development of *PICRUSt, Tax4Fun2, PanFP* and *MetGEM.* Also, converting raw counts to relative abundances is important in microbial functional profiling because microbial communities can vary widely in composition and sequencing depth. Without normalization, differences in sequencing depth can lead to misleading results, making it difficult to compare the functional potential or activity of different microbial communities. By converting counts to relative abundances, we can compare the proportional representation of different functional categories or taxa across different samples, tools, or datasets in a more meaningful way. This allows us to identify patterns in the functional potential or activity of microbial communities and to make more accurate comparisons between different samples or datasets. As a result, both the functional profiles predicted by metagenome prediction tools and those obtained through metagenomic shotgun sequencing were transformed into relative abundances before being compared. For metagenomics, the counts were converted into relative abundance using the *renorm table* function. We evaluated the Spearman correlation between the gene composition predicted by the tools and that obtained from metagenome sequencing, taking into account only the functions that were present in both the metagenomic and predicted profiles for each comparison.

3.4.7. Cohort-wise differential abundance analysis between cases

As proposed by Sun *et al.* [165], differential analysis is better suited to assess whether metagenome prediction tools are able to detect biological variation between samples. We further analysed the consistency of metagenome prediction tools and metagenome sequencing in terms of p-values they generated for null hypotheses of no association with metadata. Following their approach, we thus tested for differential KO terms abundance between sample groups. In contrast to Sun *et al.* we focused here on the comparison of disease and health as group labels, which we expect to be a challenging scenario. We removed KO terms with low prevalence (< 5% of samples). After filtering, overlapping KO terms and pathway terms between prediction tool and MGS were retained for differential analysis. For this purpose, a Wilcoxon test of the two distinguishable groups (disease versus control) was conducted in each dataset for both default and custom normalisation. Using the CRC dataset, we compared the KO terms of abundance between cancer (n=41) and healthy tissues (n=50). Using the Popgen and FoCuS datasets, we compared the differentially abundant KO terms between obese and healthy samples. Patients with a body mass index (BMI) > 30 were considered as obese and as healthy otherwise. For the KORA dataset, differentially abundant KO terms were tested between type 2 diabetes and healthy controls. The Wilcoxon rank-sum test was applied to each cohort to test the difference between groups for MGS and prediction results and predicted KO terms abundance and significant KO terms with a p-value < 0.05 were extracted. Overlapping significant KO terms between MGS and predicted tools were extracted and compared to evaluate the performance of prediction tools.

3.4.8. Simulation dataset and processing

In addition to the methods recommended by Sun *et al.* [165], we evaluated the functional profiling performance using 40 synthetic samples from the 2nd CAMI Challenge [350,351]. These represent typical microbiomes from four human body sites such as gut (n=10), skin (n=10), oral (n=10) and air (n=10). Metagenome functional profiling of simulated datasets were obtained as described under the shotgun processing steps. 16S rRNA gene sequences were reconstructed in two steps. First, 16S long reads were filtered with the help of filterReads (reference) using SILVA [101] as a reference database. Once the 16S rRNA gene sequences were obtained, functional profiling using all four tools were predicted as described above. We compared the functional prediction with the metagenome profiles and tested their agreement in principal component analysis (PCA).

3.5. Result

3.5.1. Correlation is not a suitable performance measure for metagenome prediction tools

Sun *et al.* [165] previously investigated the performance of metagenome prediction tools and observed that the comparably high Spearman correlation values are not affected by label permutation. We could confirm these findings on five independent disease cohorts where *PICRUSt2, Tax4Fun2* and *PanFP* achieved Spearman correlation values ranging from 0.65 to 0.75 **(Fig. 20)** which did not drop drastically after sample label permutation. *MetGEM* performed slightly worse than its competitors. Using rrnDB copy number normalisation, *PICRUSt2, Tax4Fun2* and *MetGEM* did not show much improvement, while the performance of *PanFP* was raised to the level of the top performing tool *PICRUSt2*. Since correlation analysis is not suited to robustly assess the performance of existing methods, alternative measures are needed.

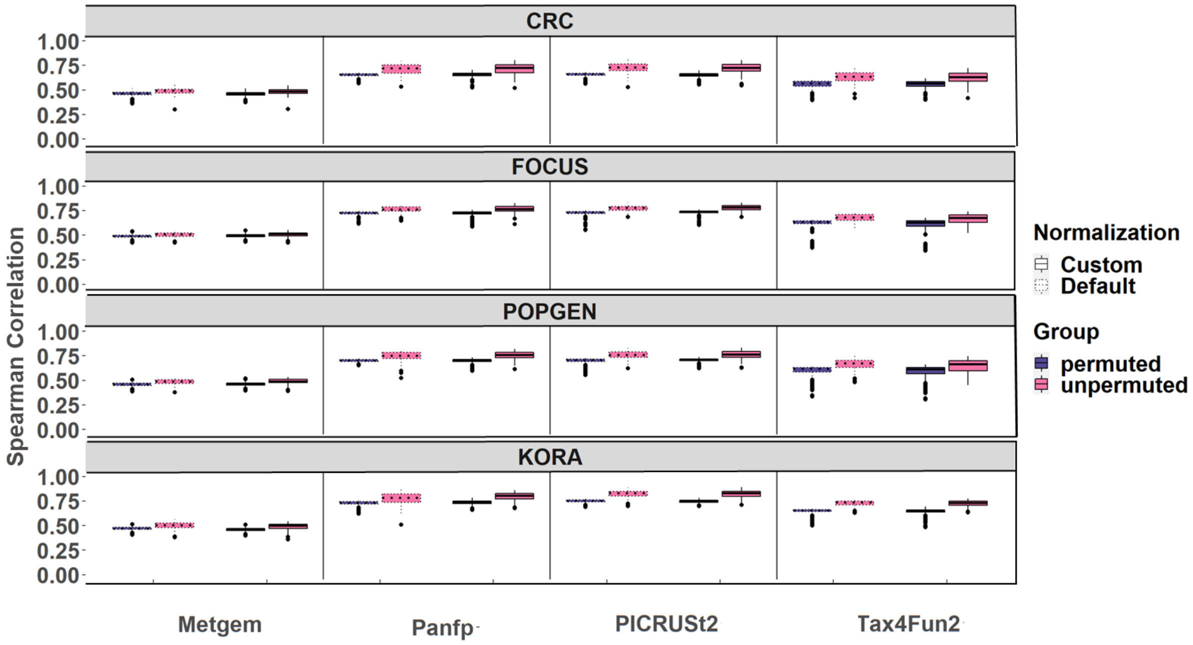


Figure 20: Spearman correlations plot between metagenome predictions and shotgun metagenome sequencing in unpermuted and permuted datasets. The analyses were conducted on both unpermuted and permuted datasets. To validate the functional prediction tools, the metagenome prediction performance was compared against the gold-standard shotgun MGS. The gene composition was estimated from the metagenome sequencing and predicted using various tools such as PICRUSt2, Tax4Fun, PanFP, and MetGEM. The analyses were conducted with default and customized normalization methods on unpermuted (blue) and permuted (red) data in all datasets. In each of the 100 permutations, the abundance of each gene was independently permuted across samples. (Source own work).

3.5.2. Metagenome prediction tools except MetGEM show high specificity in predicting KO terms

Alternatively, tool performance can also be assessed by considering which of the predicted KO terms is also identified in MGS. We would assume that KO terms that are uniquely identified by prediction tools represent false positives, whereas missing KO terms that are reported by MGS represent false negatives, allowing us to compute precision, F1, sensitivity and specificity. Overall, *Tax4Fun2, PICRUSt2* and *PanFP* showed similar performance in terms of F1, accuracy and recall (sensitivity) in contrast to *MetGEM* which showed poor performance (**Fig. 21).** However, *MetGEM* showed high specificity compared to other tools at the cost of low recall.

3.5.3. Metagenome prediction tools show low accuracy in predicting differentially abundant KO terms

It is possible to compare the sets of significant KO words identified by different tools and assess the level of overlap and consistency of the results. Multiple tools may identify the same set KO terms. This indicates that the terms are likely to be associated with the biological processes of the samples. The predictions can therefore be more accurate. If there is a lack in overlap between significant KO terms identified using different tools, it could be an indication that the predictions are not consistent. Further investigation may be necessary to determine the biological processes and pathways involved. The overlap of significant KO terms among different functional inference tools or MGS results can serve as a valuable indicator of the accuracy and provide valuable insight into the underlying biology and processes of the samples.

Analyzing several cohorts that used different tools to predict microbial communities functionally, it was discovered that *PanFP*, *Tax4Fun2*, *PICRUSt2* and *PanFP* had different degrees of overlap with the MGS results. The CRC cohort shows the greatest overlap of significant KO terms with MGS results, followed closely by *Tax4Fun2* (**Fig. 22**) and *PanFP*. Other cohorts, including KORA and FoCus, saw significant KO term overlaps drop significantly. The KORA cohort has *Tax4Fun2* having the highest overlap of significant KO term (n=112) followed by *PICRUSt2* and custom normalization (n=108), and *PICRUSt2*(n=66). *PanFP* with custom normalization (n=94) had a higher level of overlap than *PanFP* without default normalization. (n=13). The overlaps in the Popgen and FoCus cohorts decreased significantly as only a few overlapping terms of KO were found. The overlap of significant terms in KO terms was found to be high in *Tax4Fun2* and *PICRUSt2* using custom normalization. It was interesting to note that *PanFP* and *MetGEM* had poor overlaps of KO terms among all cohorts. It is worth noting the low overlap of KO terms between *MetGEM* results and MGS across all cohorts. This suggests that these tools may be less reliable in functional prediction of microbial communities.

In the CRC cohort, *PICRUSt2* shows the largest overlap of significant KO terms (n=40) with MGS results. In the FoCuS cohort, *Tax4Fun2* has the largest overlap of significant KO terms (n=41) followed by *Tax4Fun2* with custom normalization (n=27), *PICRUSt2* with custom normalization (n=26) and *PICRUSt2* (n=23). In the Popgen cohort, *PICRUSt2* and *PanFP* with customized normalization showed a comparatively large overlap of significant KO terms (n=21) followed by *Tax4Fun2, PICRUSt2* and *Tax4Fun2* with customized normalization (n=13, 12, 11). In the KO termsRA cohort, the largest overlap of significant KO terms is reported in *Tax4Fun2* followed by *Tax4Fun2* with customized normalization, *PICRUSt2*, *PICRUSt2* with customized normalization and *PanFP* with customized normalization. *PanFP* and *MetGEM* showed very poor overlap of KO terms across all cohorts.

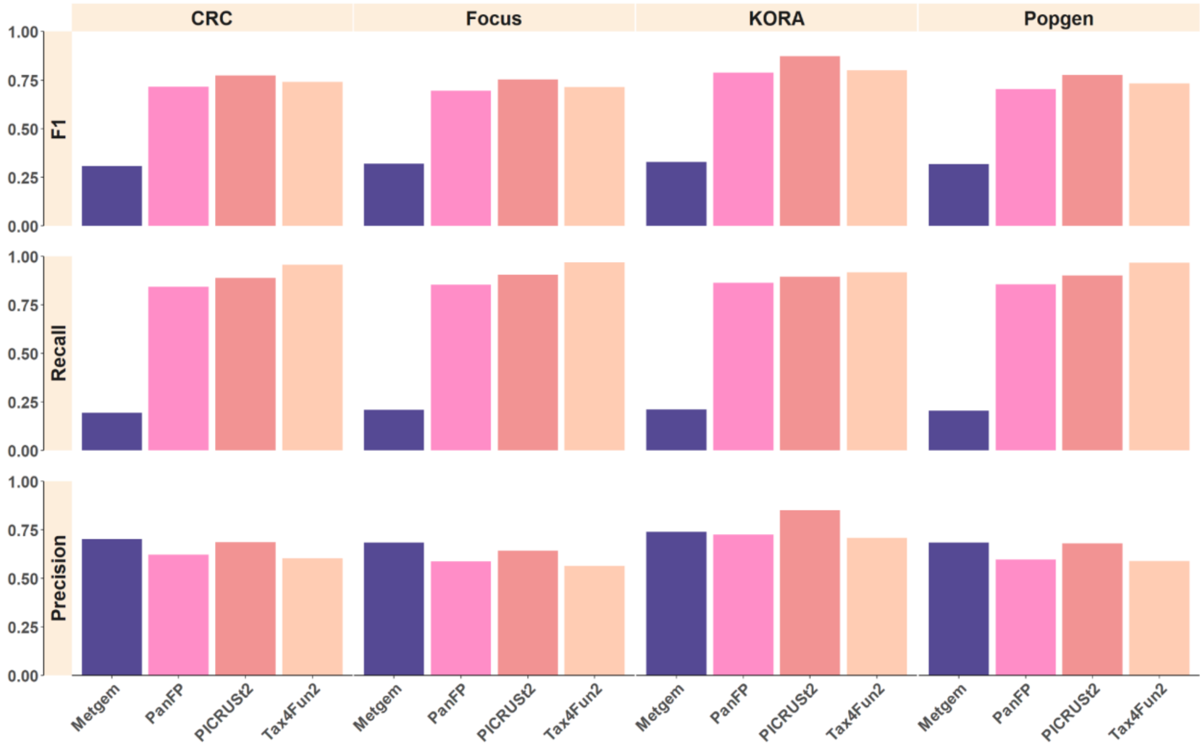


Figure 21: Comparison of detected KEGG orthologs between predicted metagenomes and MGS. Precision, F1, and Recall are reported for each category compared to the MGS data. PanFP, PICRUSt2 and Tax4Fun2 show comparable and relatively consistent performance across data sets while MetGEM shows poor recall. (Source: own work)

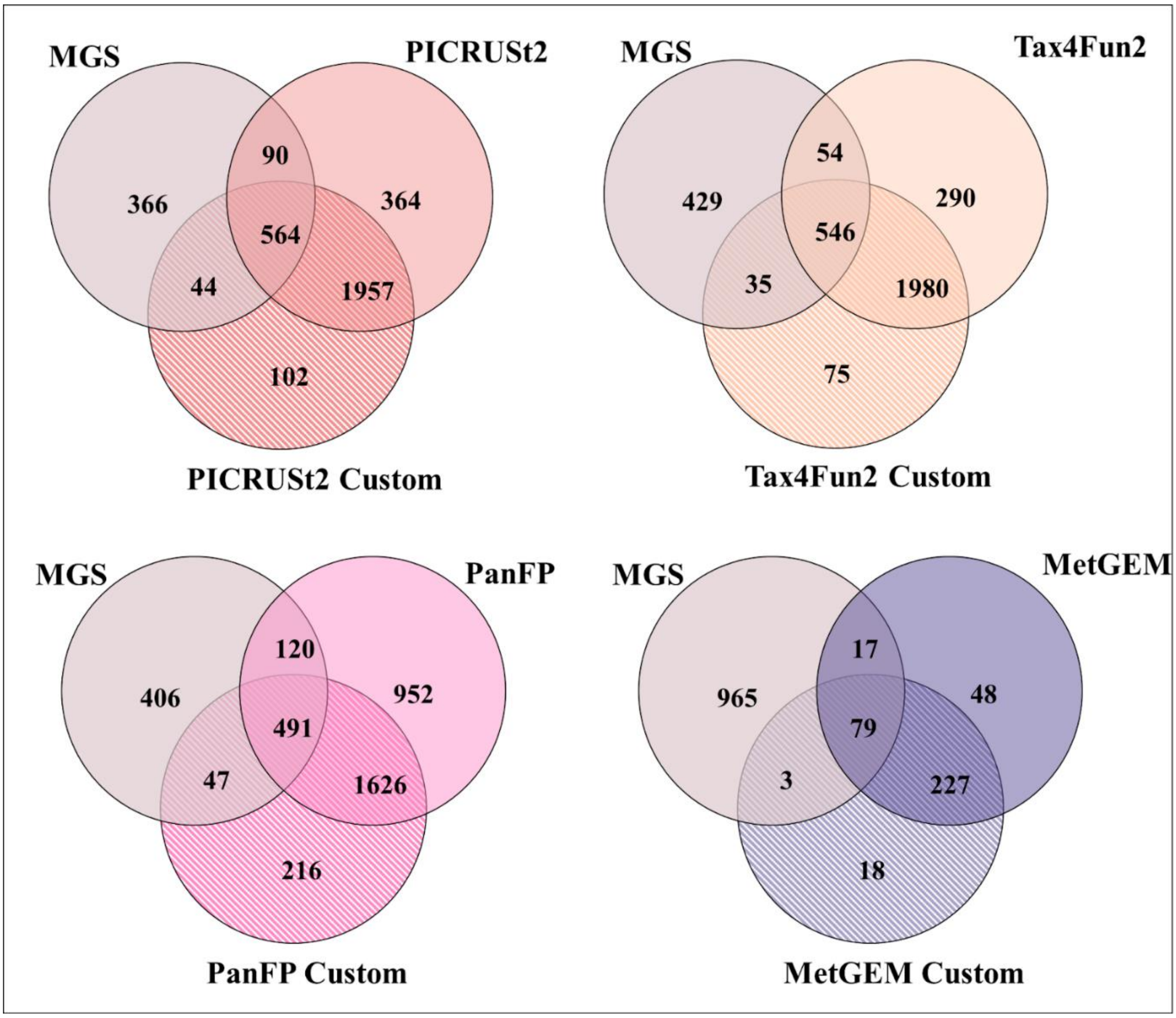


Figure 22: The overlap of significant KO terms between different functional inferred tools and MGS results in the CRC cohort. Overlapping significant KO terms can be used as an indicator of the accuracy of the predictions quantitatively (Source: own work).

Next, we examined the ability of metagenomic data to detect differences in abundance between the predicted genomic content **(Fig 23),** and used evaluation metrics like F1 score, recall, and precision. We used the CRC dataset to compare the abundance of KO terms in cancer patients and healthy patients. This was done using both functional profiling from metagenome and inferred functions. All inferred tools had a low F1 score. While there were no significant differences in F1 scores between *PICRUSt2*, Tax4Fun, and *PanFP* for both normalization techniques, *MetGEM* had a slightly lower F1 score. We found that all cohorts except CRC had very few overlapping significant KO terms. This was also evident in F1 score and recall scores, as well as precision scores. All inferred tools had low precision, which suggests that they were making false positive predictions. This means that the methods were correctly predicting certain features that weren't actually present in the metagenome. This could be due either to limitations in the prediction tools or problems with the quality data used to train and test the models.

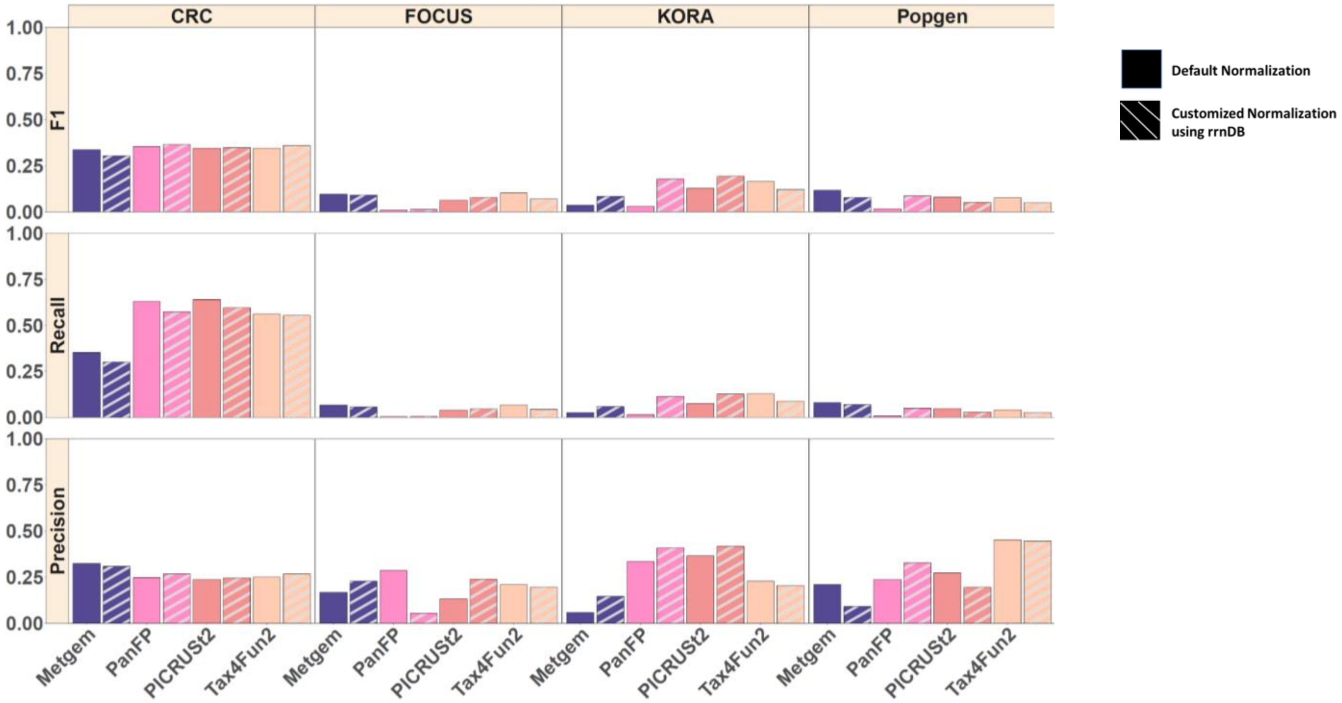


Figure 23: Comparison of significantly differentially abundant KO terms between predicted metagenomes and MGS. F1, recall and Precision scores are reported for each category compared to the MGS data. Precision corresponds to the proportion of significant KOs for that category also significant in the MGS data. Recall corresponds to the proportion of significant KO terms in the MGS data also significant for that category. The F1 score is the harmonic mean of these metrics. (Source: own work)

3.5.4. Performance of functional inference tools at the level 1 KEGG functional categories

We calculated the correlation between the Pearson correlations between the p values that were generated by testing null hypotheses for no association with the metadata for the WGS-16S paired datasets for each cohort using Wilcoxon rank Sum test to analyze their consistency. To do this, we first extracted p-values from each KO term. Next, we grouped KO terms into KEGG functional categories. Finally, we calculated Spearman correlations between the p-values derived from differential analysis in metagenome predictions and MGS results. Because level 3 and 4 functional categories have different numbers of KO terms, we decided to only perform correlation analysis on level 2. This reduces potential biases due to variations in the number KO terms within different categories and allows for a more robust analysis.

3.5.5. Metagenome prediction performance varies widely across functional categories

Our results highlighted the poor performance of prediction tools when we looked into overall differentially abundant KO terms. We next investigated if prediction performance varies across functional categories, suspecting that some categories may be easier to predict than others. We collapsed KO terms into level one KEGG functional categories and calculated the Spearman correlation between p-values obtained from metagenome prediction and MGS results **(Fig. 24)**. We decided to perform correlation analysis only on level one functional categories as level two and level three functional categories vary considerably in the number of KO terms which may affect the correlations. We observed poor correlation between p-values of MGS and prediction tools, where *PanFP* and *PICRUSt2* with customised normalisation showed improved results which indicates an advantage of 16S copy number normalisation using rrnDB in this scenario. As expected, *MetGEM* showed very few significant correlations for all types of KEGG functional categories. Most of the genes predicted by metagenome prediction tools but not detected by metagenome sequencing belong to the biosynthesis of secondary metabolism, metabolism of terpenoids and polyketides, and xenobiotics biodegradation and metabolism.

We compared the correlation of the p-values derived from metagenome with those derived from inferred tool for different functional categories within the KEGG databank. This result showed poor correlation and no pattern in inferred tools' performance towards specific functional categories across all cohorts. It was found that *PanFP*, *Tax4Fun2*, *PICRUSt2*, and *Tax4Fun2* performed better than *MetGEM* in CRC and KORA cohorts.This is mainly due to the positive correlation between the p values obtained from metagenomes compared to those from inferred instruments. Notably, *MetGEM* had a negative correlation with human disease. This could indicate that the tool is not as efficient in identifying genes and functional pathways that are related to human illnesses. The overall performance of these tools dropped even further in the Popgen and FoCuS cohorts. Negative correlations were seen in *PICRUSt2*, *Tax4Fun2* as well as *PanFP* for genetic information processing.

MetGEM also revealed a high correlation in the processing and genetic information KEGG categories. The low number of predicted KO terms in *MetGEM*' might have influenced the overall correlation, compared to other tools that have high KO terms.

All tools identified more KO terms within the Metbolism functional categorie followed by B09180 Brite hierarchies. These hierarchies are not included in pathway, brite, or 09130 environmental Information Processing. 09120 Genetic Information Processing. 09140 Cellular processes. 09160 human diseases. 09150 organismal system. Metabolism functional categories contained more KO terms than *MetGEM* detection. This indicates that the tool is more effective at identifying genes and functional pathways associated with this category. Prediction of low KO terms within a particular KEGG group may indicate that the tool has less success in identifying genes and functional pathways. A low number of predicted KO terms could limit your ability to draw meaningful conclusions, identify potential targets, or identify biomarkers.

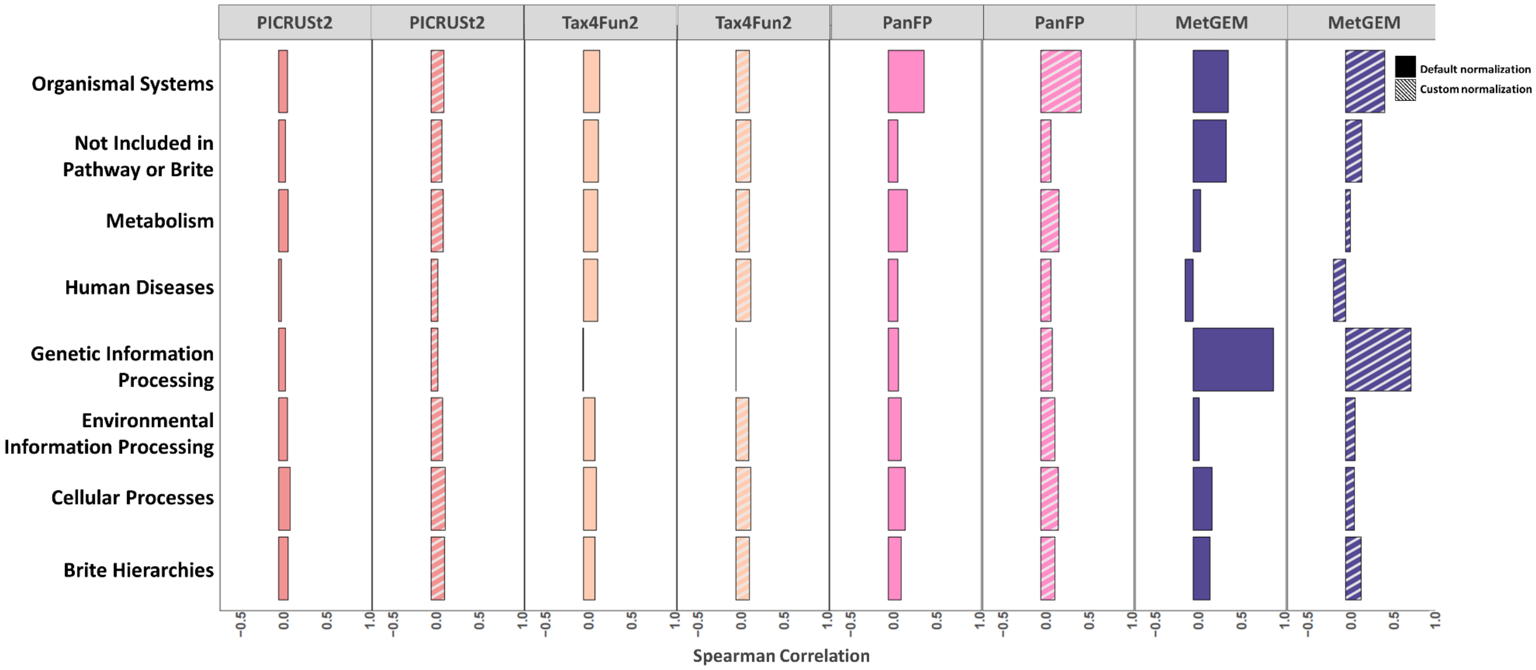


Figure 24: Spearman correlations between p-values obtained from prediction tools and metagenome sequencing for level 1 KEGG functional categories including cellular process, genetic information processing for CRC cohort. Negative correlations are not shown. (Source: own work)

We decided to concentrate on disease-specific KO terms as they are better suited for evaluating the performance of inferred instruments, since level 1 KEGG functional categories could include terms that are too generalized. We focused on KO terms that are classified under pathways in cancer and biosynthesis and metabolism glycans, lipid metabolic, for the CRC cohort. (**Fig 25 (a)**). These KO terms have been reported to be enhanced in CRC patients. In order to see if the same KO term can be predicted in an inferred tool, we selected significant KO terms (p value 0.05). This was useful in comparing the performance of inferred tools when predicting disease-specific terms for KO terms within the context of CRC. *PanFP* was the most significant overlap with MGS in the category of biosynthesis and metabolism. *Tax4Fun2* and *PICRUSt2* followed (default and customized normalization). There were only three significant KO terms that overlapped between MGS, *MetGEM*. The study also compared KO terms in lipid metabolic and found that *PICRUSt2* was the most overlapped with MGS. *Tax4Fun2* followed closely by *PanFP* (default/custom normalization). (**Fig. 25(b)).**

We focused on KO terms that are classified under pathways of carbohydrate metabolism (**Fig 26 6 (a&b)**). Similar to CRC we compared significant KO Terms (p-value 0.05), obtained from MGS results with the KO terms derived from inferred instruments. Only *PICRUSt2* with default and custom normalization showed significant overlaps in KO terms for the pathway of carbohydrate metabolic. This suggests that other inferred tools might not be able to predict disease-specific KO terms in relation to carbohydrate metabolism within the context of the KORA cohort. In the pathway of amino acids metabolism, *Tax4Fun2* with default and custom normalization showed very few significant KO terms that overlapped with MGS. This indicates that other inferred tools might not be able to predict disease-specific KO terms in relation to amino acid metabolism within the context of the KORA cohort.

The FoCus cohort studies on healthy vs obese have focused our attention on KEGG functional categories like carbohydrate metabolism and amino acid metabolic pathways. These categories have been found to be enriched in obesity (**Fig 27, 28**). The overall number of KO terms in both cohorts fell signifanylt between metagenomes and inferred instruments. Tax4Fun and PICRUSt2 only shared a few significant terms with metagenome. Additional files provided overlap information for other KEGG types. The results indicate that inferred tools may not be able to predict disease-specific KO terms for certain metabolic pathways in the KORA cohort. The interpretation of the results could depend on the MGS data and the specific tools used. Additional validation or improvement may be required for these tools. The interpretation of the results can depend on the cohort, the MGS functional profiles and the nature of inferred tools used.

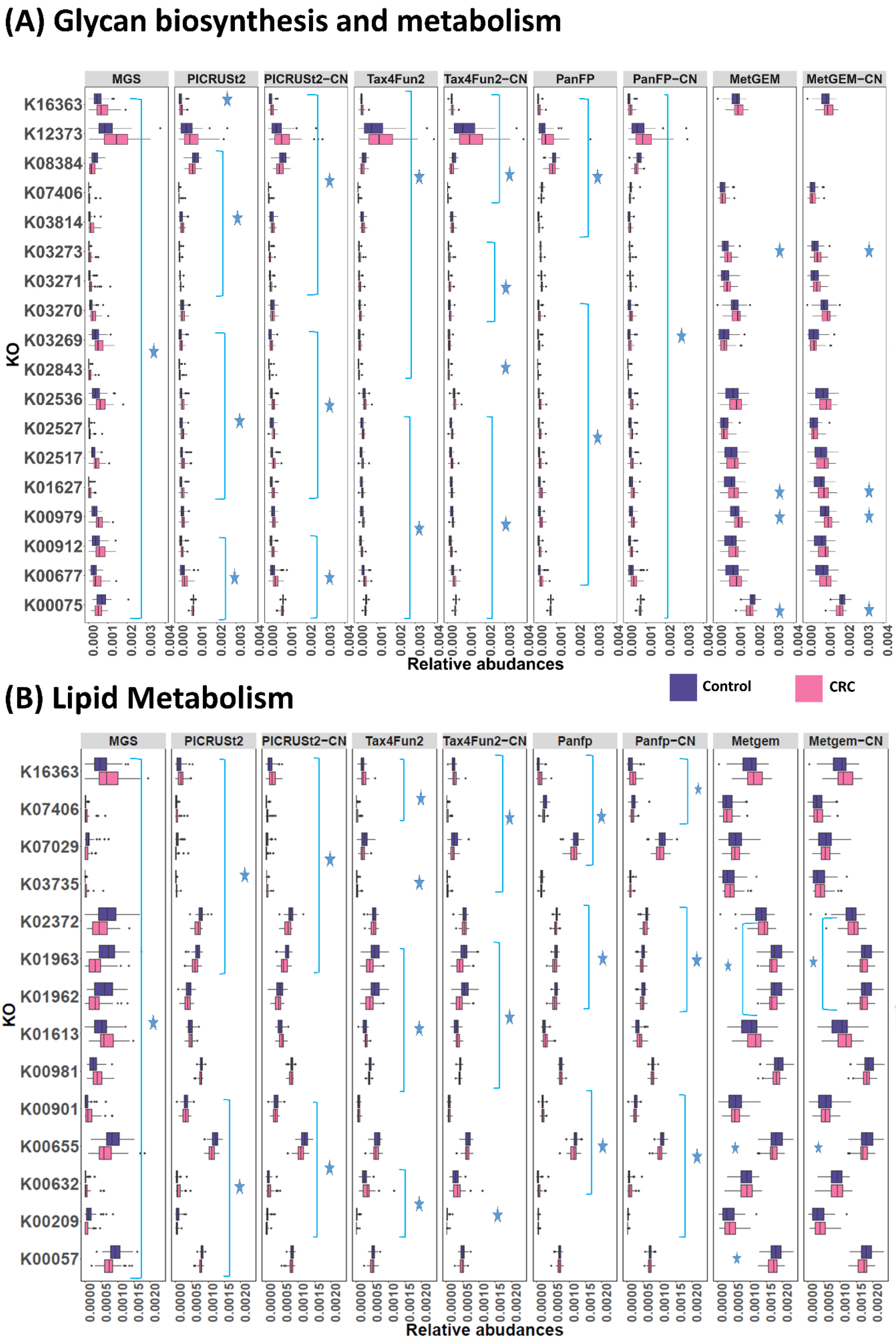


Figure 25: Comparison of relative abundance distributions among major KEGG gene categories such as (a) glycan biosynthesis metabolism and (b) lipid metabolism between

inferred tools and MGS functional profiles in CRC cohort. Relative abundance data for the two sets of profiles (inferred tools and MGS functional profiles) for the KEGG gene categories of interest and distributions of the relative abundance data for each set of profiles was plotted using a boxplot. Wilcoxon rank-sum test was performed to compare the healthy and diabetes. The null hypothesis was set that there was no significant difference in the relative abundance distributions between healthy and colorectal cancer patients. P-value of the Wilcoxon rank-sum test is less than 0.05 indicated a significant difference in the relative abundance distributions between the two sets of profiles. The significance level was displayed with an asterisk (*). (Source: own work)

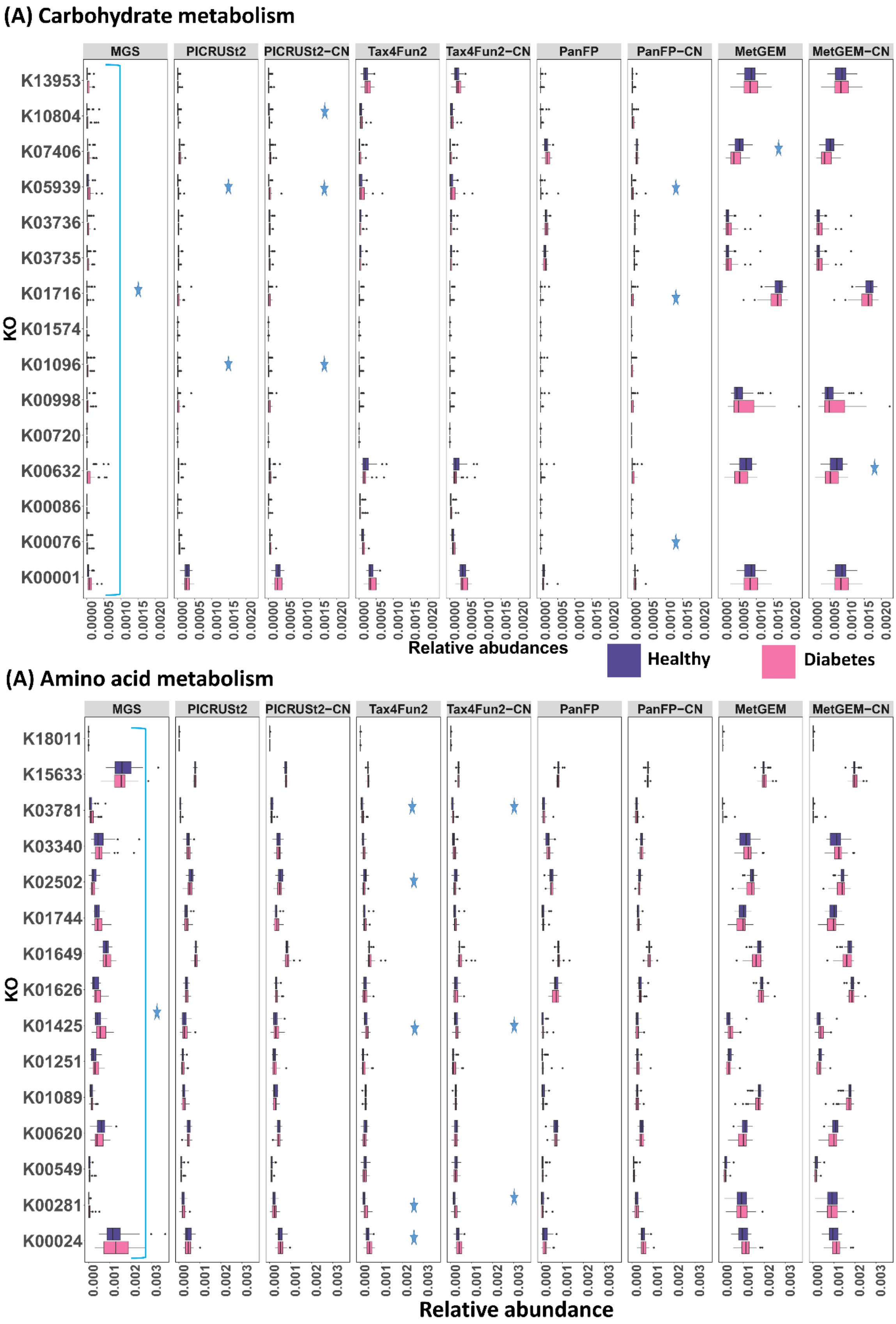


Figure 26: Comparison of relative abundance distributions among major KEGG gene categories such as (a) carbohydrate metabolism and (b) amino acid metabolism between

inferred tools and MGS functional profiles in the KORA cohort. Relative abundance data for the two sets of profiles (inferred tools and MGS functional profiles) for the KEGG gene categories of interest and distributions of the relative abundance data for each set of profiles was plotted using a boxplot. Wilcoxon rank-sum test was performed to compare the healthy and diabetes. The null hypothesis was set that there was no significant difference in the relative abundance distributions between healthy and diabetes. P-value of the Wilcoxon rank-sum test is less than 0.05 indicated a significant difference in the relative abundance distributions between the two sets of profiles. The significance level was displayed with an asterisk (*).(Source: own work)

Since the FoCus and Popgen cohort studies about healthy vs obesity, we focused on KEGG functional categories such as carbohydrate metabolism, amino acid metabolic pathways and sugar transport which have been reported as enriched in obesity [358,359] (**Fig. 24 and 25).** In both cohorts, the overall number of KO terms between metagenome and inferred tools dropped significanylt. *PICRUSt2* and Tax4Fun only shared a couple of significant terms with metagenome. Overall, the results suggest that the inferred tools may have limited ability to predict disease-specific KO terms in certain metabolic pathways for the KORA cohort. It is important to note that the interpretation of the results may depend on the specific inferred tools and MGS data used in the analysis, and additional validation and improvement may be needed for the inferred tools. It is important to note that the interpretation of the results may depend on the specific cohort and the nature of the inferred tools and MGS functional profiles being compared. Altogether, prediction tools failed to detect health-related differential abundance measures of functional categories as compared to MGS-derived results across multiple cohorts.

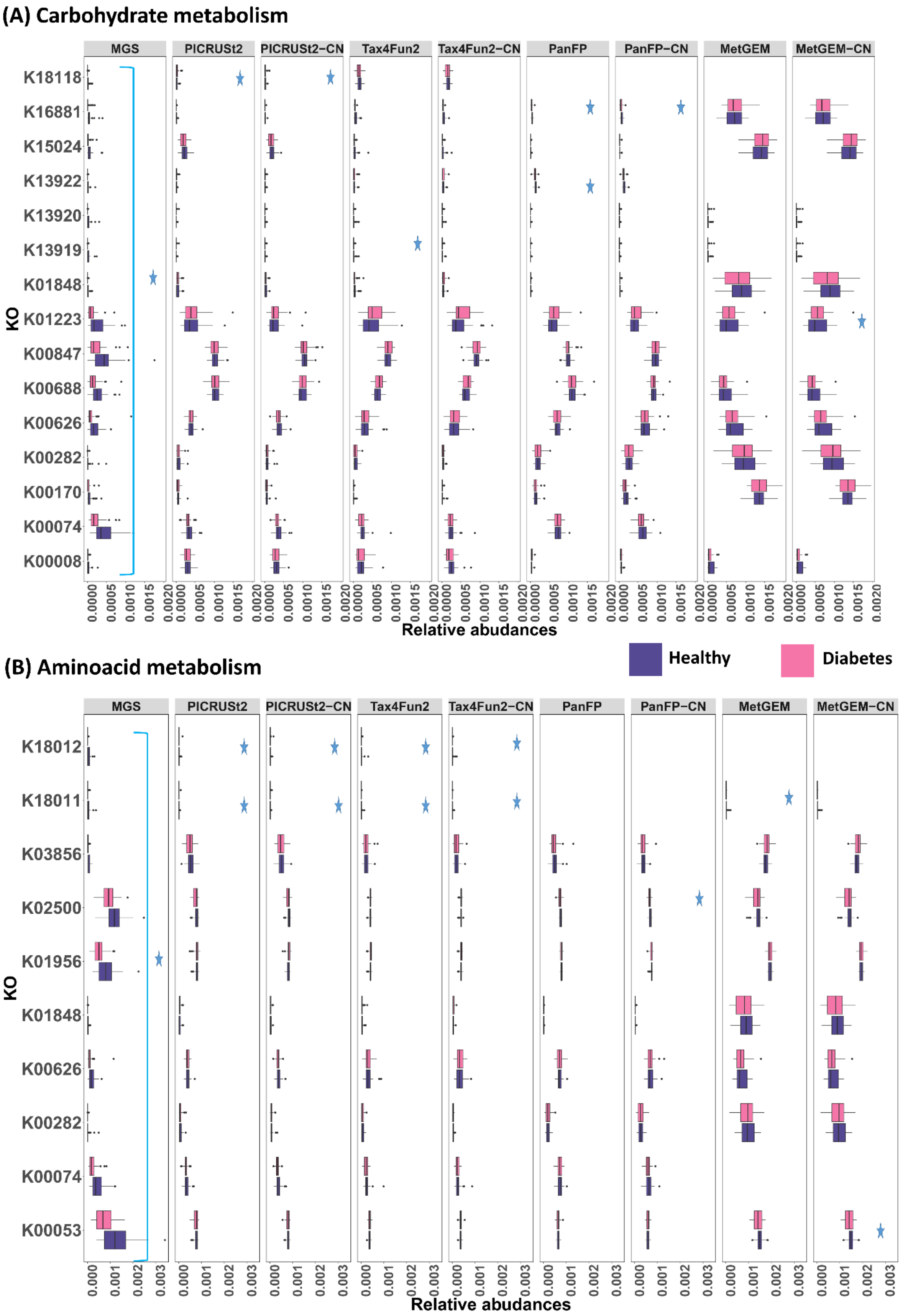


Figure 27: Comparison of relative abundance distributions among major KEGG gene categories such as (a) carbohydrate metabolism and (b) amino acid metabolism between

inferred tools and MGS functional profiles in the Popgen cohort. Relative abundance data for the two sets of profiles (inferred tools and MGS functional profiles) for the KEGG gene categories of interest (e.g., Carbohydrate metabolism and Amino acid metabolism) and distributions of the relative abundance data for each set of profiles was plotted using a boxplot. Wilcoxon rank-sum test was performed to compare the healthy and obese. The null hypothesis was set that there was no significant difference in the relative abundance distributions between health and obese. p-value of the Wilcoxon rank-sum test is less than 0.05 indicated a significant difference in the relative abundance distributions between the two sets of profiles. The significance level was displayed with an asterisk (*). (Source: own work)

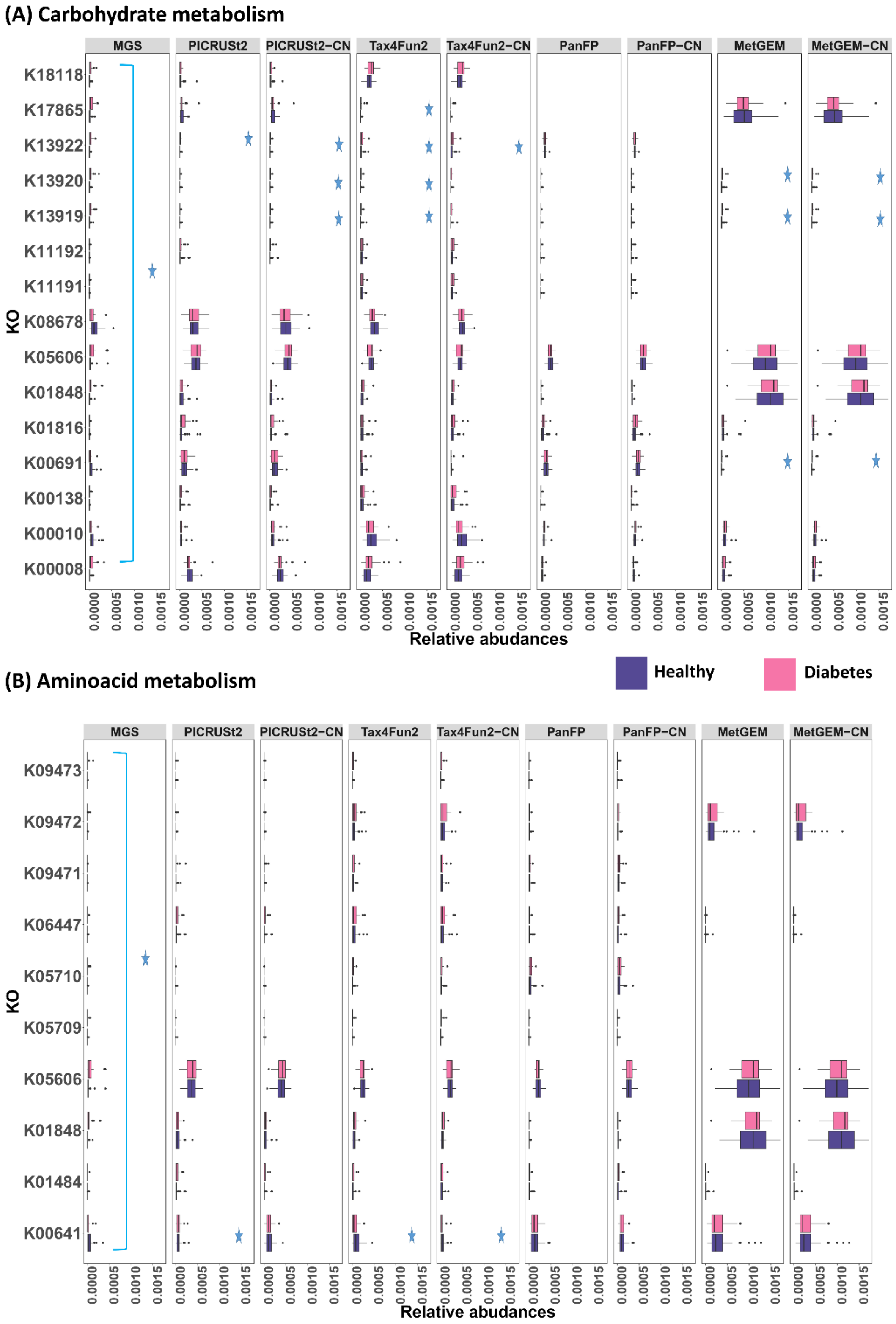


Figure 28: Comparison of relative abundance distributions among major KEGG gene categories such as (a) carbohydrate metabolism and (b) amino acid metabolism between inferred tools and MGS functional profiles in the FoCus cohort. Relative abundance data for the two sets of profiles (inferred tools and MGS functional profiles) for the KEGG gene categories of interest (e.g., Carbohydrate metabolism and Amino acid metabolism) and

distributions of the relative abundance data for each set of profiles was plotted using a boxplot. Wilcoxon rank-sum test was performed to compare the healthy and obese. The null hypothesis was set that there was no significant difference in the relative abundance distributions between health and obese. p-value of the Wilcoxon rank-sum test is less than 0.05 indicated a significant difference in the relative abundance distributions between the two sets of profiles. The significance level was displayed with an asterisk (*). (Source: own work)

3.5.6. Functional Profiling on simulation datasets

A possible explanation for the variation between predicted and MGS-derived functional profiles are technical issues. For example, due to the amplification bias induced in 16S rRNA gene polymerase chain reaction (PCR) [55,327,360] and functional profile variation among phylogenetically related genomes, microbiome functional profiles predicted from 16S amplicons can deviate greatly from MGS-derived ones. We decided to test the influence of technical variabilities using simulated MGS from the 2nd CAMI Challenge [351] which represents typical microbiomes from four human body sites. We obtained matching 16S rRNA gene profiles by filtering for 16S rRNA gene reads matching to the SILVA [101] database. Since no difference in sample processing is involved here, we expect this setup to show how close metagenome prediction tools can be expected to approximate MGS-derived functional profiles under optimal conditions. PCA plot (**Fig. 29)** revealed that *PICRUSt2* and *PanFP* using default and customized normalization clustered most closely to the MGS profiles. In contrast, *Tax4Fun2* and *MetGEM* showed a high discrepancy from their corresponding metagenome.

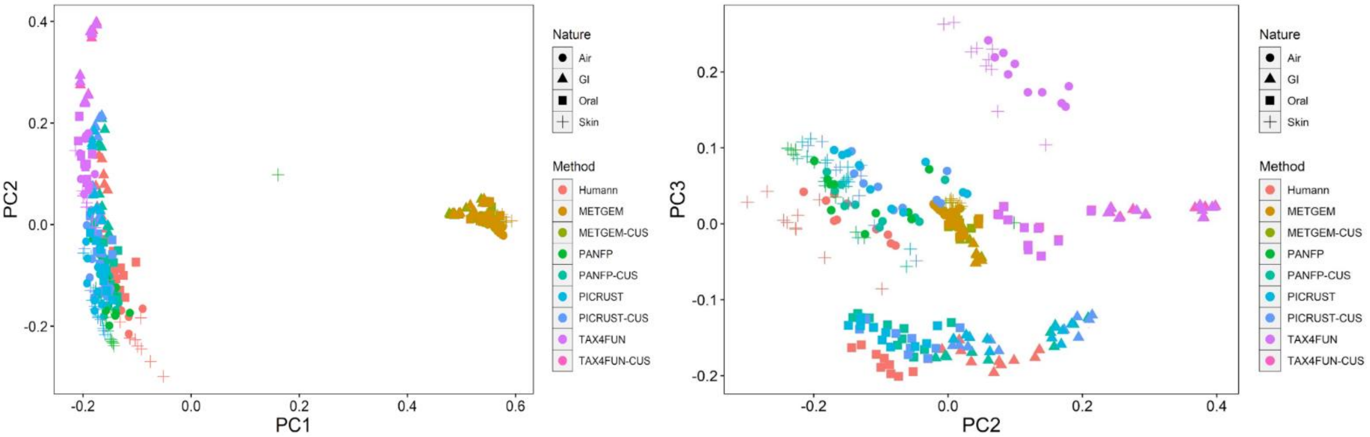


Figure 29:(a) Functional beta diversity of the simulated metagenome-amplicon sample pairs between PC1 and PC2. (b) Functional beta diversity of the simulated metagenome-amplicon sample pairs between PC2 and PC3.. (Source: own work)

We also conducted differential abundance tests to compare how results differ between predicted metagenomes and actual MGS simulation dataset. This analysis was done between GI and skin as they are clustered separately in the diversity plot such that we expected to find considerable variation in the functional profiles. The overlap between the significant KO terms of *PICRUSt2* 47% (default and custom-normalization) and MGS. Similarly, the overall percentage was also increased for other tools such as *PanFP*: 35%, *PanFP* with custom normalization: 45%, *Tax4Fun2*: 35%, *Tax4Fun2* with custom normalization 39%, *MetGEM*: 30%, *MetGEM* with with custom normalization 28%. Performance metrics such as F1, recall and precision were also compared to public datasets. Overall, the performance of simulation datasets were improved compared to the real datasets as shown previously **(Fig. 30)**

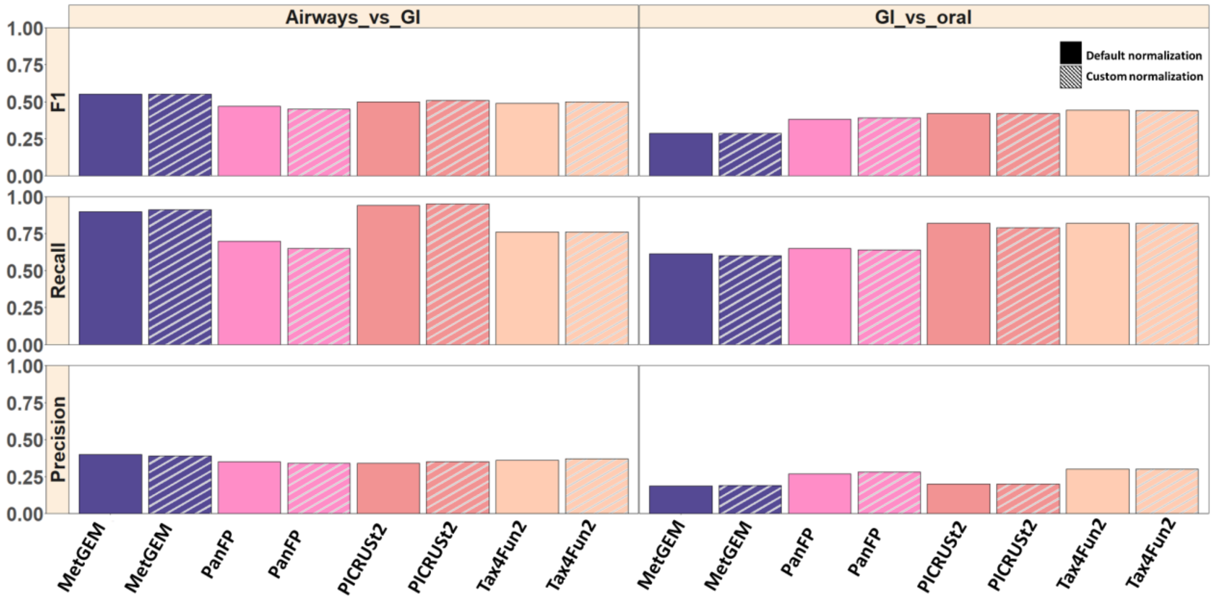


Figure 30: Pairwise differential analysis was performed between simulated GI vs Oral and Skin vs airway. Significant KO terms were identified using Wilcoxon rank-sum test with p-value < 0.05. Accuracy measurement terms among MGS, PICRUSt2 with default and custom-normalization. (Source: own work)

3.6. Discussion

In the subject of microbial ecology, knowing the roles of a microbiome allows a more comprehensive understanding of the biological processes in which it may be involved. Apart from metabolomics experiments, metagenome approaches coupled with functional enrichment analysis is the method of choice for inferring functional relationships within the microbiome and between microbiomes and their ecosystem [284,288]. However, due to the significantly lower costs, 16S rRNA gene profiling is still the choice of many researchers for studying microbial abundances [221,361,362]. Popular tools such as *PICRUSt2* or *Tax4Fun2* were developed to predict microbial functions from 16S rRNA gene amplicon sequencing datasets. They achieve this by utilizing the knowledge from large reference genome databases such as KEGG functional profiles together with ancestral state reconstruction methods [157–159]. For example, *PICRUSt2* uses an HSP approach to incorporate as many sequences as is practically possible into its prediction, whereas *Tax4Fun2* chooses to only use sequences that fall within a similarity cutoff of reference sequences. Importantly, the predictions that are produced by either tool need to be evaluated with extreme caution because taxonomic and phylogenetic certainty can only be as reliable as the curation of the databases that are employed [363,364]. Each tool follows a different algorithmic approach and relies on different references, leading to a considerable discrepancy in their predictions. As these tools can only provide functional predictions, they are not a suitable replacement for shotgun metagenomic research and should not be used as such.

Using matched MGS and 16S rRNA gene data sets, we were able to assess the prediction accuracy of *PICRUSt2, Tax4Fun2, PanFP* and *MetGEM* in a human disease setting. All metagenome prediction tools except *MetGEM* showed overall good accuracy in predicting KO terms that were also found in MGS. *PICRUSt2* was slightly better at representing gene distribution profiles than the other tools, possibly because a larger number of sequences are included in the hidden state prediction, which in turn increases the sensitivity to detect ubiquitous functions [327] that are common for the majority of microorganisms in a sample. On the other hand, *MetGEM* showed the lowest precision. One reason might be that the reference AGORA collections contain 818 GEMs from human gut microbiome which covered only 1,470 KO terms identifiers, 983 EC numbers across 226 genera and 690 species in total [162,365].

We hypothesise that the good performance is mostly driven by predicting KO terms that are typically present in a sample. This hypothesis is confirmed by the high Spearman correlation between predicted functional profiles and metagenome sequencing even after label randomization as originally shown by Sun *et al.* [165]. In this study, we focused on differential abundance of functional terms as a more challenging scenario meant to reveal if functional prediction methods can pick up biologically meaningful differences. Differential abundance testing of KO terms was performed using the Wilcoxon rank sum test at different KEGG levels. First, we performed correlation of p values from MGS-derived level 1 KEGG functional category and predicted p-values. Functional inferred tools can be affected by the unique features and complexity of biological pathways and processes. Metabolic pathways, for example, involve complex networks of reactions and regulatory mechanisms. These may differ from the signal transduction pathways. Functional inferred tools that are good at predicting functions within one category might not be as effective in another. Researchers can evaluate the performance of functionally inferred tools in different KEGG hierarchical groups to identify strengths and weaknesses. This information can be used to guide the selection and interpretation of functional genomics data. The KEGG hierarchical classifications provide a standardized classification system of gene functions that allows researchers to compare the results of different functional inferred instruments in a consistent way. This will facilitate the improvement and development of functional inferred instruments and aid in the advancement of functional genomics research. In this situation, metagenome inferred tool missed many genes that were predicted by metagenome. It also predicted many genes that are not differentially abundant in MGS data. We performed differential abundance testing of KO terms using the Wilcoxon test at different stages. As a first step, correlation of p-values from predicted and MGS-derived were level 1 KEGG functional categories was performed. In this setting, metagenome prediction tools missed a large set of genes that are predicted by metagenome and likewise predicted many genes that are actually not found to be differentially abundant in MGS data. Despite these poor results, we could show that *PanFP* and *PICRUSt2* with customized normalization showed improved results, indicating an advantage of 16S copy number normalization using rrnDB in this scenario. As expected, *MetGEM* showed very few significant correlations for all types of KEGG functional categories.

We hypothesize that the possible reasons for the huge variation between predicted and MGS-derived functional profiles are technical issues and platform-related differences introduced e.g. by PCR bias. To study this, we used a simulated dataset to compare the performance between prediction tools and MGS in the absence of technical confounders. As expected, this analysis resulted in improved performance. Since we considered the full length 16S rRNA gene in the simulation study, one can suspect that higher coverage of multiple variable regions can further improve the quality of predictions. Another reason for the good performance in the simulation experiment is that samples were obtained from two different environments, where the differences are easier to detect compared to a health *vs.* disease scenario.

3.7. Conclusion and outlook

The functional potential of microbial communities can be inferred from 16S rRNA gene sequencing. This method is based on the taxonomic composition of the communities. This approach is commonly used when whole-genome shotgun sequencing is not feasible, as 16S rRNA gene sequencing is less expensive and less computationally demanding. Functional profiling from 16S rRNA gene sequencing is based on the concept of phylogenetic conservation, which suggests that closely related microorganisms have similar metabolic capabilities. This approach typically involves mapping the taxonomic information obtained from 16S rRNA gene sequencing to a reference database of microbial genomes or metagenomes with known functional annotations. Different bioinformatic tools and databases are available for functional profiling, such as *PICRUSt2*, *Tax4Fun2*, *PanFP* and *MetGEM*. These tools use different algorithms to predict the functional capabilities of the microbial community based on the taxonomic composition.

A clear finding of this study is that functional predictions from 16S data are generally not sensitive enough to pick up differences related to changes in human health in typical settings such as CRC, obesity or type-2 diabetes. Limitations in the performance of functional prediction tools can be explained by an incomplete reference genome [363,364,366]. While functional prediction tools pick up major differences, e.g., between ecological niches, they should not be used as a replacement for MGS in the study of human health. If researchers intend to produce functional predictions from 16S rRNA gene data for hypothesis generation they should be aware of these limitations and implement control strategies such as sample label randomization. Among the available tools, we recommend using *PICRUSt2* and *Tax4Fun2* which appeared to be most robust, followed by *PanFP* which could be improved by introducing a customised copy number database.

Chapter four: Namco, a Free microbiome explorer

4. Chapter four: Namco, a Free microbiome explorer

4.1. Declaration of contribution

This chapter is the result of a project started in the Collaborative Research Centre (CRC) 1371 microbiome signature as an internal pipeline to analyse 16S rRNA gene sequencing data under the guidance of Dr. Markus List, Big Data in Biomedicine Group, Technical University of Munich (TUM). Later, it was published in the open access journal, Microbial Genomics in August 2022 [367]. The work described here has been driven by Alexander Dietrich (as a student assistant of the group, now a doctoral student) and myself in the Big Data in Biomedicine Group with equal contribution.

4.2. Introduction

Over the past decade, microbiome research has gained significant attention due to the growing understanding of the impact of the microbiome on human health. [368,369]. Several studies revealed that biochemical activities of the gut microbiome are one of the main factors for causing human diseases such as diabetes[370], cancer[371], inflammatory bowel disease (IBD)[19], breast cancer [372] and brain disorders[373]. For example, Thomas et al. [374] illustrated that the gut microbiome compromises the integrity of the gastrointestinal barrier in IBD. Thanks to the recent advancement in sequencing techniques and the development of powerful computational tools, researchers are able to conduct studies to unravel the mystery of microbial communities.

As mentioned in previous chapters, microbiome datasets are generated using two common methods: 16S rRNA gene sequencing and shotgun metagenomics sequencing. The former targets the 16S rRNA gene for charactering bacteria and many studies showed that that smallest number of raw reads as low as 18,000 to 20,000 reads per sample is sufficient to provide bacterial taxonomic classification [375]. On the other hand, shotgun metagenomic sequencing focuses on the whole microbial genome and delivers knowledge not only on the taxonomic composition but also on functional profiles which are not retrievable with 16S rRNA gene sequencing [166,167,376,377]. However, 16S rRNA gene amplicon sequencing is still the most widely used method due to the low costs [298,378,379].

The entire workflow of 16S rRNA gene data analysis falls into four important categories. The first and foremost step is to identify taxonomic composition either by OTU[94] clustering approaches or by denoising approaches [96]. Several benchmark studies proved that denoising approaches provide higher resolution and accuracy compared to traditional OTU clustering methods [380,98]. Along with taxonomic composition, microbial diversity analysis including alpha and beta diversity help in studying the species richness and evenness within a sample and between groups, respectively. The second step is to predict metagenomes and assign different functional groups to those using computational methods such as *PICRUSt2* [157], *Tax4Fun2* [158], *Piphillin* [160] and *PanFP* [159]. The third step is to identify significant differential features/functions between the conditions. And the final step is to study the microbial association through microbial co-occurrence network analyses.

Also, when analysing microbiome data, a flexible pipeline is essential in order to effectively uncover insights and discover patterns. Different types of comparisons may be necessary depending on the research question or hypothesis being tested. For example, comparing the microbiome of healthy individuals to those with a specific disease may require different types of comparisons than comparing the microbiome of different geographical regions or dietary groups.

Additionally, the pipeline should be flexible in terms of different types of data and different types of analysis, such as alpha diversity, beta diversity, and taxonomic or functional profiling. The ability to easily adapt the pipeline to different types of data and comparisons is important in order to effectively answer the research question and draw meaningful conclusions.

4.2.1. Bioinformatics pipelines for 16S rRNA gene analysis

A plethora of computational and statistical tools are being developed to analyse the large microbial data for each part of the analysis (**Fig. 31**). These tools are mostly command-based without any graphical interface. For example, *Mothur* [95] *QIIME2* [91] and *DADA2* [96] offer processing of raw sequencing files through clustering and annotation of 16S rRNA gene and provide OTU or ASV tables in text or biom format. These formats are used as inputs for downstream analysis. The *QIIME2* [91] pipeline consists of more than 20 different plugins for compositional data analysis [144], q2-longitudinal for time-series analysis, [382] and q2-sample-classifier for supervised classification and regression analysis [381].

4.2.2. Statistical analysis of microbiome data with R

Apart from the above-mentioned pipelines, several R packages have been developed, dedicated to microbiome analysis. For example, three main R packages such as *Vegan* [383], *Phyloseq* [113], and *Microbiome* [384] make up the bioinformatics pipeline used to evaluate how the gut microbiome composition is associated with various forms of gastric cancer. The *Vegan* package is developed using R programs and needs to be used in an R statistical context. Additionally, *Vegan* provides resources for multivariate and diversity analyses, among other possibly helpful features. As a result, it is well suited for the analysis of microbiome data and is frequently used to study ecological communities [385].

Microbiome and *Phyloseq* R packages also contain numerous tools and functions for evaluating microbiome profiling data. They can be combined with other statistical software and offer capabilities for analysing microbiome data sets. Additionally to fitting linear models, the *Microbiome* package offers capabilities to investigate microbiota composition and other diversity indices on microbiome data sets. In addition to statistical analysis, these packages include utilities for visualising data as graphs, plots on ordination axes, heatmaps, and other formats [386].

4.2.2.1. Phyloseq

Phyloseq is an R package to import, store, analyse, and graphically display complex phylogenetic sequencing data that has already been clustered into OTUs/ASVs, together with other processed files such as metadata, phylogenetic trees, and/or taxonomic assignments. This package utilizes existing R packages such as *Vegan, ade4, ape, picante* for phylogenetic and ecology analysis. For the publication-quality graphics, It uses as *ggplot2* R package. *Phylose*q utilizes a S4 class system to store different data types related to microbial community analysis as a single experiment level object. This makes it easier to manage and manipulate large datasets and facilitates reproducibility by enabling users to easily store and share their analysis pipelines with others. Overall, *Phyloseq* is an efficient R package, which provides reproducible analysis of microbial data.

4.2.2.2. Rhea

Rhea is a R script-based downstream analysis pipeline and scripts for individual analysis are independent of each other which makes user to run a specific analysis at any time. The pipeline covers a wide range of basic analyses such as diversity analyses, differential abundance analysis and correlation. It is available on the GitHub repository (https://github.com/Lagkouvardos/Rhea[117]).

4.2.2.3. VAMP

Visualisation and Analysis of Microbial Population Structures (*VAMPS* [387]; http://vamps.mbl.edu) is a free web-based application to facilitate research on large-scale microbial sequencing data. *VAMP* was developed by the Microbial Biodiversity Laboratory (MBL). In *VAMP*, users have the ability to upload marker gene sequences along with accompanying metadata. Reads are then subjected to quality filtering before being assigned to taxonomic structures as well as taxonomy-independent clusters. Users are able to select analysis of any combination of their own private data or the private data of their collaborators as well as data from public projects. They can then filter these by their choice of taxonomic and/or abundance criteria, and finally, explore these data using a wide range of analytical methods and visualisations. All of this is done through a user-friendly interface without any programming. Each result feature is connected with extensive hyperlinking to a different analyses and visualisation choices, which encourages data exploration, and ultimately, leads to a deeper comprehension of the ways in which data are related to one another.

4.2.3. Microbiome Analyst

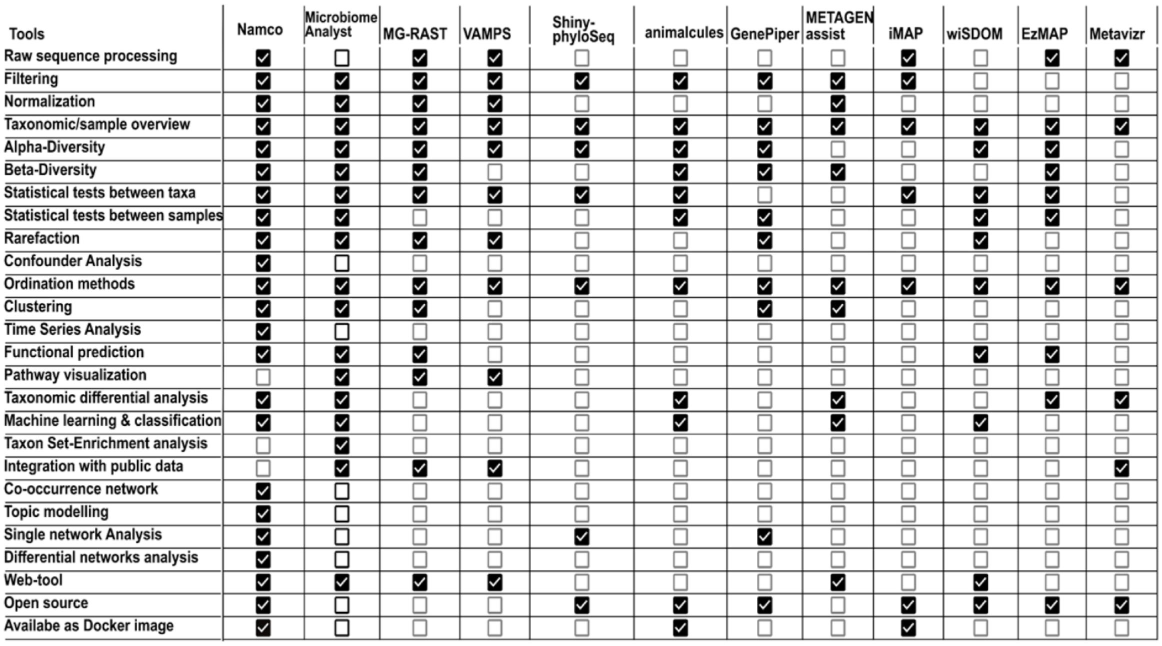
MicrobiomeAnalyst[116] is a web-based platform for comprehensive analysis specifically developed for microbial data analysis. It allows researchers especially clinicians with limited bioinformatics experience to get hands-on experience of methods for processing and analyzing microbiome data, performing functional profiling using 16S rRNA and performing statistical analysis. In addition, *MicrobiomeAnalyst* also allows users to compare their data with public datasets or known microbial signatures. It mainly has four distinct modules: The Marker-gene Data Profiling (MDP) module analyses 16S rRNA gene amplicon sequencing data, whereas the Shotgun Data Profiling (SDP) module analyses metagenomics sequencing data. Users can compare their data with publicly available datasets using the Projection with Public Data (PPD) module, while the Taxon Set Enrichment Analysis (TSEA) module provides tools for detecting enriched taxa in a specific sample group or condition. Overall, *MicrobiomeAnalyst* is a user-friendly web interface that enables researchers to thoroughly explore their preprocessed microbiome data.

4.2.4. Limitation of existing tools

While, several R packages offer a powerful approach to perform microbial data analysis, executing R scripts can be challenging to use without basic bioinformatics training or scripting knowledge. As a result, there is a demand for more accessible tools that can assist beginners to conduct complete microbial analysis. To fulfil this requirement, the aforementioned web-based tools such as (M*icrobiomanalyst* [116], *IMNGS* [104], *iMAP*[388], *MGRAST*[389], *wiSDOM* [390], *VAMPS* [387], *Shiny-phyloseq* [391]) are developed but they lack in the following aspects. (i) focusing only the downstream analysis while neglecting raw data processing, (ii) offering only basic analysis which insufficient for complex data sets, (iii) neglecting functional profiling or adapt only outdated methods such as *Tax4Fun* [323] and *PICRSUt1* [316] (iv) lacking confounder analysis capabilities (v) missing of time-series (vi) lacking machine learning and (vii) microbial association networks analysis.

To address these limitations, we developed a free, beginner-friendly, web-based tool called *Namco* which provides end-to-end analysis for microbial data. *Namco* offers a wide variety of features, ranging from the processing of raw data and basic statistics to machine learning and network analysis. As a result, it is able to cover complex data analysis tasks in a comprehensive way compared to other tools **Fig 31.**

The following is a list of the key characteristics of *Namco*: (i) it provides a point-and-click user interface to process raw sequencing data using *DADA2* and *LotuS2,* thereby relieving the user of a load of command-line and step-by-step data processing; (ii) it can be used to analyse large amounts of raw sequencing data (up to 2 gigabytes); and (iii) to the best of our knowledge, *Namco* is the only tool that enables users to carry out functional prediction with *PICRUSt2* in the context of a workflow including 16S rRNA gene s. Other techniques, such as *microbioanalyst* [116] and *wiSDOM*[390], also provide functional prediction; however, they do so only through the use of *Tax4Fun2* [158] and *PICRSUt1* [316], both of which were outperformed by *PICRUSt2* [157]. (iii) In addition to this, *Namco* offers a topic modelling technique [392], which allows users to investigate co-occurring taxa/OTU/ASVs as topics and identify topics related to known sample features. (iv) Perhaps most significantly, it gives users the ability to do differential network analysis between two separate groups. (v) Finally, *Namco* enables users to store sessions in the form of R objects at each step of the analysis process. This feature is helpful for users who want to resume the analysis without having to repeat computationally costly operations again.



Tools	Namco	Microbiome Analyst	MG-RAST	VAMPS	Shiny-phyloSeq	animalcules	GenePiper	METAGEN assist	iMAP	wiSDOM	EzMAP	Metavizr
Raw sequence processing	☑	□	☑	☑	□	□	□	□	☑	□	☑	☑
Filtering	☑	☑	☑	☑	☑	☑	☑	☑	☑	□	□	□
Normalization	☑	☑	☑	☑	□	□	□	☑	□	□	□	□
Taxonomic/sample overview	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Alpha-Diversity	☑	☑	☑	☑	☑	☑	☑	□	□	☑	☑	□
Beta-Diversity	☑	☑	☑	□	□	☑	☑	☑	□	□	☑	□
Statistical tests between taxa	☑	☑	☑	☑	☑	☑	□	□	☑	☑	☑	□
Statistical tests between samples	☑	☑	□	□	□	☑	☑	□	□	☑	☑	□
Rarefaction	☑	☑	☑	☑	□	□	☑	□	□	☑	□	□
Confounder Analysis	☑	□	□	□	□	□	□	□	□	□	□	□
Ordination methods	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Clustering	☑	☑	☑	□	□	□	☑	☑	□	□	□	□
Time Series Analysis	☑	□	□	□	□	□	□	□	□	□	□	□
Functional prediction	☑	☑	☑	□	□	□	□	□	□	☑	☑	□
Pathway visualization	□	☑	☑	☑	□	□	□	□	□	□	□	□
Taxonomic differential analysis	☑	☑	□	□	□	☑	□	☑	□	□	☑	☑
Machine learning & classification	☑	☑	□	□	□	☑	□	☑	□	☑	□	□
Taxon Set-Enrichment analysis	□	☑	□	□	□	□	□	□	□	□	□	□
Integration with public data	□	☑	☑	☑	□	□	□	□	□	□	□	☑
Co-occurrence network	☑	□	□	□	□	□	□	□	□	□	□	□
Topic modelling	☑	□	□	□	□	□	□	□	□	□	□	□
Single network Analysis	☑	□	□	□	☑	□	☑	□	□	□	□	□
Differential networks analysis	☑	□	□	□	□	□	□	□	□	□	□	□
Web-tool	☑	☑	☑	☑	□	□	□	☑	□	☑	□	□
Open source	☑	□	□	□	☑	☑	☑	□	☑	☑	☑	☑
Availabe as Docker image	☑	□	□	□	□	☑	□	□	☑	□	□	□

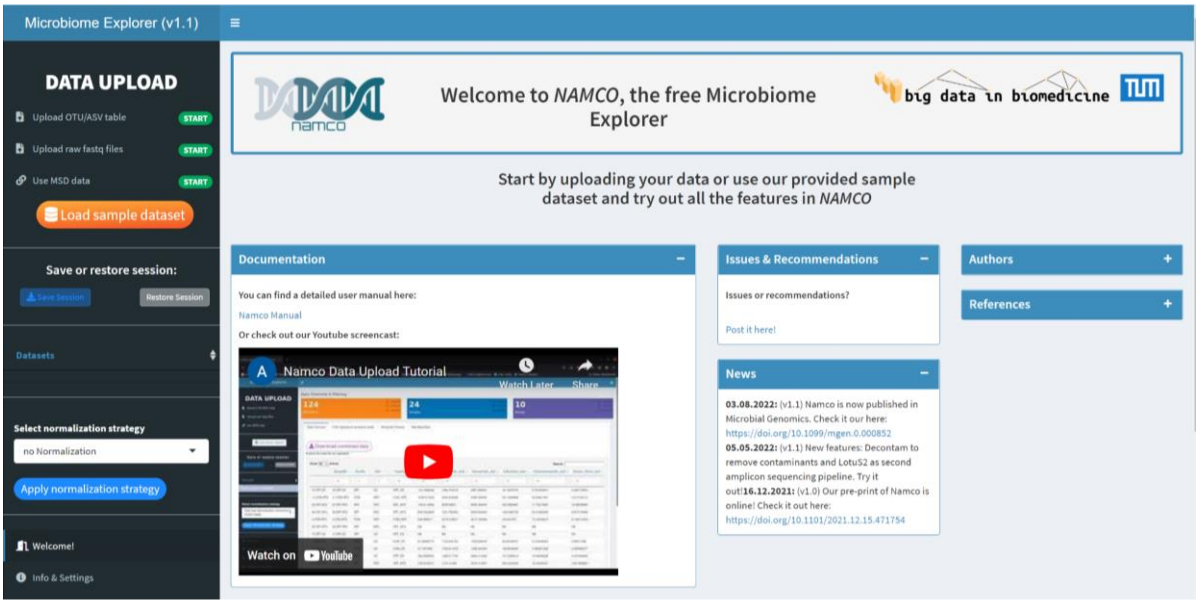
Figure 31: Namco's comparison with other web-based tools for the analysis of 16S rRNA gene data. This figure was originally published in Namco: a free microbiome explorer [367] in microbial genomics journal as an open-access article distributed under the terms of the Creative Commons Attribution License.

4.3. Materials and methods

4.3.1. Namco: framework

4.3.1.1. Development

The *Namco* web tool is implemented using R (https://www.r-project.org) and Rstudio (https://www.rstudio.com) and can be accessed via https://exbio.wzw.tum.de/namco/. It can also accessible as a Docker image, which provides easy installation locally or on a clinical server. This will facilitate the GDPR-compliant analysis of sensitive data without the need for uploading to external servers. Users can stay informed about updates and new releases on GitHub using https://github.com/biomedbigdata/namco/. The homepage of Namco is shown in **Fig .32**



Welcome to *NAMCO*, the free Microbiome Explorer

Start by uploading your data or use our provided sample dataset and try out all the features in *NAMCO*

You can find a detailed user manual here:
Namco Manual
Or check out our Youtube screencast:

Issues or recommendations?
Post it here!

03.08.2022: (v1.1) Namco is now published in Microbial Genomics. Check it our here: https://doi.org/10.1099/mgen.0.000852
05.05.2022: (v1.1) New features: Decontam to remove contaminants and LotuS2 as second amplicon sequencing pipeline. Try it out!**16.12.2021:** (v1.0) Our pre-print of Namco is online! Check it out here: https://doi.org/10.1101/2021.12.15.471754

Figure 32: Welcome page of Namco

The GitHub page of *Namco* is also the preferred way to communicate issues and request features (https://github.com/biomedbigdata/namco/issues). On the other hand, the users can contact the developers by email as displayed on the welcome page of *Namco*.

4.3.1.2. Pipeline

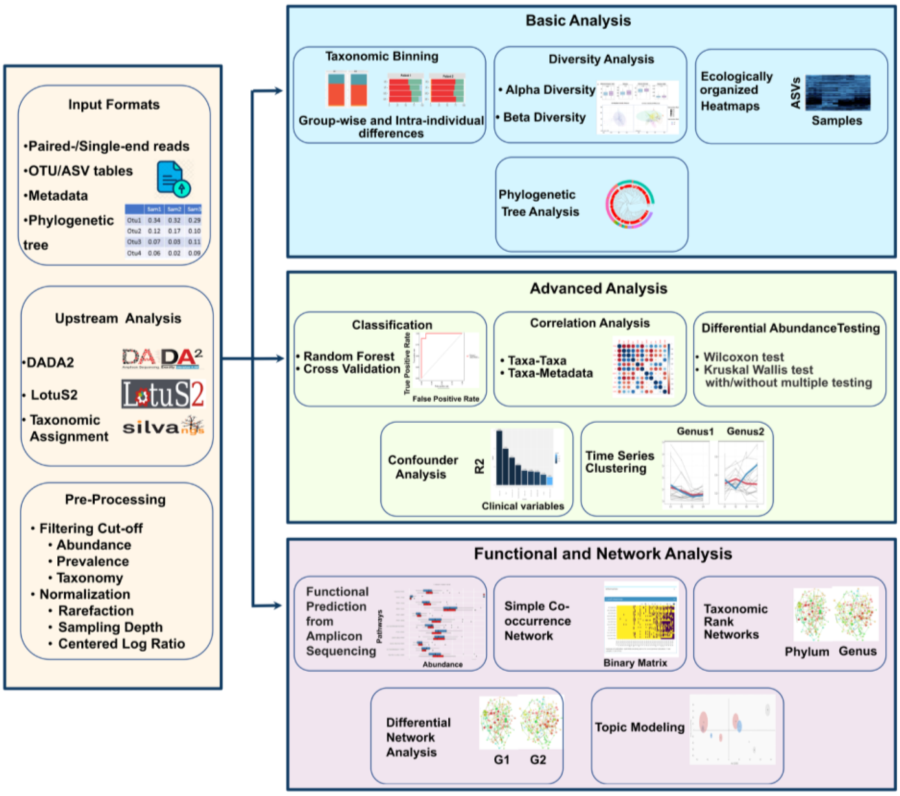


Figure 33: Overview of Namco features. The Namco workflow encompasses the entire process of microbiome data analysis, from raw FASTQ processing and filtering to

statistical, functional, and network analysis, offering a range of visualization options and tables. It allows users to seamlessly navigate through various data analysis tasks.

4.3.2. Upstream analysis

4.3.2.1. Input file formats

Main input files for *Namco* are paired-end fastq files for *DADA2/ LOTUS2* processing and three tab-delimited files such as a feature abundance table with the taxonomic information (separate tabular file with a taxonomic classification if taxonomy information is not provided with a feature abundance table) and metadata file with meta information for each sample to start the analysis at the downstream steps. In addition, users can also provide phylogenetic tree files generated from any 16S rRNA gene analysis pipeline to perform phylogenetic tree analysis or ecologically organized heatmaps. *Namco* workflow is represented in **Fig 33.**

4.3.2.2. Clustering/denoising

4.3.2.2.1. DADA2

Namco supports upstream analysis of 16S rRNA gene-based analyses. Only illumina paired-end fastq reads are supported as input files, and ASV/OTU tables with taxonomy tables are produced as output. One of the pipelines implemented is based on the *DADA2* [96] *R* package, the most widely used denoising algorithm. *DADA2* is an open-source denoising R package. It is a model-based method for correcting mistakes in amplicon sequencing data, such as 454 amplicon sequencing data, while recognizing fine-scale variation. *DADA2* does not require a reference and can be utilised on any genomic locus. It can filter, dereplicate, identify chimaeras, merge paired-end reads and contains a novel quality-aware error model for Illumina amplicon data. Users can alter the default parameters such as trimming length, and select their own choice of settings. It uses updated versions of the SILVA (v 138) [101] as default reference database for taxonomic classification. *Namco* stores *DADA2* output as a phyloseq [113] object for further downstream analysis.

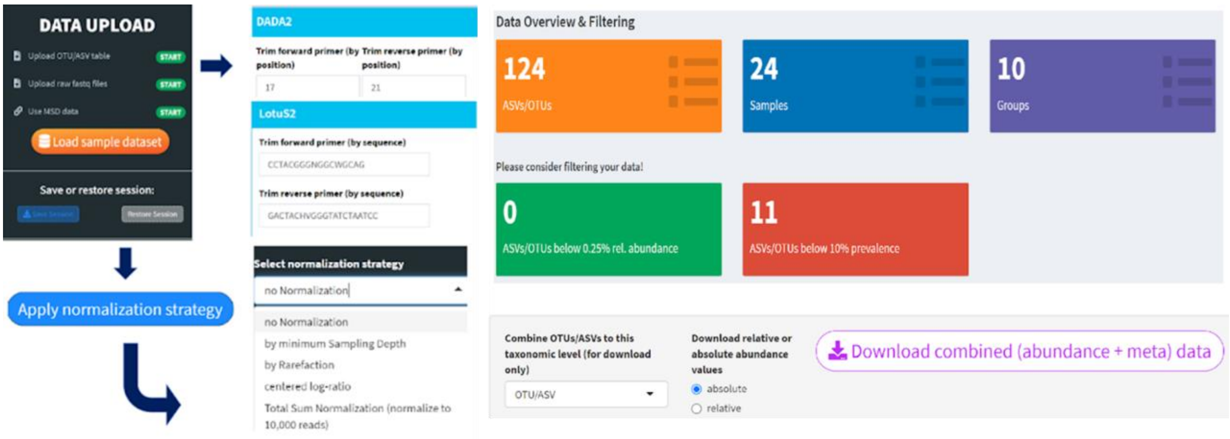
4.3.2.2.2.LotuS2

LotuS2 [103] is an open-source bioinformatics software that allows for flexible data analysis of amplicon sequencing data. It includes six sequence-clustering algorithms such as *UPARSE* [106], *UNOISE3*[232], *CD-HIT* [118], *SWARM* [393], *DADA2* [96] and *VSEARCH* [105]. In addition, in-depth pre- and post-processing options that can be adjusted by experts or used with default settings for beginners. LotuS2 follows a strict read filtering approach during the clustering step, which includes 21 different quality filtering metrics, probabilistic and Poisson binomial read filtering [106,394] and removal of reads that cannot be dereplicated. These filtered and cleared reads are then clustered into OTUs/ASVs using one of six available sequence-clustering methods.

4.3.2.3. Preprocessing

Namco employs *Rhea*[117] R scripts to perform basic analysis such as taxonomic profiling and diversity analysis. Once the denoising step is completed, users can view the data overview section in *Namco*, which summarises sample details including the total number of samples, the total number of ASVs present in all samples and the number of metadata groups available (**Fig 34**). This will give an overview of the input data and help in proceeding further with downstream analysis.

It is well aware that microbiome data is very sparse due to the presence of numerous rare taxa and frequently have zero counts in many samples. Sometimes, rare taxa may also be caused by contamination and/or by sequencing errors. Hence, filtering is an important preprocessing step to remove low-quality features to improve the accuracy of downstream statistical analysis. *Namco* offers various filtering options including filtering samples from the feature table based on metadata and filtering taxa based on absolute and relative abundance. The next important preprocessing step is the choice of normalisation methods. *Namco* provides three different normalisation methods such as centred log-ratio[128], sampling depth and rarefaction. Users can select one among three methods and compare the downstream results between them.



DATA UPLOAD

Upload OTU/ASV table START

Upload raw fastq files START

Use MSD data START

Save or restore session:

DADA2

Trim forward primer (by position): 17

Trim reverse primer (by position): 21

LotuS2

Trim forward primer (by sequence): CCTACGGGNGGCWGCAG

Trim reverse primer (by sequence): GACTACHVGGGTATCTAATCC

Select normalization strategy: no Normalization

- no Normalization
- by minimum Sampling Depth
- by Rarefaction
- centered log-ratio
- Total Sum Normalization (normalize to 10,000 reads)

Data Overview & Filtering

124 ASVs/OTUs

24 Samples

10 Groups

Please consider filtering your data!

0 ASVs/OTUs below 0.25% rel. abundance

11 ASVs/OTUs below 10% prevalence

Combine OTUs/ASVs to this taxonomic level (for download only): OTU/ASV

Download relative or absolute abundance values

● absolute

○ relative

Load sample dataset
Restore Session
Apply normalization strategy
Download combined (abundance + meta) data

Figure 34: Workflow of upstream analysis and data overview page of Namco. User can start 16S rRNA gene data analysis either by uploading fastq (raw) files (upper limit 2GB) or with pre-processed feature tables with taxonomic information. For upstream analysis, Namco offers two amplicon sequencing analysis pipelines, DADA2 (left) and LotuS2. For DADA2, user can insert the lengths of the forward and reverse primers to trim from the reads. After denoising steps, feature table automatically loaded into the Namco environment and data overview page provides overview of the samples in the dataset including total number of ASVs/OTUs, number of samples and number of sample groups. Namco also offers different normalization methods such as CLR and normalize to 10.000 reads etc.

Overview:

If users want to apply one of the advanced filtering methods, they first have to click the checkbox next to it and then select the appropriate value. Users can apply multiple methods at once, by simply clicking multiple checkboxes. They will be applied from top to bottom to the dataset.

The methods are:

- **Filtering by minimum abundance (A)**: Removes all OTUs/ASVs, which have an abundance value - over all samples - below the selected cutoff

- **Filtering by relative abundance (B)**: Removes all OTUs/ASVs, which have a relative abundance value - in each sample - below the selected cutoff (cutoff has to be given in %; 0.25% which means all OTUs which do not make up more than 0.25% of a sample will be removed.

- **Filtering by occurrence in samples (C)**: Removes all OTUs/ASVs, which appear at most the number of times as the selected cutoff. An OTU/ASV does not "appear" in a sample, if its abundance is 0.

- **Filtering by highest variance (D)**: Keeps only the number of OTUs/ASVs you selected with the highest variance over all samples.

- **Filtering by prevalence (E)**: Keeps only the OTUs/ASVs with a prevalence value over the selected cutoff (given in %).

4.3.3. Downstream analysis

4.3.3.1. Diversity analysis

Once the preprocessing is done, users can move to basic analyses such as alpha and beta diversity. Alpha diversity, which estimates the variation within one sample, can be calculated by various measures such as richness, the Simpson [395] and the Shannon effective indices [396]. The alpha-diversity feature in *Namco*, currently supports Shannon index, Simpson index and richness. The results are plotted and summarised as box plots for each group. Users can select different categorical meta-groups to visualise alpha diversity and their significance. The beta-diversity estimates the variation between the groups. *Namco* requires a phylogenetic tree to calculate dissimilarity using one of the most common distance matrices such as weighted and unweighted Unifrac distances, generalised Unifrac, Bray Curtis dissimilarity and Variance adjusted Unifrac Distance [397]. The results are presented in 2D ordination plots based on Multi-dimensional scaling (MDS) and non-metric multidimensional scaling (NMDS). In addition, *Namco* also displays the hierarchical clustering of the samples using the chosen distance method as branches in a dendrogram, which helps to identify closely related samples. The significant difference between groups is calculated by a permutational multivariate analysis of variances using adonis function from *Vegan* R-package). To account for multiple comparisons, p-values are adjusted using Benjamini-Hochberg correction method [398].

4.3.3.2. Groupwise differential abundance analysis

This analysis allows users to identify statistically-significant ASV/OTUs between the groups using *SIAMCAT* r package [399] via non-parametric Wilcoxon test. Users can choose metadata groups to compare against one with another and adjust the significance level. Users can calculate differential abundance at different taxonomic levels such as the phyla, genus onto which the feature table can be collapsed. The association plot exhibits the distribution of microbial relative abundance and also indicates significance of the relationship and a generalised fold change which serves as a non-parametric measure of effect size. Users can modify the number of significant ASVs to be displayed in the association plots and also sort the features based on fold change, p-value or prevalence shift.

4.3.3.3. **Correlation analysis**

Namco performs correlation to identify the significant positive and/or negative relationships within taxa or between taxa and metadata such as continuous experimental variables. Additionally, *Namco* takes into account the relative abundances of features at various levels (phylum, class, order, family, genus etc).

4.3.3.4. **Confounding analysis**

Confounding variables are extraneous or hidden causative variables that influence both the dependent and independent variables. The presence of confounders may hide the actual relationship between the variables in study [400]. For example, microbiomes are strongly related with several host variables including sex, body mass index (BMI), age and geographical location, and they act as the strongest potential confounders [401]. In this section, *Namco* utilises the permutational multivariate analysis of variances (adonis function of the *Vegan R*-package)[383] to elucidate many co-variables or confounding factors. The proportion of variance explained by co-variables is assessed by computing R2 values and evaluating their significance, with a p-value threshold of ≤ 0.05.

4.3.3.5. **Machine learning**

This section of *Namco* allows users to predict important features using a non-parametric machine learning algorithm called Random Forest (RF). RF has been utilized in microbial data analysis [402,403] to identify the important featuers by measuring the increase in classification errors that results from permuting the data. *Namco* integrates the *ranger* [404] R package, which provides faster implementation especially for high dimensional data. Users can select metadata for which a prediction model should be built. *Namco* also provides a flexible environment to modify the advanced parameters including resampling method, number of decision trees, the number of cross-validation, and the ratio of training and testing sets. Graphical output is created to summarise model's performance using the confusion matrix and receiver operating characteristic curve (ROC)-Plot. Users can also view the top most important features used for building the model.

4.3.3.6. **Time series clustering**

Namco offers time series clustering analysis to investigate how microbial communities either at different taxa, or at ASV/OTUs, or other features like richness change over time. This analysis can be utilized to study alterations in microbial populations over time in response to a specific treatment or during different stages of host development. This type of analysis can provide insight into the mechanisms underlying microbial changes, such as how a treatment affects the growth or survival of different microbial species, or how the host's development affects the microbial community. *Namco* provides various options to customize the inputs for time series line charts such as exhibiting changes in relative or absolute abundance or in terms of richness.

4.3.3.7. **Functional prediction**

Namco includes *PICRUSt2* [157] to infer functional profiles using 16from microbial communities. *PICRUSt2*[157] provides improved accuracy and flexibility for marker gene metagenomic inference compared to other tools including *Tax4Fun2, Piphillin, PanFP and PICRUSt1*. Since *Namco* runs in a Docker container, *PICRUSt2* is installed via the conda packaging manager, and the necessary scripts are later called using command line arguments inside the R shiny app. *Namco* retrieves representative sequences of ASV and feature tables as inputs to perform functional predictions. In addition to functional inference, *Namco* also allows users to conduct differential abundant analysis on the inferred metagenome, pathways and Enzyme Classification (EC) using *Aldex2* [131]. In a comparative study of differential abundance tests, *Aldex2* emerged as the most reliable and consistent method across multiple studies. It also showed the highest degree of agreement with the intersection of results obtained from various other approaches. [405]. In the bar charts between the groups, the number of significant functions with a BH-adjusted p-value of less than 0.05 is shown.

4.3.3.8. Network analysis

Microbial interactions are not only essential in maintaining the stability of their community, but also in maintaining the homeostasis of the host environment. Several studies highlighted the key role of these interactions in development, host immunity and metabolism [406]. Hence, exploring microbial networks became an integral part of microbial analysis. In this section, *Namco* provides two different options to build microbial co-occurrence networks. The first approach is a simple one where the OTU abundance matrix is then transformed into a binary matrix after filtering OTUs with an abundance cutoff and in this step, the number of pairs of present OTUs are counted (co-occurring OTUs). This is done separately for two groups of samples (eg. case and control), which can be chosen manually. Then, for each pair of OTUs, the log2 fold-change between two groups or difference is computed and visualized as a network. The second advanced approach is using the *NetComi* [407] R package. Users can generate single network at different taxonomy levels using different algorithms offered by *NetComi* to understand the association within microbial community. In addition, *Namco* can also perform differential network analysis to identify differentially associated single pairs of taxa between two groups.

SparCC

Sparse Correlations for Compositional data (*SparCC*) [149] is frequently used to study the microbial network, especially in human gut microbiome studies [408], as well as environmental studies [409]. SparCC is an algorithm that iteratively estimates the linear Pearson correlations between the log-transformed relative abundances of the different OTUs in a sample. The log-ratio transformation is used to reflect the true ratios of abundances present in the environment, by introducing a virtual reference point. This makes the log-ratio of two OTUs independent of the other OTUs in the sample, which helps to avoid spurious correlations. Compared to direct Pearson correlations, *SparCC* is better suited to avoid spurious correlations [410] at the cost of higher computational complexity [411].

CCLasso: correlation inference for compositional data through Lasso

CCLasso [412] is a latent variable model used to build a regularised correlation network of microbiome data. It addresses the issue of compositionality using CLR method and utilises the least squares method with an$\ell 1$-penalty. The performance of *CCLasso* is similar to *SparCC* in terms of consistency and reproducibility, but it is better in dealing with spurious relationships [412]. However, one of the major drawbacks of these correlation methods is that they fail to differentiate between direct and indirect edges. Indirect edges are edges between two species that are caused by third variables, which can be other taxa or environmental factors. Indirect edges are described as edges that occur due to third causes. In order to address both direct and indirect impacts, two distinct approaches for dealing with correlation-based methodologies were developed by Feizi et al. [413] and Barabási [414]. On the other hand, these strategies have not yet been validated utilising a microbial co-occurrence network analysis.

SPIEC-EASI

One of the main advantages of *SPIEC-EASI* (SParse InversE Covariance Estimation for Ecological Association Inference) [415] is being able to differentiate between direct and indirect relationships in microbial network inference. It achieves this using the concept of conditional independence. SPIEC-EASI employs the CLR transformation to overcome compositionality and generates a co-occurrence network using one of two methods. The first method uses sparse graphical model inference (*Glasso)* to calculate sparse inverse covariance matrix based [416]. The second approach, which is called the Meinshausen Bühlman method, is a node-wise regression model where the expression of each taxon is described by the remaining taxa [417,418]. The local neighbourhood of a node or taxa is then described by the taxa used to predict its expression. Finally, the appropriate amount of sparsity of a network is inferred by using the stability approach to regularization selection [419]. The final result is an undirected weighted graph where the edges imply the conditional dependency between two taxa.

NetCoMi

Network Construction and comparison for Microbiome data (*NetCoMi*) [407] integrates existing methods for generating single and differential microbial association networks. It implements frequently used normalisation methods such as total sum scaling, cumulative sum scaling, rarefication, CLR transformation and variance stabilising transformation for data normalization and to handle zero inflation. Various methods including *SparCC*[149], *proportionality* [420], *SPIEC-EASI* [421] and *SPRING* (the Semi-Parametric Rank-based approach for INference in Graphical model) [422] are implemented to overcome compositionality bias. In addition, *NetCoMi* features differential network analysis and differential association analysis. Differential network analysis utilizes permutation tests to evaluate the significantly different taxa between the groups. Differential association analysis uses Fisher's z-test [423], a non-parametric resampling procedure [424] and the discordant method [425] to build differential networks that are limited to differentially associated taxa. Overall, *NetCoMi* gives its users considerable flexibility to select the most suitable tools or methods for a particular dataset.

There are a variety of methods available for studying microbial interactions from a network-level perspective, and the choice of method can be challenging for researchers. These methods vary in terms of complexity and computational requirements, and the trade-offs between these factors must be carefully considered when choosing a method for a specific study. Additionally, the lack of a comprehensive benchmark dataset or commonly accepted simulated dataset makes it difficult for researchers to evaluate the performance of different methods and to choose the most appropriate for their analysis. This highlights the need for the development of standardised benchmark datasets and guidelines for evaluating the performance of different network models in order to facilitate the selection of the most suitable method for a given study **(Fig 35).**

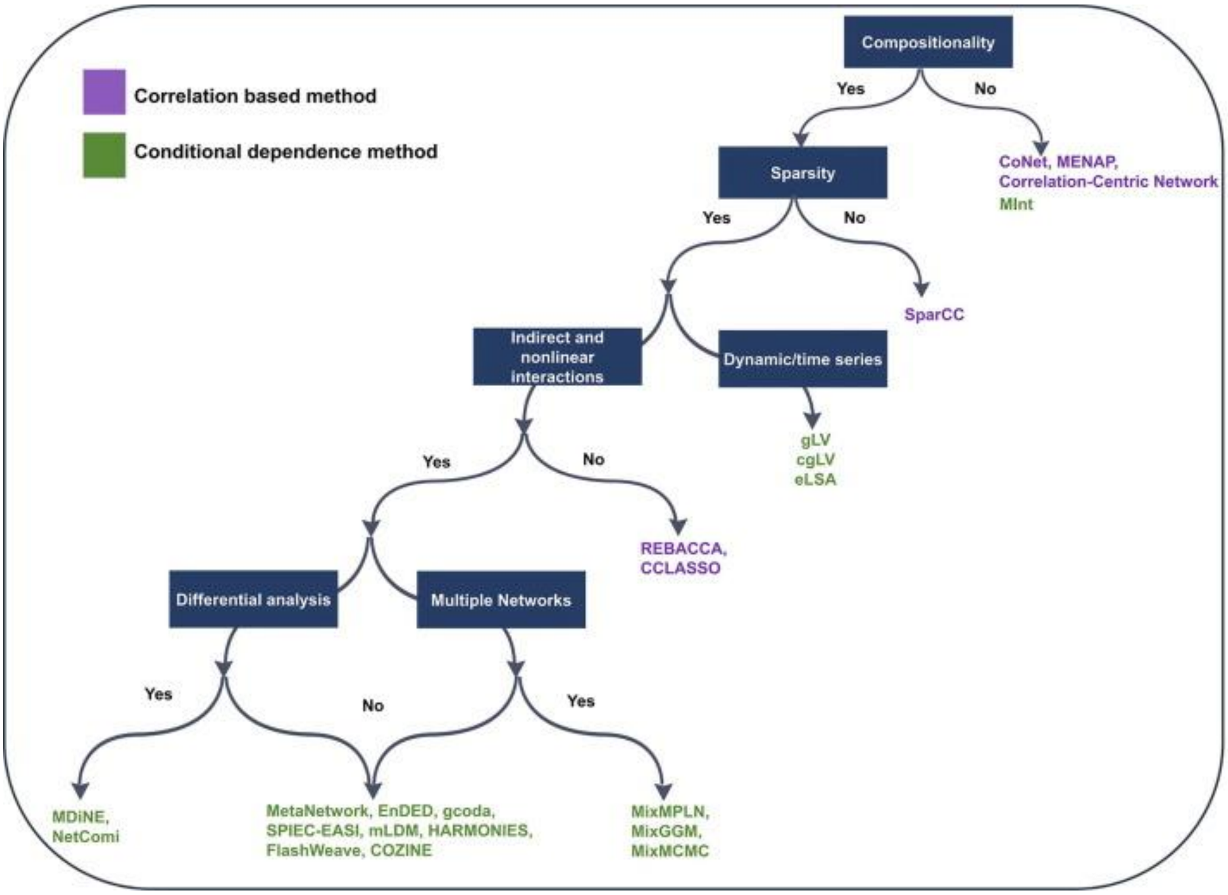


Figure 35: Workflow indicating the suitable network approaches depending on different challenges. This figure was originally published as Network analysis methods for studying microbial communities: A mini review [132] in Computational and Structural Biotechnology Journal as an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

4.3.3.9. Differential network analysis

Namco utilises the *NetComi* package for differential network analysis. The *NetComi* package uses correlation measures to compare the group differences. It uses its own R implementation for testing whether the correlation coefficients are significantly different between the groups. For multiple testing adjustment, all methods provided by p.adjust () (stats package) as well as a method for controlling the local false discovery rate provided by the fdrtool package [426](Klaus and Strimmer., 2015) are available in *NetCoMi* [407].

4.3.3.10. **Topic modeling**

Topic modelling is a computational framework which was originally designed to uncover the hidden thematic structure in document collections [427]. It is a soft clustering technique where each instance belongs to each cluster to a certain degree (**Fig 36**). It uses Latent Dirichlet allocation (LDA) [428], which is a popular algorithm in the area of natural language processing (NLP). This concept was adapted to metagenomic analysis to explore co-occurring taxa as topics and to find the association with the metadata with impressive insights. In other words, it considers a single microbiome sample as a document, every OTU/ASV as terms and community type as topics. *Namco* employs a Structural Topic Model wrapper, *find_topics* from *themetagenomics R* [429] package to predict topics and study their effects with metadata. *Namco* requires the number of topics and reference level variables as inputs from users and generates three different ordination plots representing the association of each topic with sample conditions.

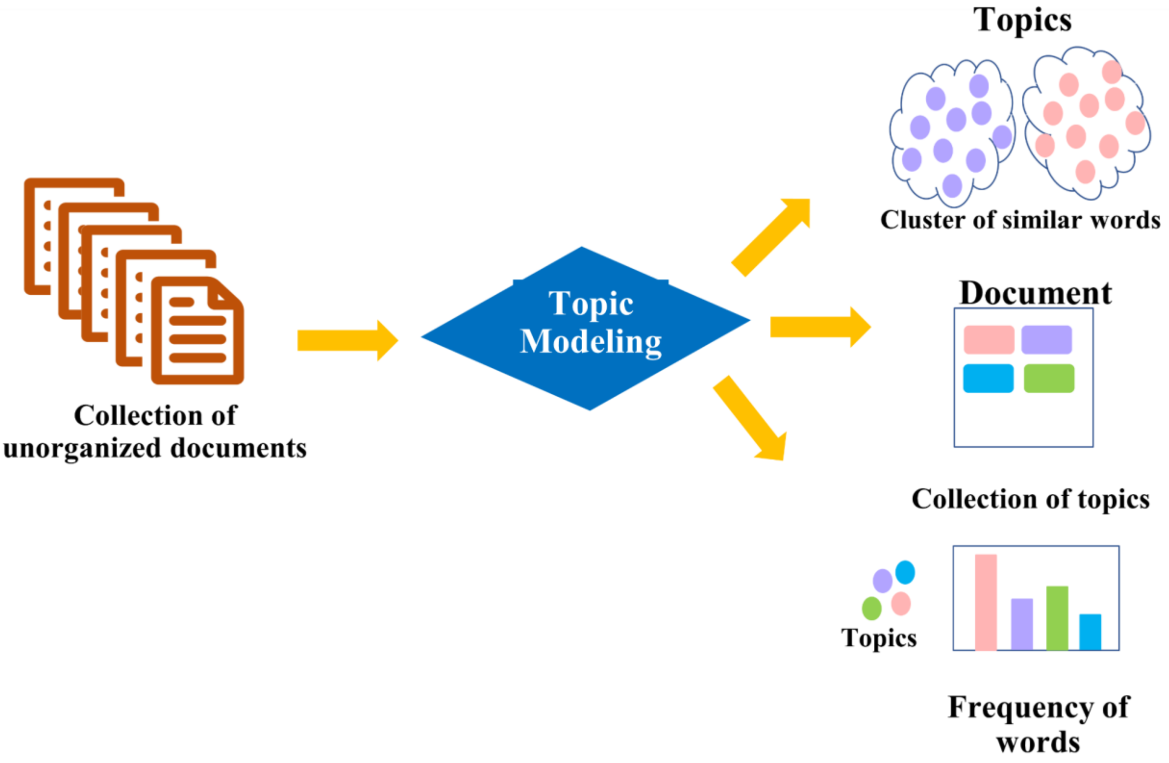


Figure 36: The basic concept of the topic modelling algorithm.Topic Modeling uses the LDA algorithm to cluster similar terms into topics. Similar topics can be categorised into relevant documents. Similarly, in microbiome data analysis, similar features or OTUs can be grouped into different topics and topics can be related to the reference metadata.

4.3.4. Use case

To demonstrate the functions of Namco, we analysed human faeces samples from a cross-over interventional research. The goal of the study is to understand how changes in diet can impact the diversity and composition of the microorganisms in the gut and to inform the development of healthier convenience food products with increased fibre content. The study also aims to assess customer acceptance of such products.

4.3.4.1. **Ethics statement**

The ethical committee of the Faculty of Medicine at the Technical University of Munich in Germany accepted the study protocol (approval number: 529/16S). Consideration was given to the guidelines of the International Conference on Harmonization of Good Clinical Practice and the Declaration of Helsinki of the World Medical Association (as updated in Fortaleza, Brazil, 2013). All study participants have provided informed consent in writing. The investigation was recorded in the German Clinical Trial Register (DRKS00011526).

4.3.4.2. Study design

This study was a single-blinded, controlled cross-over study that recruited middle-aged volunteers with elevated waist circumference and aged between 40-65 years. The study population was half male and half female. The study aims to investigate the effects of a specific diet intervention, which includes meatloaf in a bun and pizza, on the gut microbiome (**Fig 37**). The study participants visited the study centre four times, during the first visit baseline characteristics were collected. The inclusion and exclusion criteria for the study population can be found in a publication by Brandl et al [430]. The intervention was described in more detail in a publication by Rennekamp et al [431].

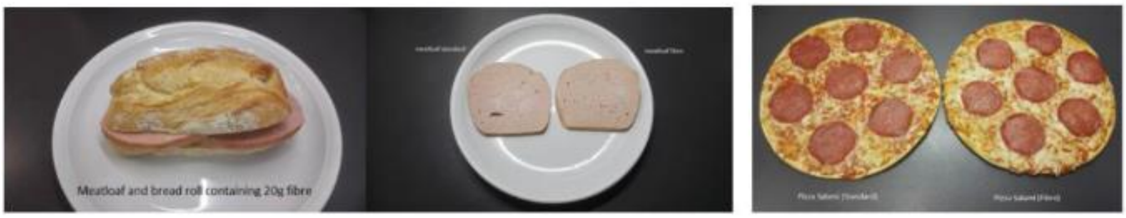


Figure 37: Design of meals for fibre-enriched Intervention (enriched) and the placebo (standard) meatloaf and salami pizza meal. Two different types of intervention meals such as (meatloaf in a bun and a pizza) both, either enriched with fibre (intervention) (IM) or not (placebo) (M), were used in this study. The first interventional meal (meatloaf in a bun, IM1) the white bread roll in the fibre-enriched meal contained an additional 5.7% wheat fibre (VITACEL® WF600) and the meatloaf (Leberkas) a mixture of 3.1% wheat fibre and 4.5% resistant dextrin.

4.3.5. Phenotypic characteristics of the study group

The study group (N = 11 females and 10 men of the same age and gender) received the same intervention and placebo (Table 5). Baseline measurements were taken in the morning, after an overnight fast, body weight and composition were measured using a Seca Medical Body Composition Analyzer, mBCA 515 (Seca GmbH & Co. KG, Hamburg, Germany). A stadiometer was used to measure body height in a standing stance without footwear (Seca GmbH & Co. KG, Hamburg, Germany). The BMI formula was weight (kg)/height (m2). The waist circumference was measured using a measuring tape halfway between the lowest rib and the iliac crest (Seca GmbH & Co. KG, Hamburg, Germany).

Table 5. Overview of the study group characteristics. Indication of the mean values and the standard deviation for the participants is given, besides the significant differences in traits between the sexes.

	Mean	**S ID**	**Differences between sexes**
Weight [kg]	90.14	11.42	0.0080 (**)
Height [m]	1.73	0.08	1.35e-05 (***)
BMI [kg/m2]	30.12	2.41	0.8490 (ns)
Fat-free mass [%]	62.98	6.74	2.40e-08 (***)
Fat mass [%]	37.02	6.74	2.40e-08 (***)
Skeletal muscle mass [kg]	27.55	6.28	9.34e-07 (***)
Visceral fat [kg]	3.24	1.32	4.55e-05 (***)
Waist circumference [cm]	101.3	7.26	0.0058 (**)

4.3.6.Sample preparation

The participants were instructed to arrive at the study centre sober (10 hours prior to the appointment) and received either the intervention or the placebo lunch. Moreover, a pill containing food colouring was provided. The consumption of the dye stains faeces green, which helps to correlate collected samples with food consumption.

The time of the meal and the time of excretion were recorded, and a mean transit time of 34.74 ± 24.69 h hours was determined. Since a coloured capsule was supplied with each meal, the faeces sample may be attributed to the corresponding meal. The dye causes green to the sample, and recognizable colouration was detected in the data.

Participants were given two different types of food (meatloaf in a bun and a pizza) both, either enriched with fibre content (intervention) (IM) or not (placebo) (M). The first interventional meal being meatloaf in a bun, (IM1) contained an additional 5.7% wheat fibre (VITACEL® WF600) and the meatloaf (Leberkas) a mixture of 3.1% wheat fibre and 4.5% resistant dextrin.

The second intervention (pizza, IM2) was also fibre-enriched containing up to 20 g of fibre with 3.0 % wheat fibre, 2.4% powdered cellulose and 2.1% inulin (Table 6). The intervention meals thus constituted a major part of the recommended daily fibre intake. As the fibre content is above 6 g per 100 g, the food products are considered as high-fibre products.

Table 6. Nutritional values per serving for the intervention (enriched) and the placebo (standard) meatloaf and salami pizza meal as well as the differences between the intervention and placebo meals.

	Portion meatloaf with bun 240 g		**Portion salami pizza 320 g**	
	Enriched	**Standard**	**Enriched**	**Standard**
Energy [kcal]	413	587	829	876
Fat [g]	13	35	41	45
Carbohydrate [g]	47	47	75	83
Total fibre [g]	19	2.9	20	6

4.3.7. Sample sequencing

In this study, gut microbiota was analyzed by sequencing the 16S rRNA gene . The sequencing was performed at the ZIEL Core Facility Microbiome, Technical University Munich, Germany. The sample preparation and sequencing method are described in detail in another publication [347]. Briefly, the DNA from the samples was isolated using an in-house developed protocol. The V3V4 region of the 16S rRNA gene was amplified and purified, and the resulting amplicons were paired-end sequenced on an Illumina MiSeq. The sequencing data is available under BioProject ID PRJNA774891.

4.4. Research question

The main aim is to investigate the effects of a high-fibre diet on the gut microbiome. It has been previously shown that increased fibre intake can be protective against the development of cardiovascular disease [432,433] and malignant diseases [434,435]. Additionally, specific health claims are associated with specific types of fibres. In this study, the researchers specifically examined the presence of butyrate-producing bacteria in the gut, which can be promoted by the fibre-enriched intervention. The goal of this analysis is to determine whether changes in diet can have a permanent effect on the composition of the gut microbiome.

4.5. Results

4.5.1. Diversity analysis

Namco was used to analyse the gut microbial changes following a dietary intervention. Paired-end FASTQ files were processed using the *DADA2* denoising step embedded in *Namco*, using default parameters. During the *DADA2* processing, an abundance-based filter of 0.25% was applied to reduce sparsity in the data. This step is used to remove low abundance reads and help in increasing the accuracy of the analysis [436].

ASVs were normalised to 10,000 reads before downstream analysis and outliers were removed. Additionally, a prevalence filter cutoff of 10% was introduced to further filter out rare ASVs. The alpha-diversity measures including Shannon, richness, Simpson Index, effective Shannon entropy and effective Simpson entropy were calculated and compared between the intervention group (IM) and the control group (M). However, no significant differences were observed in these measures between the groups. Similarly, no significantly different clusters were found in beta-diversity measures, including unweighted and weighted Unifrac between the IM and M groups. **(Fig. 38).**

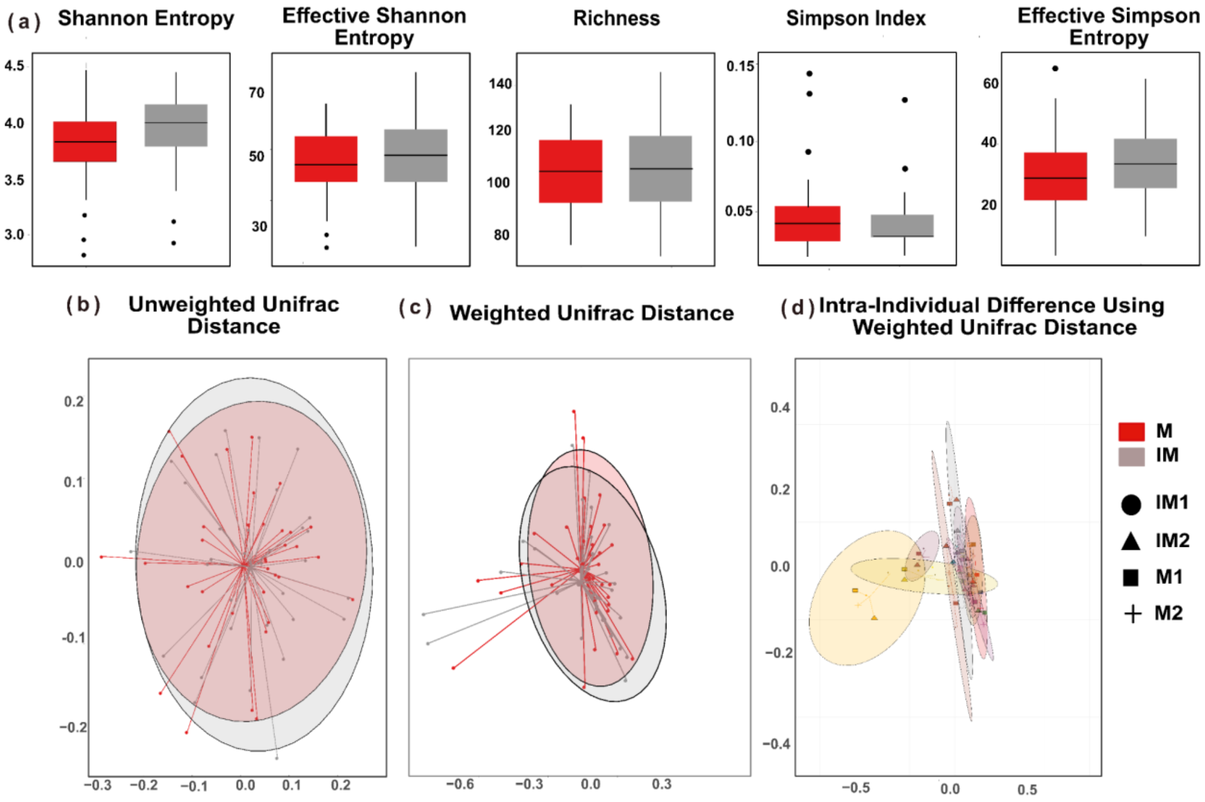


Figure 38: Microbial Diversity between intervention groups. (a): Alpha diversity measures associated with intervention (IM) and non-intervention meals (M). There was no significant difference identified between the groups. (b) nMDS visualizations of beta diversity analysis using the unweighted or weighted (c) Unifrac distance. (d) nMDS visualizations of beta diversity for intra-individual patients across two intervention meals and their respective control using weighted Unifrac distance. This figure was originally published as Namco: a microbiome explorer [367] in microbial genomics journal as an open-access article distributed under the terms of the Creative Commons Attribution License.

4.5.2. Taxonomic distribution

In the results, we mainly focused on the bacterial composition of different groups, including intervention groups (IM1 and IM2) and non-intervention groups (M1 and M2). The study found that the dominant phyla in all groups were *Firmicutes* and *Bacteroidota*, which together made up 90% of the total bacterial composition. The relative abundance of *Firmicutes* was slightly higher in the intervention groups, while *Bacteroidota* were slightly more abundant in the non-intervention groups **(Fig. 39 (a))**. Other phyla, such as *Actinobacteria*, *Verrucomicrobia*, and *Proteobacteria*, were found to have a relatively low mean relative abundance between the intervention and non-intervention groups. At the genus level, most individuals had a uniform distribution, but one individual had a high level of the genus *Prevotella* (54% of relative abundance) **(Fig. 39 (b))**. Overall, the study found that there were very few differences in bacterial composition between the intervention and non-intervention groups, but there was also a lot of intra-individual heterogeneity.

Figure 39 (c) depicts the top 20 genera, while **Fig 39 (d)** depicts the intra-individual variation. In both the M and IM groups, *Bacteroides* was the most prevalent genus, followed by *Faecalibacterium, Prevotella,* and *Agaebacter.* As a next step, we performed a non-parametric paired Wilcoxon test to identify microbial changes between the IM and M meals. The test found that five genera, *Ruminococcaceae Incertae Sedis, Butyricicoccus, Anaerostipes, Fusicatenibacter,* and *Parabacteroides* were found to significantly different between M and IM groups without multiple corrections **(Fig 40)**. However, only *Ruminococcaceae Incertae Sedis* remained significant after multiple testing correction using BH. The study showed that the *Anaerostipes,* gram-positive, a butyrate-producing bacterium, was found to be in higher abundant in IM2 than to M2. *Anaerostipes* is an anaerobic bacterium from the family of *Lachnospiraceae* and was reported as the one of the highly abundant bacteria in the normal healthy gut [437]. Earlier findings, proved that the presence of *Anaerostipes* is positively correlates with the high-fibre diets and negatively correlate with BMI [438,439].

In our study, we observed a significant difference between IM1 and M1 for two bacteria such as *Ruminococcaceae Incertae Sedis* and *Parabacteroides*. *Ruminococcaceae* has been reported as one of the main bacteria for producing short-chain fatty acids (SCFA), such as butyrate, which is an essential SCFA in maintaining the healthy gastrointestinal tract[440] and helps in preventing possible weight gain [441]. Additionally, Ruminococcus bromii is recognized as the primary species responsible for fermenting resistant starch, which has been linked to health benefits like weight regulation and diabetes prevention [442]. Parabacteroides have also been shown to have metabolic benefits and a negative correlation with BMI [443,444]. One specific Parabacteroides species, P. distasonis, is known to be part of the core gut microbiome [445–447] and can produce succinic acid, as well as promote the production of bile acid to regulate host metabolism[444,448]. Ezeji et al. also discovered that fibre-rich dietary intervention groups were enriched with Parabacteroides[449]. Dietary fibre intervention greatly enhanced the proliferation of the beneficial genera *Ruminococcaceae Incertae Sedis, Parabacteroides* and *Anaerostipes* at the genus level. We also observed that Fusicatenibacter was more prevalent in IM1 compared to M2, while Butyricicoccus was more abundant in M1 than in IM2.

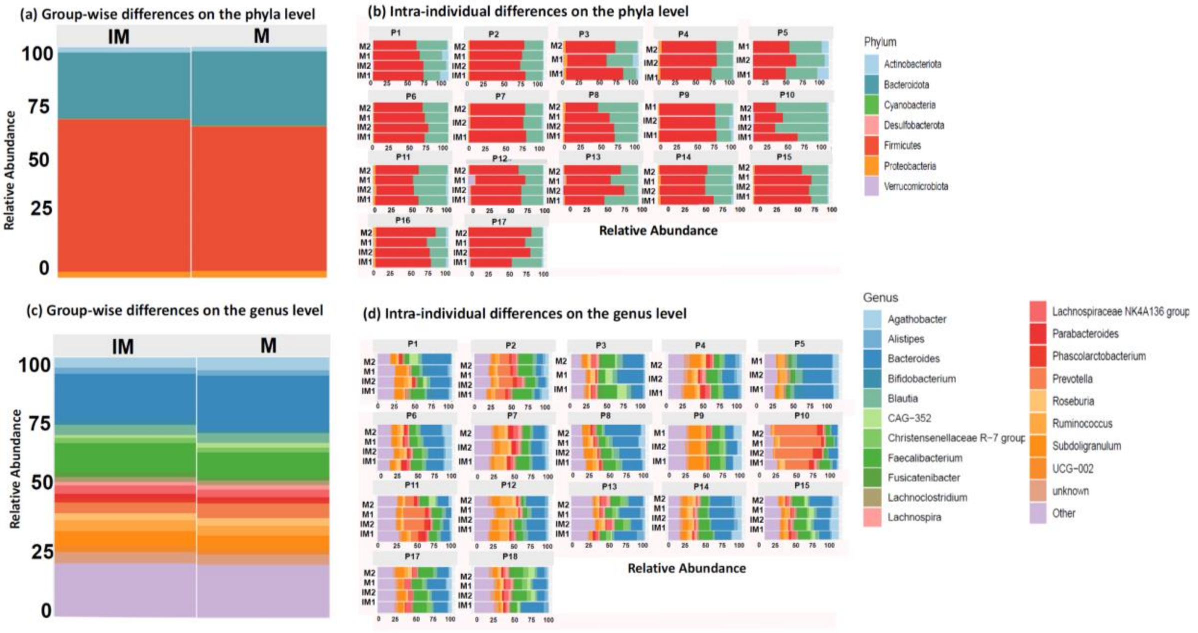


Figure 39:Taxonomic composition across intervention groups. (a, c) represents the relative frequency of phyla and genus, respectively between intervention and non-intervention groups. (b, d) Bar plots represent the inter-individual variation of gut microbiome between intervention and non-intervention groups. This figure was originally published as Namco: a microbiome explorer [367] in microbial genomics journal as an open-access article distributed under the terms of the Creative Commons Attribution License.

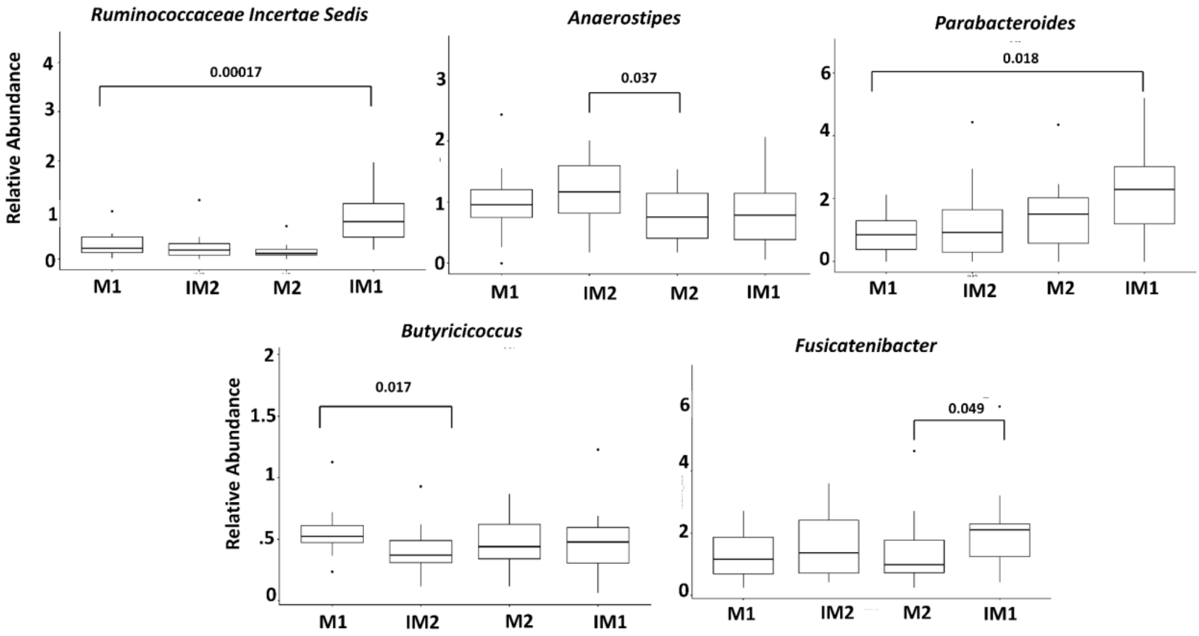


Figure 40: Differential abundance of species between intervention groups. The boxplots depict the differences in mean proportions of genera across four meal groups, as determined by the non-parametric Wilcoxon Rank test. The four groups are labeled as

IM1, IM2, M1, and M2, representing the first and second interventional meals and the first and second non-interventional meals, respectively. Significant differences between the groups are indicated by asterisks above the corresponding boxplots, with p-values less than 0.05 considered significant. This figure was originally published as Namco: a microbiome explorer [367] in microbial genomics journal as an open-access article distributed under the terms of the Creative Commons Attribution License.

4.5.3. Correlation of gut microbial composition and clinical metadata

To study the meaningful association between taxa and continuous variables of interest such as fat mass [%], skeletal muscle mass [kg], BMI, fat-free mass [%] and age, we performed Spearman correlation at phyla level **(Fig 41 (a))**. On the phyla level, our results showed that *Deltaproteobacteria Verrucomicrobiota*, *Firmuicutes,* and *Actinobacteriota* are negatively correlate with BMI. Whereas on the genus level, *Lachnospira*, *the Lachnospiraceae FCS020 group, Phascolarctobacterium*, *Alistipes*, *Oscillospiraceae UCG-005* and *Prevotella,* were positively correlated with BMI **(Fig 41 (b))**. On the other hand, *Ruminococcus, Lachnoclostridium* and *Bacteroides* were significantly negatively correlated with BMI. When compared to the M groups, the IM group had a slightly lower abundance of *Prevotella, Lachnospiraceae FCS020 group*, and *Phascolarctobacterium,* all of which had a positive correlation with BMI. Additionally, *Anaerostipes*, *Lachnospira and Eubacterium ruminantium group* were found to have a favourable association with the percentage of fat mass. On the other hand, the *Clostridia UCG-014 and Rikenellaceae RC9* gut group were found to have a negative association with the proportion of fat mass

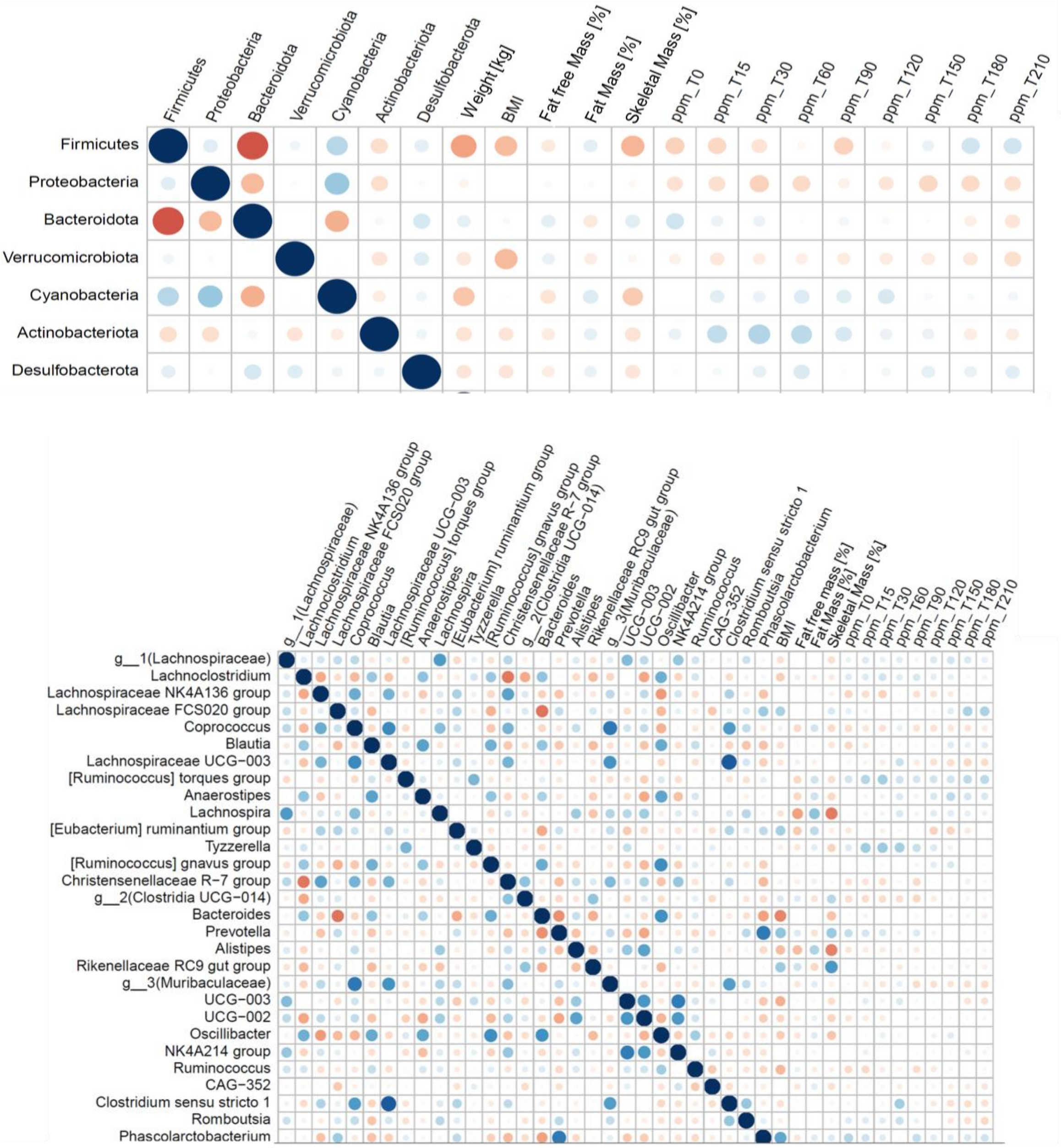


Figure 41: Corrplot explains the positive and negative correlation between gut microbes and selected clinical variables at the phyla (a) and (b) genus level. This figure was originally published as Namco: a microbiome explorer [367] in microbial genomics journal as an open-access article distributed under the terms of the Creative Commons Attribution License.

4.5.4. Functional analysis

To infer how the M and IM groups' microbial populations differ in terms of function, we referred to Namco's built-in *PICRUSt2* feature. A paired Wilcoxon rank test was utilized to determine statistically significant differences between the M and IM and groups, similar to differential abundance analysis at the taxonomic level. In total, after adjusting for the multiple test, we identified 82 KO terms that showed statistically significant differences across the groups. Most of the 76 KO terms were classified as metabolic (level 1) terms, and these were further distributed into 11 sub-categories, including amino acid metabolism, carbohydrate metabolism, lipid metabolism, energy metabolism (Oxidative phosphorylation), glycan biosynthesis and metabolism and cofactor and vitamin metabolism, and (**Fig 42)**. Notably, the IM groups had a higher abundance of the KO term K00845 (glucokinase), which is involved in the metabolism of amino sugars and nucleotide sugars, than the M groups did. It has been hypothesized in prior research that a high-fiber diet has a beneficial effect on glucose and fat metabolism in humans [450]. Also, the expression of ATP-binding cassette (ABC) transporters K10823, K02018, K15583 and K15580 and was increased in IM, lending credence to the aforementioned findings.

Evidence from the previous study suggests that the Firmicutes phyla which was identified as the highest abundant phyla in the IM group encode ABC transporters which is located in the transport on the bacterial plasma membrane. These glucose transporters are essential for facilitating the transfer of glucose across the plasma membrane [451]. Abundance differences of ABC transporters and glucokinase between the groups are shown in **Fig 43.** *Namco* helped to identify a potential mechanism linking fibre intake to improved liver health by showing that Aspartate aminotransferase (AST), a biomarker for liver damage, was found less in the IM group compared to the M group. Previous researches have shown that high fibre diets decrease AST level. *Namco* also helped to discover twelve other significant pathways between IM and M groups, including those involved in fatty acid elongation, N-acetylglucosamine and N-acetylmannosamine degradation, lipid IVA biosynthesis, O-antigen building blocks biosynthesis, acetylene degradation, polyisoprenoid biosynthesis, urate biosynthesis, Kdo transfer to lipid IVA III, guanosine ribonucleotides de novo biosynthesis, GDP-mannose-derived O-antigen building blocks biosynthesis, UDP-N-acetylglucosamine-derived O-antigen building blocks biosynthesis and thiamin diphosphate biosynthesis I. We also repeated the differential analysis using the *Aldex2* method provided in *Namco* to find out the significance of applying CLR transformation on the microbial functional profiles and no significant KO terms were detected. This indicates that CLR transformation could potentially enhance the specificity in functional analysis, but it could come with the expense of sensitivity. Hence users should carefully consider and evaluate their method of choice while interpreting their results.

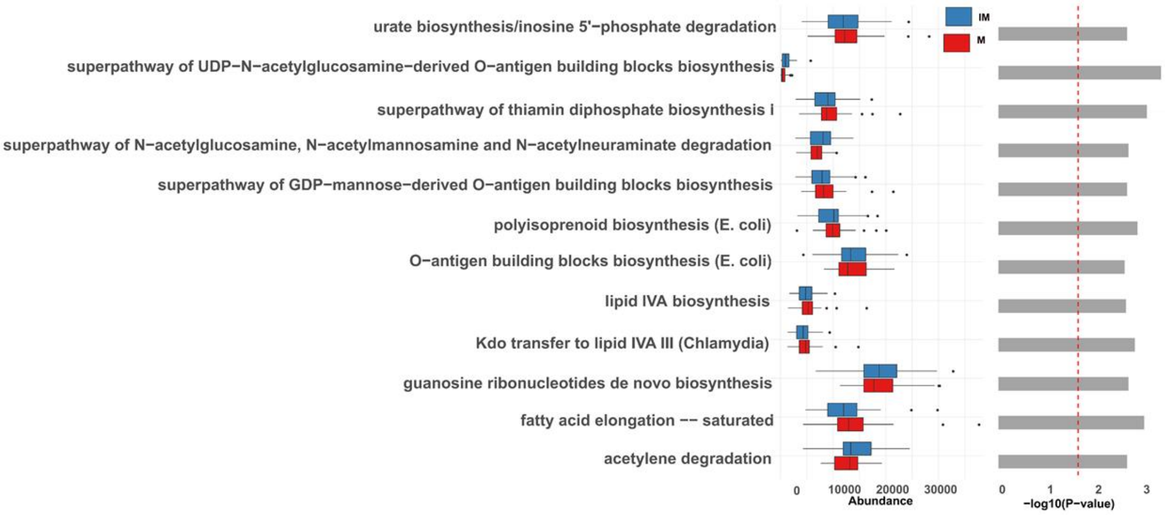


Figure 42: Bar plot represents the abundance difference in the KEGG pathways between intervention and non-intervention groups. The Wilcoxon test was performed to analyze relative abundance, and extended error bar plots were employed to compare the IM and M groups. The rightside of the barplot shows the significant p-value in –log10 scale for each pathways. This figure was originally published as Namco: a microbiome explorer [367] in microbial genomics journal as an open-access article distributed under the terms of the Creative Commons Attribution License.

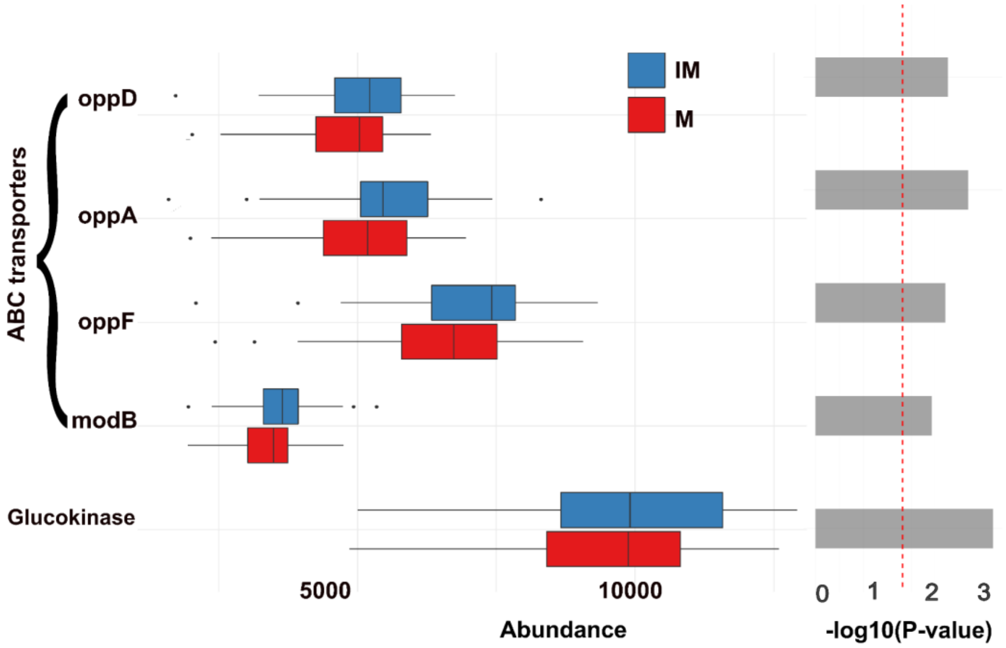


Figure 43: Bar plot depicts a bar plot exhibiting the variation in relative abundance of ABC transporters and glucokinase between the IM and M groups.The p-values were computed using the Wilcoxon Rank test on the abundance values. This figure was originally published as Namco: a microbiome explorer [367] in microbial genomics journal as an open-access article distributed under the terms of the Creative Commons Attribution License.

4.5.5. Network analysis

In order to characterise microbial interactions that take place between the IM and M groups, the interaction networks were generated and analyzed using *Namco* with the *SPRING* [422] mehod for network creation serving as the default option. A criterion of 0.25% for abundance and a cutoff of 10% for prevalence were applied to focus on the most prevalent ASVs. **Fig 44** illustrates the complete microbial genus-level association network between IM and M. The SPRING [422] technique was used as the association measure in the generation of a network at the genus level (replication numbers were set to 100 and nlambda as 50). All four measures of centrality did not show any statistically significant differences. Both of the interaction networks had the same characteristics, and there were no hubs or nodes to be found in either of the groupings. The genera *Eubacterium Coprostanoligenes, Lachnospira Lachnospiraceae UCG-004, CAG-56* and *Ruminococcaceae Incertae Sedis* were responsible for the greatest variances in closeness centrality between the IM and M groups. Only the *Ruminococcaceae Incertae Sedis* family was discovered to be abundant in the IM group. In conclusion, there is not a substantial difference seen between the IM and the M group when comparing the network at the genus level.

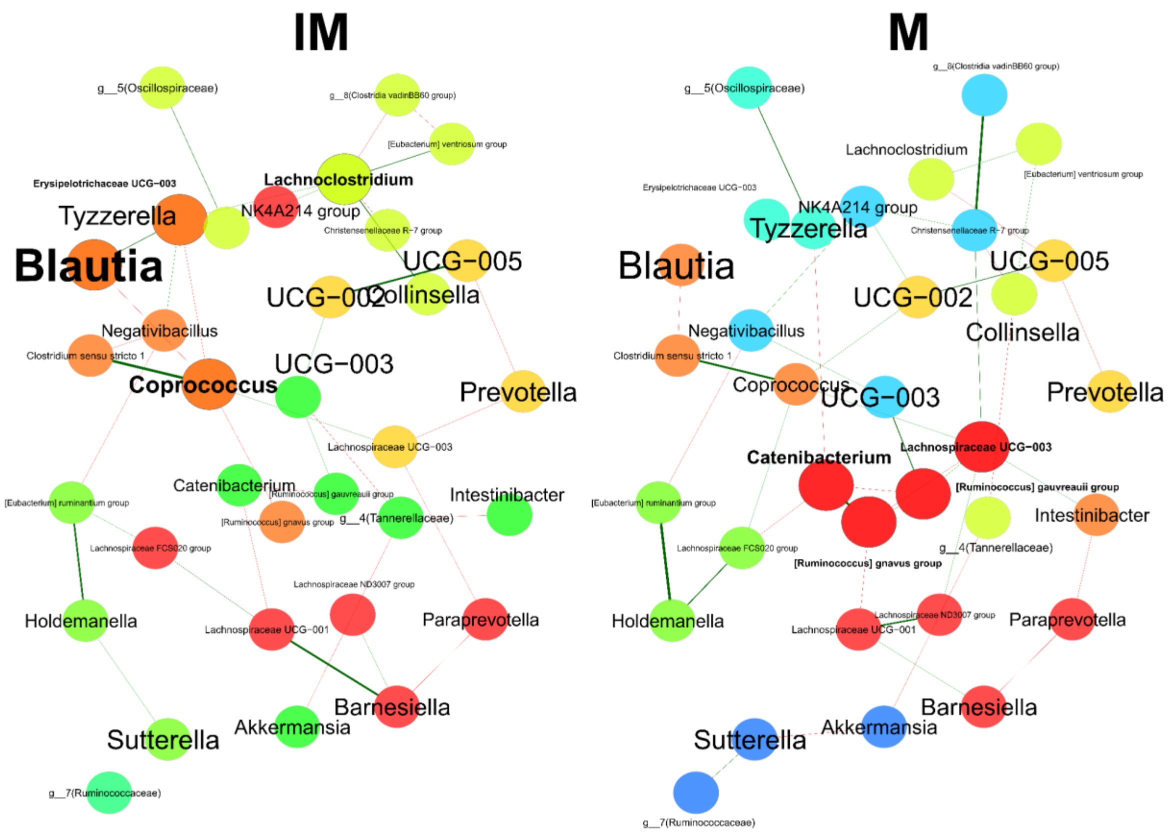


Figure 44: Differential network of genus between intervention groups. It shows the bacterial associations on the genus level for the intervention (IM) and non-intervention (M) groups using the SPRING method. The green edges represent positive associations, while the red edges represent negative associations. The node colours represent clusters determined by using greedy modularity optimization. The networks are shown with only the 50 nodes with the highest degree and 50 edges with the highest weight. This figure was originally published as Namco: a microbiome explorer [367] in microbial genomics journal as an open-access article distributed under the terms of the Creative Commons Attribution License.

4.6. Discussion

There have been numerous algorithms, statistical methods, and software programs to reveal or extract valuable and relevant biological and clinical insights from the from the vast amount of available microbiome sequencing data. To address this, *Namco* was developed as a one-stop data analysis platform for microbial datasets, capable of performing both raw data processing as well as basic and advanced downstream analyses. By integrating existing tools into a cohesive computational workflow, *Namco* enables users to efficiently construct, analyze, and interpret microbial composition without the need for command-line arguments. Namco is accessed through a web browser and does not require the installation of any software packages. Additionally, the platform allows users to save the results of each analysis as an R session that can be resumed at any time, facilitating the sharing of research findings. Namco is available as a Docker image, making it possible to install it locally or on a clinical server behind a firewall for GDPR-compliant sensitive data analysis without uploading to the public Namco instance. Detailed comparisons with other web-based tools have shown that Namco offers unique functions, including time series clustering, function profiling with PICRUSt2, confounder analysis, and topic modeling.

We utilised the dietary intervention study as a case study to elucidate *Namco*'s features. We studied the relationship between a high-fibre diet and gut microbes using both fundamental and advanced analyses in *Namco*. We evaluated the changes in the relative abundance of microbes between the IM and M groups, comparing the most abundant species and also examining intra-individual variation with response to fibre-rich diets.

Following the analysis of the intervention study with respect to microbial abundance and diversity, we discovered that genera which are significantly abundant between IM and M groups, were mostly associated with butyrate synthesis, a SCFA plays a major a role in maintaining the healthy gut through antimicrobial and anti-inflammatory actions [452–454]. *Namco* helped in finding the vastly varied KO terms and pathways but also offered details about the composition of a wide variety of microbes. *Namco* revealed that the presence of the enzyme glucokinase, which catalyses the conversion of amino sugars and nucleotide sugars into energy, was positively correlated with the IM group's performance. Enhanced glucose metabolism is observed in people who consume a high-fibre diet. Several studies also showed that constant fibre intake has been shown to enhance glucose homeostasis [455]. In addition to that, *Namco* found a substantial correlation between the IM subgroup and ABC transporters, which are involved in the translocation of glucose across plasma membranes [451]. Furthermore, *Namco* allowed for the generation of genus-level differential microbial co-occurrence networks, which may be used in studies of microbial interactions. In the end, the inferred associations showed just a little difference in topological properties between the IM and M groups' differential network. In essence, *Namco* supplied a much-required interface for a more natural analysis of data about the microbial community.

4.7. Conclusion and outlook

As the amount of genomic data generated from microbial studies continues to grow, it becomes increasingly important to have software tools that can process and analyze this data efficiently and accurately. There are several challenges associated with analyzing microbial data, such as dealing with high levels of genetic diversity, identifying novel genetic elements, and characterizing the functional roles of genes within microbial communities. To address these challenges, a range of bioinformatics tools have been developed, including sequence alignment software, gene prediction tools, functional annotation software, and metagenomic analysis pipelines. However, many of these tools require significant computational expertise to use effectively, which can be a barrier to entry for researchers who are not trained in bioinformatics.

Therefore, it is essential to develop user-friendly interfaces and workflows that make it easier for researchers to access and utilize these tools. One way to achieve this is by creating web-based interfaces that allow users to upload their data and run pre-configured analysis pipelines without needing to install any software or configure complex computing environments. Many bioinformatics tools are available as web services, making it easy for researchers to access them from anywhere with an internet connection. Another approach is to develop software with intuitive graphical interfaces that allow users to interact with their data in a more visual and intuitive way. This can be particularly helpful for researchers who are new to bioinformatics or who do not have a strong computational background. In summary, extensive and user-friendly bioinformatics tools are critical in microbial data analysis. By developing software that is easy to use and accessible to a broad range of researchers, we can help to accelerate progress in this field and enable more comprehensive and insightful studies of microbial communities.

We present *Namco*, a shiny-based R application made specifically to perform a comprehensive microbiome study by investigating the 16S rRNA gene . To support researchers in efficiently characterising and understanding the structure of the microbial communities included in their data, we have integrated state-of-the-art R packages for both upstream and downstream analysis into a streamlined framework. *Namco* will be further developed to incorporate additional analysis methods and to enable the connection of microbial abundances with other data sources, such as metabolomics and transcriptomics. This indicates that the application will be continuously updated to keep up with the latest developments in microbiome research and to provide researchers with a comprehensive tool for analyzing their data.

5. General discussion

In this thesis, we mainly focused on computational challenges that can be encountered during different stages in the 16S rRNA gene analysis workflow and aimed to provide reliable solutions.

Currently, the predominant method for investigating the microbial communities in various environments ranging from the open ocean to soil to human gut involves utilizing a single gene: 16S rRNA gene [456,457]. However, 16S rRNA gene-based techniques has inherent limitations such as sequencing errors, short read lengths, [458], variability introduced by selecting different variable regions [55,220], and challenges associated with clustering sequences into OTUs [459]. Furthermore, using a single marker gene to evaluate species diversity is challenging due to the prevalence of horizontal gene transfer, the difficulty classifying bacterial species [460], and limited resolution between closely related species when looking at 16S rRNA gene sequences. These factors combine to make the use of a single marker gene for species diversity assessment challenging. One major challenge of most marker gene studies is that they typically focus on only one or a few universal genes, leaving metabolic or other functional capacities of microorganisms underestimated [316]. Overall, the analysis of 16S rRNA gene data requires careful consideration of experimental design, data processing and downstream analysis methods in order to guarantee accurate and dependable results. To our knowledge, there has not been a systematic evaluation of these challenges in high-throughput 16S rRNA gene sequencing nor guidelines on how to address them effectively

In chapter two, we intended to provide the scientific community with up-to-date recommendations for experimental design and data analysis, thus one of our goals was to highlight the contribution that different parameters make to the precision with which taxonomic assignments can be made. In order to draw conclusions beforehand, it is necessary to evaluate the optimal performance of each experimental setting by employing a variety of experimental procedures and settings. Hence, we analysed the impact of choosing different primers, reference databases, clustering methods, and specific pipeline settings using combinations of human stool samples and mock communities with increasing levels of complexity.

We derived the following recommendations from our results. Primers for frequently used V-regions were selected following an analysis of the relevant literature. From our results, we found that the classification of Bacteroidetes, Proteobacteria, and Firmicutes at the phylum level was robust when using different primer combinations except for the V4 region (515F-944R). In contrast, the detection of Actinobacteria, Tenericutes, Lentisphaerae, and Verrucomicrobia changed with the use of different primer pairings, emphasising the importance of selecting the appropriate primer.

As we moved forward, we examined the influence of different bioinformatics pipelines on microbiota composition. After preprocessing raw reads, we generated features [233] by clustering or denoising approaches [94,96]. In this study, results from mock communities revealed that the number of features identified by each method was nearly identical. However, denoising approaches performed better on human data compared to traditional OTU clustering methods. Recent studies often recommend ASV or zOTU approaches, which can detect sequencing errors and provide single nucleotide resolution [231]. We also compared the effect of using reference databases when making taxonomic assignments. SILVA and RDP databases proved the most accurate 16S rRNA gene databases, offering similar consistently good performances compared to GRD, LTP, and GG databases. We no longer recommend using GG since its last update was in 2013; using it may result in false positives [55,70]. As for pipeline parameters, we suggest testing truncated length combinations for both forward and reverse reads for each primer pair used in each study. Finally, we observed that using different mock communities can help detect potential biases in the methods employed to analyse samples in a preferred environment. For instance, if a method fails to accurately detect or quantify one or more organisms present in a mock community, this could indicate that it is unsuitable for the sample being studied. Furthermore, comparing results from the mock community with those from actual samples provides insight into their accuracy and helps identify any potential sources of error.

Chapter three addressed a highly debated topic: how reliable functional profiling from 16S rRNA gene sequences is. Metagenome sequencing can provide valuable information about microbial communities in a sample, including strain-level classification and functional details; however, it is more expensive and technically demanding than other methods for analysing microbial communities - making it less suitable for large-scale research such as population-based studies. Due to the lower costs, 16S rRNA gene profiling remains popular among researchers studying microbial functional profiling due to its much lower costs. This study concluded that 16S data-based functional predictions are often not sensitive enough to detect changes in human health. This was one the most important findings from this study. These functional prediction methods are capable of detecting significant variations between ecological niches. However, they are not an alternative option for Metagenomic approaches and researchers should not be relied on this alone. Researchers who want to develop hypotheses using functional predictions derived from 16S rRNA gene information should be aware that they have limitations and use control measures such as sample label randomization. We recommended *Tax4Fun* and *PICRUSt2* as the best tools. *PanFP* may benefit from a custom copy number database.

Functional prediction tools often perform poorly due to technical and computational biases. One of the primary causes is the absence of sufficient reference genomes [363,364,366] accuracy in functional predictions depends on both number and quality [316]. Functional prediction is the process of inferring the functional capabilities of a microbiome based on the presence or absence of specific genes within their genomes. The more reference genomes that exist for a particular clade of microorganisms, the more confidently one can infer their functional capabilities. However, most microorganisms in nature have yet to be cultivated and thus lack a complete genome. At present, only around 20,000 entire microbial genomes exist that can be used for 16S-based functional profile prediction [461]. Although the number of known 16S rRNA genes [101] is significantly smaller than that, with reference full-length 16S rRNA gene already exceeding 2 million, calibrating functional profiles based on 16S data for environmental microbiomes where reference genomes are less plentiful can prove challenging.

The second important technical bias that could affect the performance of prediction tools is PCR bias. As we have already shown the impact of primer choice in taxonomic classification in the second chapter [55], in return also affects the functional prediction. In order to investigate this, we utilised the simulated datasets from CAMI [350] to evaluate how well various prediction tools and MGS performed in the absence of any technical biases. As was to be expected, the outcome of this analysis was a performance improvement.

Apart from these technical biases, the major concern is that the majority of microbial genes still remain uncharacterized and their functions are unknown. This represents an important challenge in the field of microbial genomics. These uncharacterized genes have the potential to hold immense value for biotechnology and medicine, as they may be used as genome manipulation tools, antimicrobials, delivery systems, and more. Deciphering the function of uncharacterized genes remains a complex and ongoing challenge in microbial genomics. Hence there is a great demand to perform experimental studies to characterise uncharacterized genes or to develop computational predictive models to decipher microbial gene functions.

5.1. Are short-read amplicons suitable for the prediction of microbiome functional potential?

While short reads generated by next-generation sequencing (NGS) platforms are known to be more precise, recent research comparing the output of short-read and long-read sequencing technologies has indicated that the latter offers superior taxonomic classification accuracy at both the genus and species levels [462,463]. Studies including ours have demonstrated that short read amplicon sequencing often fails to differentiate closely related strains and sometimes species. Studies conducted by Ash *et al* [464] and Sergei G *et al* [465] failed to distinguish between clinically important species such as Bacillus anthracis and the Bacillus cereus group. Sequencing full 16S rRNA gene sequences on B. anthracis and B. cereus isolates revealed that only a few had distinguishable 16S rRNA gene sequences, demonstrating that taxonomic resolution of this marker gene may not be accurate for closely related strains [466].

5.2. Full-length 16S rRNA gene sequencing

Results from the first chapter stressed the importance of working with full-length 16S rRNA gene to eliminate region-specific bias. This can be accomplished either through full-length sequencing using third-generation sequencing technologies like Oxford Nanopore or PacBio; additionally, sequencing this commonly used marker for microbial identification [269].

PacBio's Single Molecule Real-Time (SMRT) technology [65] permits full 16S rRNA gene sequencing, providing access to all nine variable regions of the gene. This provides a deeper comprehension of the microbial community being studied and can provide more insights into the connections between different microorganisms and their functional capacities. Furthermore, SMRT can be utilised for full genome sequencing of microbial isolates to provide insight into their genetic diversity, antibiotic resistance genes, virulence factors and other functional genes[467,468]. In conclusion, long-read sequencing technologies such as SMRT and Oxford Nanopore provide a more precise and detailed understanding of microbial communities by enabling the reconstruction of ASVs and full-length sequencing of 16S rRNA gene.

5.3. Shallow metagenomics: Maybe a better alternative to 16S rRNA gene sequencing

Shallow metagenomics [469,470], on the other hand, involves sequencing a small portion of each organism's genome from an entire sample. This provides more in-depth knowledge about microbial communities than 16S rRNA gene sequencing alone can provide, such as functional genes or metabolic pathways. However, it remains more expensive than 16S rRNA gene sequencing and may not offer as comprehensive an overview of the community as deeper sequencing would. The advantages of shallow metagenomics include targeted analysis: By focusing on specific genes or pathways, shallow metagenomics allows for faster and cost-effective identification of microorganisms or genetic markers associated with disease or health.

(2) greater sensitivity: Because shallow metagenomics focuses on specific regions, it can be more sensitive in detecting low-abundance microorganisms or genetic markers. (3) easy comparison: Shallow metagenomics allows for easy comparison of genetic sequences across different samples, which can be useful when identifying similarities or differences between healthy and diseased populations. (4) cost-effectiveness: Shallow metagenomics is typically less costly than shotgun metagenomics as it requires less sequencing. Furthermore, shallow metagenomics allows for the identification of specific species as well as any variations from the reference genome. Shallow metagenomics can also be performed using deeper sequencing mode, known as deep metagenomics. While this provides even more detailed information on the microbiome, the cost and computational demand will be higher. Furthermore, shallow metagenomics also has limitations such as missing or underrepresenting certain microorganisms or genetic markers that may be essential for understanding the sample.

5.4. Employing multi-omics approaches toward microbiome research

Measurement of microbial function requires microbes that are functionally active during sampling with an adequate population size and sampling method, or can be accessible for functional measurements. As not all microbes meet these criteria, real-time PCR may help quantify their genes regarding their functions; however, it must be remembered that genetic potentials do not always accurately reflect actual microbial activities [283,471]. In this context, multi-omics approaches such as transcriptomics, proteomics and metabolomics offer a powerful way to explore microbial functions.

One major limitation of shotgun metagenomics is its inability to differentiate between active and inactive microbiomes. Unfortunately, it can be difficult to differentiate between microbiome members that are contributing to ecosystem behaviour and those simply present, waiting for more favourable conditions. Even if a gene is present, that does not guarantee its expression. Consequently, direct measurement of proteins and/or transcripts through metaproteomics or etatranscriptomics has become an increasingly useful addition for metagenomics [472]. Metatranscriptomics utilises similar strategies as metagenomics, so these same tools can be utilised for the analysis pipeline as well [473]. This straightforward technique assists in combining metagenomics and metatranscriptomics to explore microbial functional potentials [290,291]. Their combination not only facilitates the assembly of microbial genomes and the prediction of gene functions [474–476], but it also has the potential to identify which genes are up/down regulated under particular circumstances. Furthermore, identifying active transcripts within a genome can differentiate metabolically active bacteria from inactive strain [477].

Metaproteomics is another useful omics approach for studying microbial functions. Unlike metagenomics and metatranscriptomics, this omics utilizes high-resolution mass spectrometry (MS) to quantify expressed proteins. Metaproteomics tools such as *Galaxy-P* [478], *MetaProteomeAnalyzer* [479] *and MetaLab* [480] enabled the rapid identification of over 50,000 unique microbiome protein compositions from a single study [481–483]. Metaproteomics offers one major advantage, enabling it to identify and quantify proteins from all organisms present in a microbiome sample - irrespective of their phylogenetic relationship. This includes both microbial and host proteins, which can provide important insights into the interactions between host and microbiome. Metaproteomics can also be employed to investigate the functional activity of the microbiome, by identifying and quantifying enzymes, transporters, and other proteins involved in key metabolic processes. Furthermore, it has applications in host-microbe interactions as well, measuring host proteins involved in immune response, communication between them, and nutrient uptake [484].

Combining the results from these various omic techniques with metagenomics can provide a deeper insight into microbial functioning. For instance, metagenomics can discover genes participated in specific metabolic pathways while transcriptomics reveals which are currently being expressed, proteomics shows which of those genes have been translated to proteins and metabolomics shows which proteins are active and producing metabolic products. This multidimensional data helps us better comprehend how microbes adapt to various environmental conditions, interact with their hosts, and perform specific tasks.

Bioinformatics researchers are actively engaged in developing tools for integrative analysis and visualisation of multi-omics datasets. Some examples include MiBiOmics [485], MOFA+ [486,487], DIABLO [488] and gNOMO [489]. These tools usually employ machine learning and statistical techniques to detect patterns and relationships among various types of omics data, then provide visualisation tools to assist users in understanding the results. Studies using an integrated approach have sought to uncover a potential connection between microbiome and metabolites in diseases. For instance, Ali, R *et al* (2022) [490] demonstrated that regulation of fatty acid metabolism is disrupted in patients with hepatitis C when their gut-liver axis is disrupted. The authors speculate that this discovery could offer new insight into the causes of hepatitis C infection and may pave the way to the identification of new therapeutic targets. Yu, Zheng, et al (2022) [491] combined metabolomics data and 16S rRNA gene sequencing to examine the role of the gut microbiome and serum metabolome in polycystic ovary syndrome pathophysiology.

5.5. Adaptation of machine learning models coupled with other models in microbiome studies

Our understanding of how the microbiome functions in health and disease has the potential to be greatly improved by machine learning and text mining techniques like natural language processing (NLP). This could lead to more precise diagnostic and treatment strategies, propelling microbiome research forward.

For instance, deciphering the function of microbial genes necessitates extensive experimental and computational analysis, including an exploration of their genomic context, protein structure, and interactions with other genes and proteins. Machine learning and deep learning methods, combined with text mining, have the potential to drastically improve our ability to predict gene function and identify new targets for study and intervention. These results indicate that combining language models with microbial genomics is a promising approach for uncovering gene functions in microbes. This study emphasises the potential of NLP techniques in microbiome research, suggesting they could facilitate our comprehension of the intricate interactions between microbes and their hosts. Miller, D et al [492] used Natural Language Processing and a neural network model to predict microbial gene functions. They were successful in classifying 56,617 genes among 444,521 uncategorized ones. This study demonstrated the power of combining microbial genomics with language models to uncover gene functions in microbes. The authors believe this approach offers a promising avenue for discovering new targets and functions within microbiome research. Furthermore, this work emphasises the significance of applying machine learning and NLPs techniques when analysing microbial genomic data.

Another text mining approach which is being used in microbiome research is topic modelling [392,493,494]. They are used to identify underlying topics in a corpus of documents by grouping words into topics based on their co-occurrence patterns. These topics can then be utilized to summarize and comprehend the contents of the corpus. This same model can also be applied to microbiome data, where "documents" refer to samples while "words" correspond to taxa or OTUs. Our tool *Namco* makes this model available so researchers can quickly identify dominant microbial communities within samples and understand their relationship to various biological and environmental variables. Studies such as Movassagh, M.*et al* (2021) [495], Tataru, C. *et al* (2022) [493] and Xiong, X *et al* (2022) [496] have already shown the potential application of topic modeling in microbiome research. Hence this is also integrated in our tool *Namco,* as a part of advanced analysis.

5.6. Extensive bioinformatics tools are needed

Microbiomes are highly interdisciplinary fields requiring expertise across several disciplines, such as microbiology and bioinformatics, genetics, immunology, computer science and statistics. Studying the complex interactions between microbiomes, their hosts, and environments necessitates a sophisticated bioinformatics toolbox. Unfortunately, the tools available for studying various microbes such as fungi, protists and viruses are still relatively limited [497,498], hindering our comprehension of their roles within microbiomes. Additionally, more sophisticated tools are necessary for studying how the components of a microbiome interact with each other, how they evolve over time, and how environmental changes influence them. For microbiome research to reach its full potential, longitudinal study of host-microbiome interactions and integration of environmental elements are necessary. To meet these challenges, collaborations between various fields of expertise are essential as well as interdisciplinary research initiatives.

In the final chapter, we introduced our tool called *Namco*, a user-friendly, shiny R-based web interface with a aim to provide all-in-one solution for 16S rRNA gene analysis. The outcome from the first two chapters also reminded us that there are no guided tools for the users which can give suggestions to the users when and how to use particular analysis. We aimed to develop a user-friendly tool with end-to-end analysis. A plethora of open-source tools are available and each tool has its strengths and weaknesses. One of the main advantages of *Namco* is that it is completely coding-free. Even though many R statistical packages are available to carry out statistical analysis, implementation of R scripts might be challenging for individuals without scripting expertise and bioinformatics experience. Most web-based tools also do not provide end-to-end analysis. The primary drawbacks associated with these tools include: (i) Their exclusive focus on downstream analysis, with little to no attention given to raw data processing; (ii) their reliance on standard statistical methods, which may not be adequate for analyzing more intricate data sets; (iii) a lack of support for functional profiling or utilization of outdated approaches; (iv) inadequate attention paid to confounder analysis and the unique requirements of time-series or longitudinal studies; and (v) limited support for advanced machine learning techniques, as well as the creation of microbial association and differential networks across various taxonomic ranks.

A dietary intervention study was used to help us understand the features of *Namco*. We performed both fundamental and advanced analyses in *Namco* in order to examine the relationship between consuming high fibre diets and gut microbiome composition. We compared the abundances of taxa in the IM group and examined intra-individual variation in gut microbiome in relation to fibre-rich diets. Our findings showed that the IM and the M groups had significant differences in genus that were involved in the production and use of butyrate. Butyrate is a SCFA, which helps in maintaining the gut's homeostasis through anti-inflammatory and antimicrobial actions [454]. *Namco* provided information on the difference in microbial composition and also helped in identifying the most significantly different metagenome and pathways. Overall, *Namco* fills a gap in the market by offering an interface that makes it easier to analyse data about microbial communities.

6. Future perspective of the thesis

My thesis focused primarily on solving computational problems that 16S rRNA gene analyses face and provided guidelines for selecting primers and pipelines. The second part of my thesis, which was also based on the 16S-rRNA gene sequences, did a benchmark analysis and found that *Tax4Fun2* as well as *PICRUSt2* performed better even though they were not true to metagenomics' microbial functions. The outcome could be improved by integrating functional profiles from both these tools. Hence, the future perspective of my thesis would be integrating *Tax4Fun2* and *PICRUSt2* into *Namco* to create a pipeline that facilitates seamless data transition between the two tools. This would enable users to easily compare results obtained from both applications, taking advantage of each one's distinct features. The pipeline should also include data preprocessing and postprocessing steps to guarantee that data is in the appropriate format for each tool, and that results are easily interpretable. Additionally, statistical analysis should be performed on the results to assess the efficiency of both the pipeline and tools. The KEGG Orthology database, which is widely used as a reference source for functional annotation, has undergone multiple updates and revisions since the development of tools. These updates may cause discrepancies in the functional annotations assigned to different genes and organisms, which in turn impacts the accuracy and comparability of functional profiles predicted by these tools. Therefore, when comparing functional profiles across different tools, it is essential to take into account the reference databases and annotation methods each tool utilises. We could only conclude that 16S functional prediction tools are insensitive to detecting functional differences in microbial genes associated with human disease conditions. This leaves open the opportunity to further assess their accuracy in other areas.

Constructing a machine learning model that can accurately calibrate functional profiles for numerous 16S-amplicon samples with only a limited number of WGS-16S amplicon paired samples as training data is an invaluable asset for large-scale, function-focused microbiome sequencing initiatives. This strategy would enable the cost-effective sequencing of 16S amplicons while still ensuring the higher precision offered by WGS sequencing for functional reconstruction. A machine learning model can be trained using a supervised learning approach, using pairs of WGS-16S amplicon samples as training data. With this information, the model could be programmed to predict microbiome functional profiles based on 16S-amplicon sequencing data. Once trained, the model can be applied to a large number of 16S-amplicon samples in order to anticipate their functional profiles. It is essential to remember that the performance of the model relies on the quality of training data and similarity between training samples and those to be analysed. Therefore, it is imperative to carefully select training samples which are representative of the population being analysed.

Machine learning models can also be designed to address PCR bias by targeting blind spots of different primer pairs or correcting for systematic differences caused by differences in DNA extraction protocols. For instance, if a certain primer pair fails to amplify DNA from certain samples, machine learning algorithms could detect the sequence or other characteristics preventing the primer from binding effectively. Machine learning approaches typically utilize large datasets with both positive and negative controls, as well as samples with known biases or other confounding factors. To accomplish this goal, machine learning approaches typically rely on large datasets. By analyzing these datasets, the algorithm can learn to recognize underlying patterns and relationships that could be used to correct biases in future experiments. Machine learning algorithms can also be employed to detect and correct for biases caused by differences in DNA extraction protocols or other experimental variables. Kayama, K et al (2021 [499] recently used recurrent neural network (RNN) prediction to replace preliminary PCR experimentation.

Namco can be further developed to enable novel analysis techniques and the correlation of microbial abundances with other data sources like metabolomics or transcriptomics. One way to integrate metabolomics and metatranscriptomics data with *Namco* would be by providing the ability to import and visualise these data alongside microbial abundance data. To achieve this, data import and visualisation modules tailored specifically for metabolomics and metatranscriptomics data could be added. Additionally, incorporating existing multi-omics statistical frameworks like MOFA+, DIABLO and gNOMO into the *Namco* framework or developing new statistical methods to correlate microbial abundances with metabolomics and metatranscriptomics data could offer a deeper insight into the microbiome community and its interactions with the host. Another way to enhance *Namco* is to incorporate other analysis techniques such as machine learning and graph-based methods to detect key features and correlations in the data. Doing so could provide new insights into the microbial community and its interactions with the host. *Namco* should strive to make itself user-friendly, offering appropriate documentation and tutorials so users can understand the new features and how to utilise them efficiently. Furthermore, providing support and regular updates to the software will guarantee it remains a valuable tool for microbial community analysis.

Overall, while 16S rRNA gene sequencing will likely remain a useful method for microbial community analysis, other methods such as metagenomics and metatranscriptomics that provide higher resolution have become increasingly well received. In five or ten years, it appears likely that 16S rRNA gene sequencing will remain an effective tool for microbial community analysis, however, it may not be the only approach available.

Even though there is extensive ongoing research being carried out in the microbiome field, there are numerous challenges and unanswered questions, leaving plenty of research opportunities open for exploration. Some of the challenges, which I have outlined in each chapter, may be completely overcome with advanced sequencing/ machine learning methods, while others may persist. As stated in the previous section, machine learning/ AI based approaches can be developed to improve the accuracy of PCR amplification. Example for the latter, reference databases will always present challenges related to completeness, accuracy and consistency; hence it is necessary for periodic benchmark studies like mine to identify high quality databases.

Also, it is possible that the compositional nature microbiome data will never be completely resolved. There are still efforts to improve microbiome studies' reliability and reproducibility by developing new analytical methods. Single-cell sequencing technology may be a viable option. This allows for the determination of absolute abundances within samples. However, there is a constant need for a comprehensive benchmark of 16S rRNA gene analysis tools or metagenomics tools to assess their performance and benchmark them against other tools. Hence I strongly believe my work will act as a baseline to carry out the similar benchmark analysis to weigh the advantages and drawbacks of different approaches before selecting which one best suits their research question and objectives.

References

1. Ridgman, W. J. Fungi in Biological Control Systems. Edited by M. N. Burge. x 269 pages. Manchester: Manchester University Press. 1988. Price (hard covers) £39.50. ISBN 0 7190 1979 6. *The Journal of Agricultural Science* vol. 113 124–124 Preprint at https://doi.org/10.1017/s0021859600084744 (1989).
2. Wong, P. T. W. Book Review - Fungi in Biological Control Systems. Edited by M.N. Burge. Manchester University Press, Manchester and New York, 1988. ISBN 0 7190 1979 6. *Australasian Plant Pathology* vol. 18 107 Preprint at https://doi.org/10.1071/app9890106c (1989).
3. Marchesi, J. R. & Ravel, J. The vocabulary of microbiome research: a proposal. *Microbiome* **3**, 31 (2015).
4. Lloyd-Price, J., Abu-Ali, G. & Huttenhower, C. The healthy human microbiome. *Genome Med.* **8**, 51 (2016).
5. Thursby, E. & Juge, N. Introduction to the human gut microbiota. *Biochem. J* **474**, 1823–1836 (2017).
6. Gill, S. R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312**, 1355–1359 (2006).
7. Bäckhed, F., Ley, R. E., Sonnenburg, J. L., Peterson, D. A. & Gordon, J. I. Host-bacterial mutualism in the human intestine. *Science* **307**, 1915–1920 (2005).
8. Aagaard, K. *et al.* The placenta harbors a unique microbiome. *Sci. Transl. Med.* **6**, 237ra65 (2014).
9. Zakis, D. R. *et al.* The evidence for placental microbiome and its composition in healthy pregnancies: A systematic review. *J. Reprod. Immunol.* **149**, 103455 (2022).
10. Olaniyi, K. S., Moodley, J., Mahabeer, Y. & Mackraj, I. Placental Microbial Colonization and Its Association With Pre-eclampsia. *Front. Cell. Infect. Microbiol.* **10**, 413 (2020).
11. Bäckhed, F. Programming of host metabolism by the gut microbiota. *Ann. Nutr. Metab.* **58 Suppl 2**, 44–52 (2011).
12. Dethlefsen, L. & Relman, D. A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl**

1, 4554–4561 (2011).

13. Claesson, M. J. *et al.* Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl 1**, 4586–4591 (2011).
14. Lynch, S. V., Ng, S. C., Shanahan, F. & Tilg, H. Translating the gut microbiome: ready for the clinic? *Nat. Rev. Gastroenterol. Hepatol.* **16**, 656–661 (2019).
15. Mörbe, U. M. *et al.* Human gut-associated lymphoid tissues (GALT); diversity, structure, and function. *Mucosal Immunol.* **14**, 793–802 (2021).
16. Tulic, M. K., Piche, T. & Verhasselt, V. Lung-gut cross-talk: evidence, mechanisms and implications for the mucosal inflammatory diseases. *Clin. Exp. Allergy* **46**, 519–528 (2016).
17. Leiva-Gea, I. *et al.* Gut Microbiota Differs in Composition and Functionality Between Children With Type 1 Diabetes and MODY2 and Healthy Control Subjects: A Case-Control Study. *Diabetes Care* **41**, 2385–2395 (2018).
18. Sanchez-Rodriguez, E. *et al.* The Gut Microbiota and Its Implication in the Development of Atherosclerosis and Related Cardiovascular Diseases. *Nutrients* **12**, (2020).
19. Glassner, K. L., Abraham, B. P. & Quigley, E. M. M. The microbiome and inflammatory bowel disease. *J. Allergy Clin. Immunol.* **145**, 16–27 (2020).
20. Manichanh, C., Borruel, N., Casellas, F. & Guarner, F. The gut microbiota in IBD. *Nat. Rev. Gastroenterol. Hepatol.* **9**, 599–608 (2012).
21. Carding, S., Verbeke, K., Vipond, D. T., Corfe, B. M. & Owen, L. J. Dysbiosis of the gut microbiota in disease. *Microb. Ecol. Health Dis.* **26**, 26191 (2015).
22. van Vliet, M. J., Harmsen, H. J. M., de Bont, E. S. J. M. & Tissing, W. J. E. The role of intestinal microbiota in the development and severity of chemotherapy-induced mucositis. *PLoS Pathog.* **6**, e1000879 (2010).
23. Rodríguez, J. M. *et al.* The composition of the gut microbiota throughout life, with an emphasis on early life. *Microb. Ecol. Health Dis.* **26**, 26050 (2015).
24. Koenig, J. E. *et al.* Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl 1**, 4578–4585 (2011).
25. Collado, M. C., Rautava, S., Aakko, J., Isolauri, E. & Salminen, S. Human gut colonisation may be

initiated in utero by distinct microbial communities in the placenta and amniotic fluid. *Sci. Rep.* **6**, 23129 (2016).

26. Dominguez-Bello, M. G. *et al.* Partial restoration of the microbiota of cesarean-born infants via vaginal microbial transfer. *Nat. Med.* **22**, 250–253 (2016).

27. Tamburini, S., Shen, N., Wu, H. C. & Clemente, J. C. The microbiome in early life: implications for health outcomes. *Nat. Med.* **22**, 713–722 (2016).

28. Gomez de Agüero, M. *et al.* The maternal microbiota drives early postnatal innate immune development. *Science* **351**, 1296–1302 (2016).

29. Arboleya, S. *et al.* C-section and the Neonatal Gut Microbiome Acquisition: Consequences for Future Health. *Ann. Nutr. Metab.* **73 Suppl 3**, 17–23 (2018).

30. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).

31. Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.* **14**, 20–32 (2016).

32. de Menezes, E. W., Giuntini, E. B., Dan, M. C. T., Sardá, F. A. H. & Lajolo, F. M. Codex dietary fibre definition **-** Justification for inclusion of carbohydrates from 3 to 9 degrees of polymerisation. *Food Chem.* **140**, 581–585 (2013).

33. Ilyés, T., Silaghi, C. N. & Crăciun, A. M. Diet-Related Changes of Short-Chain Fatty Acids in Blood and Feces in Obesity and Metabolic Syndrome. *Biology* **11**, (2022).

34. Murphy, E. A., Velazquez, K. T. & Herbert, K. M. Influence of high-fat diet on gut microbiota: a driving force for chronic disease risk. *Curr. Opin. Clin. Nutr. Metab. Care* **18**, 515–520 (2015).

35. Goodrich, J. K. *et al.* Human genetics shape the gut microbiome. *Cell* **159**, 789–799 (2014).

36. Blekhman, R. *et al.* Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* **16**, 191 (2015).

37. Katz, K. *et al.* The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res.* **50**, D387–D390 (2022).

38. Gevers, D. *et al.* The Human Microbiome Project: a community resource for the healthy human microbiome. *PLoS Biol.* **10**, e1001377 (2012).

39. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).

40. Sinha, R., Abnet, C. C., White, O., Knight, R. & Huttenhower, C. The microbiome quality control project: baseline study design and future directions. *Genome Biol.* **16**, 276 (2015).

41. Santiago, A. *et al.* Processing faecal samples: a step forward for standards in microbial community analysis. *BMC Microbiol.* **14**, 112 (2014).

42. McDonald, D. *et al.* American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* **3**, (2018).

43. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).

44. Erickson, A. R. *et al.* Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS One* **7**, e49138 (2012).

45. Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).

46. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).

47. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).

48. Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* **16**, 276–289 (2014).

49. Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).

50. Zhou, W. *et al.* Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature* **569**, 663–671 (2019).

51. Leinonen, R., Sugawara, H., Shumway, M. & International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res.* **39**, D19–21 (2011).

52. Tryka, K. A. *et al.* NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids*

Research vol. 42 D975–D979 Preprint at https://doi.org/10.1093/nar/gkt1211 (2014).

53. Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project. *Nature* **569**, 641–648 (2019).

54. Kittelmann, S. *et al.* Simultaneous amplicon sequencing to explore co-occurrence patterns of bacterial, archaeal and eukaryotic microorganisms in rumen microbial communities. *PLoS One* **8**, e47879 (2013).

55. Abellan-Schneyder, I. *et al.* Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing. *mSphere* **6**, (2021).

56. Tian, R.-M., Cai, L., Zhang, W.-P., Cao, H.-L. & Qian, P.-Y. Rare Events of Intragenus and Intraspecies Horizontal Transfer of the 16S rRNA Gene. *Genome Biol. Evol.* **7**, 2310–2320 (2015).

57. Johnson, J. S. *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **10**, 5029 (2019).

58. Chen, L. *et al.* Rapid Sanger sequencing of the 16S rRNA gene for identification of some common pathogens. *PLoS One* **9**, e88886 (2014).

59. Barb, J. J. *et al.* Development of an Analysis Pipeline Characterizing Multiple Hypervariable Regions of 16S rRNA Using Mock Samples. *PLoS One* **11**, e0148047 (2016).

60. Tremblay, J. *et al.* Primer and platform effects on 16S rRNA tag sequencing. *Front. Microbiol.* **6**, 771 (2015).

61. Cai, L., Ye, L., Tong, A. H. Y., Lok, S. & Zhang, T. Biased diversity metrics revealed by bacterial 16S pyrotags derived from different primer sets. *PLoS One* **8**, e53649 (2013).

62. Srinivasan, R. *et al.* Use of 16S rRNA gene for identification of a broad range of clinically relevant bacterial pathogens. *PLoS One* **10**, e0117617 (2015).

63. Wagner, J. *et al.* Evaluation of PacBio sequencing for full-length bacterial 16S rRNA gene classification. *BMC Microbiol.* **16**, 274 (2016).

64. Chakravorty, S., Helb, D., Burday, M., Connell, N. & Alland, D. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J. Microbiol. Methods* **69**, 330–339 (2007).

65. Ardui, S., Ameur, A., Vermeesch, J. R. & Hestand, M. S. Single molecule real-time (SMRT)

sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res.* **46**, 2159–2168 (2018).

66. Jones, C. B., White, J. R., Ernst, S. E., Sfanos, K. S. & Peiffer, L. B. Incorporation of Data From Multiple Hypervariable Regions when Analyzing Bacterial 16S rRNA Gene Sequencing Data. *Front. Genet.* **13**, 799615 (2022).
67. Jeong, J. *et al.* The effect of taxonomic classification by full-length 16S rRNA sequencing with a synthetic long-read technology. *Sci. Rep.* **11**, 1727 (2021).
68. Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V. & Polz, M. F. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *J. Bacteriol.* **186**, 2629–2635 (2004).
69. Cole, J. R. *et al.* Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42**, D633–42 (2014).
70. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
71. Sikolenko, M. A. & Valentovich, L. N. RiboGrove: a database of full-length prokaryotic 16S rRNA genes derived from completely assembled genomes. *Res. Microbiol.* **173**, 103936 (2022).
72. Verbeke, T. J. *et al.* Predicting relatedness of bacterial genomes using the chaperonin-60 universal target (cpn60 UT): application to Thermoanaerobacter species. *Syst. Appl. Microbiol.* **34**, 171–179 (2011).
73. Case, R. J. *et al.* Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl. Environ. Microbiol.* **73**, 278–288 (2007).
74. Walsh, D. A., Bapteste, E., Kamekura, M. & Doolittle, W. F. Evolution of the RNA Polymerase B′ Subunit Gene (rpoB′) in Halobacteriales: a Complementary Molecular Marker to the SSU rRNA Gene. *Mol. Biol. Evol.* **21**, 2340–2351 (2004).
75. Hadziavdic, K. *et al.* Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. *PLoS One* **9**, e87624 (2014).
76. Stoeck, T. *et al.* Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol. Ecol.* **19 Suppl 1**, 21–31 (2010).
77. Hu, S. K. *et al.* Estimating protistan diversity using high-throughput sequencing. *J. Eukaryot.*

Microbiol. **62**, 688–693 (2015).

78. Choi, J. & Park, J. S. Comparative analyses of the V4 and V9 regions of 18S rDNA for the extant eukaryotic community using the Illumina platform. *Sci. Rep.* **10**, 6519 (2020).
79. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
80. Wensel, C. R., Pluznick, J. L., Salzberg, S. L. & Sears, C. L. Next-generation sequencing: insights to advance clinical investigations of the microbiome. *J. Clin. Invest.* **132**, (2022).
81. Honaas, L. A., Altman, N. S. & Krzywinski, M. Study Design for Sequencing Studies. *Methods in Molecular Biology* 39–66 Preprint at https://doi.org/10.1007/978-1-4939-3578-9_3 (2016).
82. Vandeputte, D. *et al.* Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507–511 (2017).
83. Kadam, P. & Bhalerao, S. Sample size calculation. *Int. J. Ayurveda Res.* **1**, 55–57 (2010).
84. Goodrich, J. K. *et al.* Conducting a microbiome study. *Cell* **158**, 250–262 (2014).
85. Martin, T. C., Visconti, A., Spector, T. D. & Falchi, M. Conducting metagenomic studies in microbiology and clinical research. *Appl. Microbiol. Biotechnol.* **102**, 8629–8646 (2018).
86. Laukens, D., Brinkman, B. M., Raes, J., De Vos, M. & Vandenabeele, P. Heterogeneity of the gut microbiome in mice: guidelines for optimizing experimental design. *FEMS Microbiol. Rev.* **40**, 117–132 (2016).
87. Raising standards in microbiome research. *Nat Microbiol* **1**, 16112 (2016).
88. Nekrutenko, A. & Taylor, J. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nat. Rev. Genet.* **13**, 667–672 (2012).
89. Kunin, V., Engelbrektson, A., Ochman, H. & Hugenholtz, P. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ. Microbiol.* **12**, 118–123 (2010).
90. Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L. & Welch, D. M. Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.* **8**, R143 (2007).
91. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).

92. Bolyen, E. *et al.* Author Correction: Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 1091 (2019).

93. Callahan, B. J., Sankaran, K., Fukuyama, J. A., McMurdie, P. J. & Holmes, S. P. Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses. *F1000Res.* **5**, 1492 (2016).

94. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

95. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).

96. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).

97. Amir, A. *et al.* Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* **2**, (2017).

98. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* **11**, 2639–2643 (2017).

99. FastQC. FastQC: a quality control tool for high throughput sequence data. (2016).

100. Gordon, A. & Hannon, G. J. Fastx-toolkit. *FASTQ/A short-reads preprocessing tools (unpublished*.

101. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–6 (2013).

102. Weisman, D., Yasuda, M. & Bowen, J. L. FunFrame: functional gene ecological analysis pipeline. *Bioinformatics* **29**, 1212–1214 (2013).

103. Özkurt, E. *et al.* LotuS2: An ultrafast and highly accurate tool for amplicon sequencing analysis. *bioRxiv* 2021.12.24.474111 (2021) doi:10.1101/2021.12.24.474111.

104. Lagkouvardos, I. *et al.* IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. *Sci. Rep.* **6**, 33721 (2016).

105. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).

106. Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* **10**, 996–998 (2013).

107. Djemiel, C. *et al.* BIOCOM-PIPE: a new user-friendly metabarcoding pipeline for the characterization of microbial diversity from 16S, 18S and 23S rRNA gene amplicons. *BMC Bioinformatics* **21**, 492 (2020).

108. Tong, W. M. & Chan, Y. GenePiper, a Graphical User Interface Tool for Microbiome Sequence Data Mining. *Microbiol Resour Announc* **9**, (2020).

109. Zhao, Y. *et al.* animalcules: interactive microbiome analytics and visualization in R. *Microbiome* **9**, 76 (2021).

110. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261–5267 (2007).

111. Bedell, J., Korf, I. & Yandell, M. Basic Local Alignment Search Tool. (2003).

112. Chaudhary, N., Sharma, A. K., Agarwal, P., Gupta, A. & Sharma, V. K. 16S classifier: a tool for fast and accurate taxonomic classification of 16S rRNA hypervariable regions in metagenomic datasets. *PLoS One* **10**, e0116106 (2015).

113. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**, e61217 (2013).

114. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol.* **12**, R60 (2011).

115. Parks, D. H., Tyson, G. W., Hugenholtz, P. & Beiko, R. G. STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* **30**, 3123–3124 (2014).

116. Dhariwal, A. *et al.* MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res.* **45**, W180–W188 (2017).

117. Lagkouvardos, I., Fischer, S., Kumar, N. & Clavel, T. Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons. *PeerJ* **5**, e2836 (2017).

118. Huang, Y., Niu, B., Gao, Y., Fu, L. & Li, W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680–682 (2010).

119. Kioukis, A., Pourjam, M., Neuhaus, K. & Lagkouvardos, I. Taxonomy Informed Clustering, an Optimized Method for Purer and More Informative Clusters in Diversity Analysis and Microbiome Profiling. *Front Bioinform* **2**, 864597 (2022).

120. Edgar, R. C. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *bioRxiv* 081257 (2016) doi:10.1101/081257.

121. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194–2200 (2011).

122. Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R. K. & Schmidt, T. M. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Res.* **43**, D593–8 (2015).

123. Hsieh, Y.-P., Hung, Y.-M., Tsai, M.-H., Lai, L.-C. & Chuang, E. Y. 16S-ITGDB: An integrated database for improving species classification of prokaryotic 16S ribosomal RNA sequences. *Front. Bioinform.* **2**, 905489 (2022).

124. Riesenfeld, C. S., Schloss, P. D. & Handelsman, J. Metagenomics: genomic analysis of microbial communities. *Annu. Rev. Genet.* **38**, 525–552 (2004).

125. Aitchison, J. The statistical analysis of compositional data. *J. R. Stat. Soc.* **44**, 139–160 (1982).

126. Aitchison, J. Irregular compositional data. *The Statistical Analysis of Compositional Data* 256–280 Preprint at https://doi.org/10.1007/978-94-009-4109-0_11 (1986).

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

128. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* **8**, 2224 (2017).

129. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27 (2017).

130. McMurdie, P. J. & Holmes, S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* **10**, e1003531 (2014).

131. Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G. & Gloor, G. B. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. *PLoS One* **8**, e67019 (2013).

132. Matchado, M. S. *et al.* Network analysis methods for studying microbial communities: A mini review. *Comput. Struct. Biotechnol. J.* (2021) doi:10.1016/j.csbj.2021.05.001.

133. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**, e1005752 (2017).

134. van den Boogaart, K. G. & Tolosana-Delgado, R. *Analyzing Compositional Data with R*. (Springer Berlin Heidelberg).

135. Greenacre, M., Martínez-Álvaro, M. & Blasco, A. Compositional Data Analysis of Microbiome and Any-Omics Datasets: A Validation of the Additive Logratio Transformation. *Front. Microbiol.* **12**, 727398 (2021).

136. Aitchison, J. & Bacon-Shone, J. Log contrast models for experiments with mixtures. *Biometrika* **71**, 323–330 (1984).

137. Palarea-Albaladejo, J. & Martín-Fernández, J. A. zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems* vol. 143 85–96 Preprint at https://doi.org/10.1016/j.chemolab.2015.02.019 (2015).

138. Gloor, G. ALDEx2: ANOVA-Like Differential Expression tool for compositional data. https://bioconductor.statistik.tu-dortmund.de/packages/3.6/bioc/vignettes/ALDEx2/inst/doc/ALDEx2_vignette.pdf (2017).

139. Dillies, M.-A. *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* **14**, 671–683 (2013).

140. Joseph, N., Paulson, C., Corrada Bravo, H. & Pop, M. Robust methods for differential abundance analysis in marker gene surveys. *Nat. Methods*.

141. Xu, L., Paterson, A. D., Turpin, W. & Xu, W. Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data. *PLOS ONE* vol. 10 e0129606 Preprint at https://doi.org/10.1371/journal.pone.0129606 (2015).

142. Silverman, J. D., Roche, K., Mukherjee, S. & David, L. A. Naught all zeros in sequence count data are the same. *Comput. Struct. Biotechnol. J.* **18**, 2789–2798 (2020).

143. Tsilimigras, M. C. B. & Fodor, A. A. Compositional data analysis of the microbiome:

fundamentals, tools, and challenges. *Ann. Epidemiol.* **26**, 330–335 (2016).

144. Mandal, S. *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **26**, 27663 (2015).

145. Baruzzo, G., Patuzzi, I. & Di Camillo, B. Beware to ignore the rare: how imputing zero-values can improve the quality of 16S rRNA gene studies results. *BMC Bioinformatics* **22**, 618 (2022).

146. Leite, M. F. A. & Kuramae, E. E. You must choose, but choose wisely: Model-based approaches for microbial community analysis. *Soil Biol. Biochem.* **151**, 108042 (2020).

147. Jiang, R., Li, W. V. & Li, J. J. mbImpute: an accurate and robust imputation method for microbiome data. *Genome Biol.* **22**, 192 (2021).

148. Xiao, J., Chen, L., Yu, Y., Zhang, X. & Chen, J. A Phylogeny-Regularized Sparse Regression Model for Predictive Modeling of Microbial Community Data. *Front. Microbiol.* **9**, 3112 (2018).

149. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, e1002687 (2012).

150. Jiang, D. *et al.* Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities. *Front. Genet.* **10**, 995 (2019).

151. Jiang, S. *et al.* HARMONIES: A Hybrid Approach for Microbiome Networks Inference via Exploiting Sparsity. *Front. Genet.* **11**, 445 (2020).

152. Besser, J., Carleton, H. A., Gerner-Smidt, P., Lindsey, R. L. & Trees, E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin. Microbiol. Infect.* **24**, 335–341 (2018).

153. Morgan, X. C. *et al.* Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* **13**, R79 (2012).

154. Barott, K. L. *et al.* Microbial to reef scale interactions between the reef-building coral Montastraea annularis and benthic algae. *Proc. Biol. Sci.* **279**, 1655–1664 (2012).

155. Chaffron, S., Rehrauer, H., Pernthaler, J. & von Mering, C. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.* **20**, 947–959 (2010).

156. Konstantinidis, K. T. & Tiedje, J. M. Genomic insights that advance the species definition for

prokaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 2567–2572 (2005).

157. Douglas, G. M. *et al. PICRUSt2* for prediction of metagenome functions. *Nat. Biotechnol.* **38**, 685–688 (2020).

158. Wemheuer, F. *et al. Tax4Fun2*: prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. *Environ Microbiome* **15**, 11 (2020).

159. Jun, S.-R., Robeson, M. S., Hauser, L. J., Schadt, C. W. & Gorin, A. A. *PanFP*: pangenome-based functional profiles for microbial communities. *BMC Res. Notes* **8**, 479 (2015).

160. Narayan, N. R. *et al.* Piphillin predicts metagenomic composition and dynamics from DADA2-corrected 16S rDNA sequences. *BMC Genomics* **21**, 56 (2020).

161. Wilkinson, T. J. *et al.* CowPI: A Rumen Microbiome Focussed Version of the PICRUSt Functional Inference Software. *Front. Microbiol.* **9**, 1095 (2018).

162. Patumcharoenpol, P. *et al. MetGEM*s Toolbox: Metagenome-scale models as integrative toolbox for uncovering metabolic functions and routes of human gut microbiome. *PLoS Comput. Biol.* **17**, e1008487 (2021).

163. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

164. Heinken, A. *et al.* Systematic assessment of secondary bile acid metabolism in gut microbes reveals distinct metabolic capabilities in inflammatory bowel disease. *Microbiome* **7**, 75 (2019).

165. Sun, S., Jones, R. B. & Fodor, A. A. Inference-based accuracy of metagenome prediction tools varies across sample types and functional categories. *Microbiome* **8**, 46 (2020).

166. Weinstock, G. M. Genomic approaches to studying the human microbiota. *Nature* **489**, 250–256 (2012).

167. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).

168. Verberkmoes, N. C. *et al.* Shotgun metaproteomics of the human distal gut microbiota. *ISME J.* **3**, 179–189 (2009).

169. Chistoserdova, L. Functional Metagenomics: Recent Advances and Future Challenges. *Biotechnol. Genet. Eng. Rev.* **26**, 335–352 (2009).

170. Forsberg, K. J. *et al.* Functional metagenomics-guided discovery of potent Cas9 inhibitors in the human microbiome. *Elife* **8**, (2019).

171. Lam, K. N., Cheng, J., Engel, K., Neufeld, J. D. & Charles, T. C. Current and future resources for functional metagenomics. *Front. Microbiol.* **6**, 1196 (2015).

172. Chevrette, M. G. & Handelsman, J. From Metagenomes to Molecules: Innovations in Functional Metagenomics Unlock Hidden Chemistry in the Human Microbiome. *Biochemistry* vol. 59 729–730 (2020).

173. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).

174. van Kessel, M. A. H. J. *et al.* Complete nitrification by a single microorganism. *Nature* **528**, 555–559 (2015).

175. Norman, J. M. *et al.* Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160**, 447–460 (2015).

176. Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15**, 382–392 (2014).

177. Donia, M. S. *et al.* A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. *Cell* **158**, 1402–1414 (2014).

178. Ma, S. *et al.* Metagenomic analysis reveals oropharyngeal microbiota alterations in patients with COVID-19. *Signal Transduct Target Ther* **6**, 191 (2021).

179. Ke, S., Weiss, S. T. & Liu, Y.-Y. Dissecting the role of the human microbiome in COVID-19 via metagenome-assembled genomes. *Nat. Commun.* **13**, 5235 (2022).

180. Baker, M. De novo genome assembly: what every biologist should know. *Nat. Methods* **9**, 333–337 (2012).

181. Simpson, J. T. & Pop, M. The Theory and Practice of Genome Sequence Assembly. *Annu. Rev. Genomics Hum. Genet.* **16**, 153–172 (2015).

182. Ayling, M., Clark, M. D. & Leggett, R. M. New approaches for metagenome assembly with short reads. *Brief. Bioinform.* **21**, 584–594 (2020).

183. Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment

assembly. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 9748–9753 (2001).

184. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics* **27**, i94–101 (2011).

185. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).

186. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

187. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

188. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).

189. Stewart, E. J. Growing unculturable bacteria. *J. Bacteriol.* **194**, 4151–4160 (2012).

190. Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59–64 (2014).

191. Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99–103 (2013).

192. Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).

193. Segata, N. *et al.* Computational meta'omics for microbial community studies. *Mol. Syst. Biol.* **9**, 666 (2013).

194. Huson, D. H., Mitra, S., Ruscheweyh, H.-J., Weber, N. & Schuster, S. C. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* **21**, 1552–1560 (2011).

195. Brady, A. & Salzberg, S. L. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Methods* **6**, 673–676 (2009).

196. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

197. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes.

Nat. Methods **10**, 1196–1199 (2013).

198. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).

199. Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* **38**, e132 (2010).

200. Abubucker, S. *et al.* Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* **8**, e1002358 (2012).

201. Curtis, H. *et al.* Human Microbiome Project Consortium: Structure, Function and Diversity of the Healthy Human Microbiome. *Nature*.

202. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199–205 (2014).

203. Consortium, T. U. & The UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research* vol. 42 7486–7486 Preprint at https://doi.org/10.1093/nar/gku469 (2014).

204. de Abreu, V. A. C., Perdigão, J. & Almeida, S. Metagenomic Approaches to Analyze Antimicrobial Resistance: An Overview. *Front. Genet.* **11**, 575592 (2020).

205. Liu, B. & Pop, M. ARDB--Antibiotic Resistance Genes Database. *Nucleic Acids Res.* **37**, D443–7 (2009).

206. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* **9**, 207–216 (2015).

207. Love, M., Anders, S. & Huber, W. Differential analysis of count data--the DESeq2 package. *Genome Biol.* **15**, 10–1186 (2014).

208. Dixon, P. VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* vol. 14 927–930 Preprint at https://doi.org/10.1111/j.1654-1103.2003.tb02228.x (2003).

209. Paulson, J. N., Pop, M. & Bravo, H. C. metagenomeSeq: statistical analysis for sparse high-throughput sequencing. Bioconductor Package. (2013).

210. Jimeno, R., Brailey, P. M. & Barral, P. Quantitative Polymerase Chain Reaction-based Analyses of Murine Intestinal Microbiota After Oral Antibiotic Treatment. *Journal of Visualized Experiments*

Preprint at https://doi.org/10.3791/58481 (2018).

211. Zheng, Y. *et al.* Identifying individual-specific microbial DNA fingerprints from skin microbiomes. *Front. Microbiol.* **13**, 960043 (2022).

212. Bhagat, N. *et al.* Microbiome Fingerprint as Biomarker for Geographical Origin and Heredity in Crocus sativus: A Feasibility Study. *Frontiers in Sustainable Food Systems* **5**, (2021).

213. Poretsky, R., Rodriguez-R, L. M., Luo, C., Tsementzi, D. & Konstantinidis, K. T. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PLoS One* **9**, e93827 (2014).

214. Gupta, S. *et al.* Amplicon sequencing provides more accurate microbiome information in healthy children compared to culturing. *Commun Biol* **2**, 291 (2019).

215. Fricker, A. M., Podlesny, D. & Florian Fricke, W. What is new and relevant for sequencing-based microbiome research? A mini-review. *Journal of Advanced Research* vol. 19 105–112 Preprint at https://doi.org/10.1016/j.jare.2019.03.006 (2019).

216. Wilson, M. J., Weightman, A. J. & Wade, W. G. Applications of molecular ecology in the characterization of uncultured microorganisms associated with human disease. *Reviews and Research in Medical Microbiology* **8**, 91 (1997).

217. Bharti, R. & Grimm, D. G. Current challenges and best-practice protocols for microbiome analysis. *Brief. Bioinform.* **22**, 178–193 (2021).

218. Fouhy, F., Clooney, A. G., Stanton, C., Claesson, M. J. & Cotter, P. D. 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiol.* **16**, 123 (2016).

219. Pollock, J., Glendinning, L., Wisedchanwet, T. & Watson, M. The Madness of Microbiome: Attempting To Find Consensus ‘Best Practice’ for 16S Microbiome Studies. *Appl. Environ. Microbiol.* **84**, (2018).

220. Bukin, Y. S. *et al.* The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Sci Data* **6**, 190007 (2019).

221. Knight, R. *et al.* Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* **16**, 410–422 (2018).

222. Kennedy, K., Hall, M. W., Lynch, M. D. J., Moreno-Hagelsieb, G. & Neufeld, J. D. Evaluating bias of illumina-based bacterial 16S rRNA gene profiles. *Appl. Environ. Microbiol.* **80**, 5717–5722 (2014).

223. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).

224. Ahn, J.-H., Kim, B.-Y., Song, J. & Weon, H.-Y. Effects of PCR cycle number and DNA polymerase type on the 16S rRNA gene pyrosequencing analysis of bacterial communities. *J. Microbiol.* **50**, 1071–1074 (2012).

225. Dos Santos, H. R. M., Argolo, C. S., Argôlo-Filho, R. C. & Loguercio, L. L. A 16S rDNA PCR-based theoretical to actual delta approach on culturable mock communities revealed severe losses of diversity information. *BMC Microbiol.* **19**, 74 (2019).

226. Schloss, P. D. & Westcott, S. L. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. *Appl. Environ. Microbiol.* **77**, 3219–3226 (2011).

227. Schloss, P. D., Gevers, D. & Westcott, S. L. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS One* **6**, e27310 (2011).

228. Ye, Y. Identification and Quantification of Abundant Species from Pyrosequences of 16S rRNA by Consensus Alignment. *Proceedings* **2010**, 153–157 (2011).

229. Lawley, B. & Tannock, G. W. Analysis of 16S rRNA Gene Amplicon Sequences Using the QIIME Software Package. *Methods Mol. Biol.* **1537**, 153–163 (2017).

230. Navas-Molina, J. A. *et al.* Advancing our understanding of the human microbiome using QIIME. *Methods Enzymol.* **531**, 371–444 (2013).

231. Caruso, V., Song, X., Asquith, M. & Karstens, L. Performance of Microbiome Sequence Inference Methods in Environments with Varying Biomass. *mSystems* **4**, (2019).

232. Nearing, J. T., Douglas, G. M., Comeau, A. M. & Langille, M. G. I. Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ* **6**, e5364 (2018).

233. Prodan, A. *et al.* Comparing bioinformatic pipelines for microbial 16S rRNA amplicon

sequencing. *PLoS One* **15**, e0227434 (2020).

234. Agnihotry, S., Sarangi, A. N. & Aggarwal, R. Construction & assessment of a unified curated reference database for improving the taxonomic classification of bacteria using 16S rRNA sequence data. *Indian J. Med. Res.* **151**, 93–103 (2020).

235. Cole, J. R. *et al.* The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* **37**, D141–5 (2009).

236. Yoon, S.-H. *et al.* Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int. J. Syst. Evol. Microbiol.* **67**, 1613–1617 (2017).

237. F Escapa, I. *et al.* Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. *Microbiome* **8**, 65 (2020).

238. Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J. & Weightman, A. J. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl. Environ. Microbiol.* **71**, 7724–7736 (2005).

239. Garrity, G. M. *et al.* Taxonomic Outline of the Bacteria and Archaea, Release 7.7 March 6, 2007. Part 7--The Bacteria: Phylum ‘Firmicutes’: Class ‘Clostridia’.

240. *Bergey’s Manual of Systematic Bacteriology*. (Springer New York).

241. Ludwig, W. *et al.* ARB: a software environment for sequence data. *Nucleic Acids Res.* **32**, 1363–1371 (2004).

242. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41**, D36–42 (2013).

243. McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6**, 610–618 (2012).

244. Myer, P. R. *et al.* Classification of 16S rRNA reads is improved using a niche-specific database constructed by near-full length sequencing. *PLoS One* **15**, e0235498 (2020).

245. Munoz, R. *et al.* Release LTPs104 of the All-Species Living Tree. *Syst. Appl. Microbiol.* **34**, 169–170 (2011).

246. Park, S.-C. & Won, S. Evaluation of 16S rRNA Databases for Taxonomic Assignments Using Mock Community. *Genomics Inform.* **16**, e24 (2018).

247. Sierra, M. A. *et al.* The Influences of Bioinformatics Tools and Reference Databases in Analyzing

the Human Oral Microbial Community. *Genes* **11**, (2020).

248. Dueholm, M. S. *et al.* Generation of Comprehensive Ecosystem-Specific Reference Databases with Species-Level Resolution by High-Throughput Full-Length 16S rRNA Gene Sequencing and Automated Taxonomy Assignment (AutoTax). *MBio* **11**, (2020).
249. Meola, M. *et al.* DAIRYdb: a manually curated reference database for improved taxonomy annotation of 16S rRNA gene sequences from dairy products. *BMC Genomics* **20**, 560 (2019).
250. Bokulich, N. A. *et al.* Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* vol. 6 Preprint at https://doi.org/10.1186/s40168-018-0470-z (2018).
251. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* vol. 17 10 Preprint at https://doi.org/10.14806/ej.17.1.200 (2011).
252. Peltzer, A., Straub, D. & Patel, H. *nf-core/ampliseq: Ampliseq Version 1.1.1.* (2019). doi:10.5281/zenodo.3568091.
253. Ewels, P. A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology* vol. 38 276–278 Preprint at https://doi.org/10.1038/s41587-020-0439-x (2020).
254. Jumpstart Consortium Human Microbiome Project Data Generation Working Group. Evaluation of 16S rDNA-based community profiling for human microbiome research. *PLoS One* **7**, e39315 (2012).
255. Human Microbiome Project Consortium. A framework for human microbiome research. *Nature* **486**, 215–221 (2012).
256. Costea, P. I. *et al.* Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**, 1069–1076 (2017).
257. Ma, J. *et al.* Variations of Gut Microbiome Profile Under Different Storage Conditions and Preservation Periods: A Multi-Dimensional Evaluation. *Front. Microbiol.* **11**, 972 (2020).
258. Thaiss, C. A. *et al.* Transkingdom control of microbiota diurnal oscillations promotes metabolic homeostasis. *Cell* **159**, 514–529 (2014).
259. Bellali, S., Lagier, J.-C., Raoult, D. & Bou Khalil, J. Among Live and Dead Bacteria, the Optimization of Sample Collection and Processing Remains Essential in Recovering Gut

Microbiota Components. *Front. Microbiol.* **10**, 1606 (2019).

260. Plummer, E., Twin, J. & Bulach, D. M. A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. *of Proteomics & …*.

261. Meisel, J. S. *et al.* Skin Microbiome Surveys Are Strongly Influenced by Experimental Design. *J. Invest. Dermatol.* **136**, 947–956 (2016).

262. Bjerre, R. D. *et al.* Effects of sampling strategy and DNA extraction on human skin microbiome investigations. *Sci. Rep.* **9**, 17287 (2019).

263. Teng, F. *et al.* Impact of DNA extraction method and targeted 16S-rRNA hypervariable region on oral microbiota profiling. Sci Rep 8: 16321. Preprint at (2018).

264. Xue, Z., Kable, M. E. & Marco, M. L. Impact of DNA Sequencing and Analysis Methods on 16S rRNA Gene Bacterial Community Analysis of Dairy Products. *mSphere* **3**, (2018).

265. Thijs, S. *et al.* Comparative Evaluation of Four Bacteria-Specific Primer Pairs for 16S rRNA Gene Surveys. *Front. Microbiol.* **8**, 494 (2017).

266. Rausch, P. *et al.* Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms. *Microbiome* **7**, 133 (2019).

267. Yang, B., Wang, Y. & Qian, P.-Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics* **17**, 135 (2016).

268. Martijn, J. *et al.* Confident phylogenetic identification of uncultured prokaryotes through long read amplicon sequencing of the 16S-ITS-23S rRNA operon. *Environ. Microbiol.* **21**, 2485–2498 (2019).

269. Karst, S. M. *et al.* Retrieval of a million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias. *Nat. Biotechnol.* **36**, 190–195 (2018).

270. Loit, K. *et al.* Relative Performance of MinION (Oxford Nanopore Technologies) versus Sequel (Pacific Biosciences) Third-Generation Sequencing Instruments in Identification of Agricultural and Forest Fungal Pathogens. *Applied and Environmental Microbiology* vol. 85 Preprint at https://doi.org/10.1128/aem.01368-19 (2019).

271. Almeida, A., Mitchell, A. L., Tarkowska, A. & Finn, R. D. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments.

Gigascience **7**, (2018).

272. Ducarmon, Q. R., Hornung, B. V. H., Geelen, A. R., Kuijper, E. J. & Zwittink, R. D. Toward Standards in Clinical Microbiota Studies: Comparison of Three DNA Extraction Methods and Two Bioinformatic Pipelines. *mSystems* **5**, (2020).
273. Edgar, R. C. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* **34**, 2371–2375 (2018).
274. Thomas, A. M. *et al.* Author Correction: Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 1948 (2019).
275. Poyet, M. *et al.* A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat. Med.* **25**, 1442–1452 (2019).
276. Forster, S. C. *et al.* A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nat. Biotechnol.* **37**, 186–192 (2019).
277. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).
278. Singh, R. K. *et al.* Influence of diet on the gut microbiome and implications for human health. *J. Transl. Med.* **15**, 73 (2017).
279. Rowland, I. *et al.* Gut microbiota functions: metabolism of nutrients and other food components. *Eur. J. Nutr.* **57**, 1–24 (2018).
280. Uebanso, T., Shimohata, T., Mawatari, K. & Takahashi, A. Functional Roles of B-Vitamins in the Gut and Gut Microbiome. *Mol. Nutr. Food Res.* **64**, e2000426 (2020).
281. Ghosh, S., Whitley, C. S., Haribabu, B. & Jala, V. R. Regulation of Intestinal Barrier Function by Microbial Metabolites. *Cell Mol Gastroenterol Hepatol* **11**, 1463–1482 (2021).
282. Kamada, N., Seo, S.-U., Chen, G. Y. & Núñez, G. Role of the gut microbiota in immunity and inflammatory disease. *Nat. Rev. Immunol.* **13**, 321–335 (2013).
283. Heintz-Buschart, A. & Wilmes, P. Human Gut Microbiome: Function Matters. *Trends Microbiol.* **26**, 563–574 (2018).
284. Doolittle, W. F. & Booth, A. It’s the song, not the singer: an exploration of holobiosis and

evolutionary theory. *Biol. Philos.* **32**, 5–24 (2017).

285. Gilbert, J. A. *et al.* Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature* **535**, 94–103 (2016).

286. Gosalbes, M. J. *et al.* Metatranscriptomic approach to analyze the functional human gut microbiota. *PLoS One* **6**, e17447 (2011).

287. Ferrer, M. *et al.* Microbiota from the distal guts of lean and obese adolescents exhibit partial functional redundancy besides clear differences in community structure. *Environ. Microbiol.* **15**, 211–226 (2013).

288. Langille, M. G. I. Exploring Linkages between Taxonomic and Functional Profiles of the Human Microbiome. *mSystems* **3**, (2018).

289. Si, X. *et al.* The importance of accounting for imperfect detection when estimating functional and phylogenetic community structure. *Ecology* **99**, 2103–2112 (2018).

290. Heintz-Buschart, A. *et al.* Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat Microbiol* **2**, 16180 (2016).

291. Franzosa, E. A. *et al.* Relating the metatranscriptome and metagenome of the human gut. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E2329–38 (2014).

292. Shafranskaya, D. *et al.* MetaGT: A pipeline for de novo assembly of metatranscriptomes with the aid of metagenomic data. *Front. Microbiol.* **13**, 981458 (2022).

293. Aguiar-Pulido, V. *et al.* Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis. *Evol. Bioinform. Online* **12**, 5–16 (2016).

294. Moran, M. A. Metatranscriptomics: Eavesdropping on complex microbial communities. *Microbe Wash. DC* **4**, 329–335 (2009).

295. Gilbert, J. A. *et al.* Detection of Large Numbers of Novel Sequences in the Metatranscriptomes of Complex Marine Microbial Communities. *Handbook of Molecular Microbial Ecology II* 277–286 Preprint at https://doi.org/10.1002/9781118010549.ch27 (2011).

296. Poretsky, R. S. *et al.* Analysis of microbial gene transcripts in environmental samples. *Appl. Environ. Microbiol.* **71**, 4121–4126 (2005).

297. Franzosa, E. A. *et al.* Identifying personal microbiomes using metagenomic codes. *Proc. Natl.*

Acad. Sci. U. S. A. **112**, E2930–8 (2015).

298. Laudadio, I. *et al.* Quantitative Assessment of Shotgun Metagenomics and 16S rDNA Amplicon Sequencing in the Study of Human Gut Microbiome. *OMICS* **22**, 248–254 (2018).
299. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
300. Gemayel, K., Lomsadze, A. & Borodovsky, M. MetaGeneMark-2: Improved Gene Prediction in Metagenomes. *bioRxiv* 2022.07.25.500264 (2022) doi:10.1101/2022.07.25.500264.
301. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
302. Rho, M., Tang, H. & Ye, Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* **38**, e191 (2010).
303. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
304. Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* **10**, (2021).
305. Minich, J. J. *et al.* KatharoSeq Enables High-Throughput Microbiome Analysis from Low-Biomass Samples. *mSystems* **3**, (2018).
306. Jervis-Bardy, J. *et al.* Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data. *Microbiome* **3**, 19 (2015).
307. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
308. Pereira-Marques, J. *et al.* Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis. *Front. Microbiol.* **10**, 1277 (2019).
309. Devoid, S., Overbeek, R., DeJongh, M. & Vonstein, V. Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. *: methods and protocols* (2013).
310. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated

orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2018).

311. Galperin, M. Y., Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43**, D261–D269 (2014).

312. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).

313. Peterson, D. *et al.* Comparative Analysis of 16S rRNA Gene and Metagenome Sequencing in Pediatric Gut Microbiomes. *Front. Microbiol.* **12**, 670336 (2021).

314. Qiao, Y. *et al.* Effects of different Lactobacillus reuteri on inflammatory and fat storage in high-fat diet-induced obesity mice model. *J. Funct. Foods* **14**, 424–434 (2015).

315. Wu, G. *et al.* Genomic Microdiversity of Bifidobacterium pseudocatenulatum Underlying Differential Strain-Level Responses to Dietary Carbohydrate Intervention. *MBio* **8**, (2017).

316. Langille, M. G. I. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* **31**, 814–821 (2013).

317. Thornton, J. W. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat. Rev. Genet.* **5**, 366–375 (2004).

318. McLennan, D. A. How to Read a Phylogenetic Tree. *Evolution: Education and Outreach* **3**, 506–519 (2010).

319. Fujimura, K. E. *et al.* Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. *Nat. Med.* **22**, 1187–1191 (2016).

320. Boursier, J. *et al.* The severity of nonalcoholic fatty liver disease is associated with gut dysbiosis and shift in the metabolic function of the gut microbiota. *Hepatology* **63**, 764–775 (2016).

321. Pannaraj, P. S. *et al.* Association Between Breast Milk Bacterial Communities and Establishment and Development of the Infant Gut Microbiome. *JAMA Pediatr.* **171**, 647–654 (2017).

322. Kostic, A. D. *et al.* The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* **17**, 260–273 (2015).

323. Aßhauer, K. P., Wemheuer, B., Daniel, R. & Meinicke, P. Tax4Fun: predicting functional profiles

from metagenomic 16S rRNA data. *Bioinformatics* **31**, 2882–2884 (2015).

324. Sakamoto, T. & Ortega, J. M. Taxallnomy: an extension of NCBI Taxonomy that produces a hierarchically complete taxonomic tree. *BMC Bioinformatics* **22**, 388 (2021).

325. Ortiz-Estrada, Á. M., Gollas-Galván, T., Martínez-Córdova, L. R. & Martínez-Porchas, M. Predictive functional profiles using metagenomic 16S rRNA data: a novel approach to understanding the microbial ecology of aquaculture systems. *Rev. Aquac.* **11**, 234–245 (2019).

326. Djemiel, C. *et al.* Inferring microbiota functions from taxonomic genes: a review. *Gigascience* **11**, (2022).

327. Toole, D. R., Zhao, J., Martens-Habbena, W. & Strauss, S. L. Bacterial functional prediction tools detect but underestimate metabolic diversity compared to shotgun metagenomics in southwest Florida soils. *Appl. Soil Ecol.* **168**, 104129 (2021).

328. Aßhauer, K. P. & Meinicke, P. On the estimation of metabolic profiles in metagenomics. in *German Conference on Bioinformatics 2013* 13 (Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik GmbH, Wadern/Saarbruecken, Germany, 2013).

329. Zheng, W. *et al.* Changes in the soil bacterial community structure and enzyme activities after intercrop mulch with cover crop for eight years in an orchard. *Eur. J. Soil Biol.* **86**, 34–41 (2018).

330. Weng, F. C.-H., Yang, Y.-J. & Wang, D. Functional analysis for gut microbes of the brown tree frog (Polypedates megacephalus) in artificial hibernation. *BMC Genomics* **17**, 1024 (2016).

331. Koo, H. *et al.* Comparison of two bioinformatics tools used to characterize the microbial diversity and predictive functional attributes of microbial mats from Lake Obersee, Antarctica. *J. Microbiol. Methods* **140**, 15–22 (2017).

332. Iwai, S. *et al.* Piphillin: Improved Prediction of Metagenomic Content by Direct Inference from Human Microbiomes. *PLoS One* **11**, e0166104 (2016).

333. He, Y. *et al.* Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat. Med.* **24**, 1532–1535 (2018).

334. Suzuki, M. T. & Giovannoni, S. J. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl. Environ. Microbiol.* **62**, 625–630 (1996).

335. Hong, S., Bunge, J., Leslin, C., Jeon, S. & Epstein, S. S. Polymerase chain reaction primers miss

half of rRNA microbial diversity. *ISME J.* **3**, 1365–1373 (2009).

336. Haas, B. J. *et al.* Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research* vol. 21 494–504 Preprint at https://doi.org/10.1101/gr.112730.110 (2011).

337. Kembel, S. W., Wu, M., Eisen, J. A. & Green, J. L. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput. Biol.* **8**, e1002743 (2012).

338. Louca, S., Doebeli, M. & Parfrey, L. W. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. *Microbiome* **6**, 41 (2018).

339. Rastogi, R., Wu, M., Dasgupta, I. & Fox, G. E. Visualization of ribosomal RNA operon copy number distribution. *BMC Microbiol.* **9**, 208 (2009).

340. Lee, Z. M.-P., Bussema, C., 3rd & Schmidt, T. M. rrnDB: documenting the number of rRNA and tRNA genes in bacteria and archaea. *Nucleic Acids Res.* **37**, D489–93 (2009).

341. Farrelly, V., Rainey, F. A. & Stackebrandt, E. Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. *Applied and Environmental Microbiology* vol. 61 2798–2801 Preprint at https://doi.org/10.1128/aem.61.7.2798-2801.1995 (1995).

342. Venter, J. C. *et al.* Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).

343. Wu, M. & Eisen, J. A. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.* **9**, R151 (2008).

344. Bowman, J. S. & Ducklow, H. W. Microbial Communities Can Be Described by Metabolic Structure: A General Framework and Application to a Seasonally Variable, Depth-Stratified Microbial Community from the Coastal West Antarctic Peninsula. *PLoS One* **10**, e0135868 (2015).

345. Holle, R., Happich, M., Löwel, H., Wichmann, H. E. & MONICA/KORA Study Group. KORA--a research platform for population based health research. *Gesundheitswesen* **67 Suppl 1**, S19–25 (2005).

346. Krawczak, M. *et al.* PopGen: population-based recruitment of patients and controls for the analysis

of complex genotype-phenotype relationships. *Community Genet.* **9**, 55–61 (2006).

347. Reitmeier, S. *et al.* Arrhythmic Gut Microbiome Signatures Predict Risk of Type 2 Diabetes. *Cell Host Microbe* **28**, 258–272.e6 (2020).
348. Rühlemann, M. C. *et al.* Genome-wide association study in 8,956 German individuals identifies influence of ABO histo-blood groups on gut microbiome. *Nat. Genet.* **53**, 147–155 (2021).
349. Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
350. Fritz, A. *et al.* CAMISIM: simulating metagenomes and microbial communities. *Microbiome* **7**, 17 (2019).
351. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
352. Saus, E., Iraola-Guzmán, S., Willis, J. R., Brunet-Vega, A. & Gabaldón, T. Microbiome and colorectal cancer: Roles in carcinogenesis and clinical potential. *Mol. Aspects Med.* **69**, 93–106 (2019).
353. Rebersek, M. Gut microbiome and its role in colorectal cancer. *BMC Cancer* **21**, 1325 (2021).
354. Huang, R. *et al.* Changes of Intestinal Microflora in Colorectal Cancer Patients after Surgical Resection and Chemotherapy. *Comput. Math. Methods Med.* **2022**, 1940846 (2022).
355. Fang, C.-Y. *et al.* Colorectal Cancer Stage-Specific Fecal Bacterial Community Fingerprinting of the Taiwanese Population and Underpinning of Potential Taxonomic Biomarkers. *Microorganisms* **9**, (2021).
356. Dai, Z. *et al.* Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome* **6**, 70 (2018).
357. Coleman, O., Ecker, M. & Haller, D. Dysregulated lipid metabolism in colorectal cancer. *Curr. Opin. Gastroenterol.* **38**, 162–167 (2022).
358. Duan, M. *et al.* Characteristics of gut microbiota in people with obesity. *PLoS One* **16**, e0255446 (2021).
359. Bombin, A., Yan, S., Bombin, S., Mosley, J. D. & Ferguson, J. F. Obesity influences composition of salivary and fecal microbiota and impacts the interactions between bacterial taxa. *Physiol Rep*

10, e15254 (2022).

360. Walker, A. W. *et al.* 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. *Microbiome* **3**, 26 (2015).
361. Morgan, X. C. & Huttenhower, C. Chapter 12: Human microbiome analysis. *PLoS Comput. Biol.* **8**, e1002808 (2012).
362. Su, X., Jing, G., Zhang, Y. & Wu, S. Method development for cross-study microbiome data mining: Challenges and opportunities. *Comput. Struct. Biotechnol. J.* **18**, 2075–2080 (2020).
363. Mukherjee, S. *et al.* 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat. Biotechnol.* **35**, 676–683 (2017).
364. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
365. Magnúsdóttir, S. *et al.* Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat. Biotechnol.* **35**, 81–89 (2017).
366. Chuvochina, M. *et al.* The importance of designating type material for uncultured taxa. *Syst. Appl. Microbiol.* **42**, 15–21 (2019).
367. Dietrich, A. *et al.* Namco: a microbiome explorer. *Microb Genom* **8**, (2022).
368. Ogunrinola, G. A., Oyewale, J. O., Oshamika, O. O. & Olasehinde, G. I. The Human Microbiome and Its Impacts on Health. *Int. J. Microbiol.* **2020**, 8045646 (2020).
369. Liang, D., Leung, R. K.-K., Guan, W. & Au, W. W. Involvement of gut microbiome in human health and disease: brief overview, knowledge gaps and research opportunities. *Gut Pathog.* **10**, 3 (2018).
370. Devaraj, S., Hemarajata, P. & Versalovic, J. The human gut microbiome and body metabolism: implications for obesity and diabetes. *Clin. Chem.* **59**, 617–628 (2013).
371. Sepich-Poore, G. D. *et al.* The microbiome and human cancer. *Science* **371**, (2021).
372. Urbaniak, C. *et al.* The Microbiota of Breast Tissue and Its Association with Breast Cancer. *Appl. Environ. Microbiol.* **82**, 5039–5048 (2016).
373. Morais, L. H., Schreiber, H. L., 4th & Mazmanian, S. K. The gut microbiota-brain axis in behaviour and brain disorders. *Nat. Rev. Microbiol.* **19**, 241–255 (2021).

374. Thomas, S. *et al.* The Host Microbiome Regulates and Maintains Human Health: A Primer and Perspective for Non-Microbiologists. *Cancer Res.* **77**, 1783–1812 (2017).

375. Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K. & Schloss, P. D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* **79**, 5112–5120 (2013).

376. Gao, B. *et al.* An Introduction to Next Generation Sequencing Bioinformatic Analysis in Gut Microbiome Studies. *Biomolecules* **11**, (2021).

377. Sharpton, T. J. An introduction to the analysis of shotgun metagenomic data. *Front. Plant Sci.* **5**, 209 (2014).

378. Durazzi, F. *et al.* Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Sci. Rep.* **11**, 3030 (2021).

379. Kuniyoshi, G. & Laura, M. Comparison between 16S rRNA and shotgun sequencing data for the characterization of the human gut microbiota. *thesis.unipd.it* (2022).

380. Eren, A. M. *et al.* Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J.* **9**, 968–979 (2015).

381. Bokulich, N. A. *et al.* q2-sample-classifier: machine-learning tools for microbiome classification and regression. *J Open Res Softw* **3**, (2018).

382. Bokulich, N. A. *et al.* q2-longitudinal: Longitudinal and Paired-Sample Analyses of Microbiome Data. *mSystems* **3**, (2018).

383. Oksanen, J. *et al.* The vegan package. *Community ecology package* **10**, 719 (2007).

384. Lahti, L. & Shetty, S. Introduction to the microbiome R package. http://bioconductor.statistik.tu-dortmund.de/packages/3.6/bioc/vignettes/microbiome/inst/doc/vignette.html (2018).

385. Oksanen, J., Blanchet, F. G., Friendly, M. & Kindt, R. Package 'Vegan'Title Community Ecology Package Version 2.5-6. 2019. *is a statistical package for R.*

386. Xia, Y. & Sun, J. Hypothesis Testing and Statistical Analysis of Microbiome. *Genes Dis* **4**, 138–148 (2017).

387. Huse, S. M. *et al.* VAMPS: a website for visualization and analysis of microbial population structures. *BMC Bioinformatics* **15**, 41 (2014).

388. Buza, T. M. *et al.* iMAP: an integrated bioinformatics and visualization pipeline for microbiome data analysis. *BMC Bioinformatics* **20**, 374 (2019).

389. Wilke, A. *et al.* The MG-RAST metagenomics database and portal in 2015. *Nucleic Acids Res.* **44**, D590–4 (2016).

390. Su, S.-C., Galvin, J. E., Yang, S.-F., Chung, W.-H. & Chang, L.-C. wiSDOM: a visual and statistical analytics for interrogating microbiome. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab057.

391. McMurdie, P. J. & Holmes, S. Shiny-phyloseq: Web application for interactive microbiome analysis with provenance tracking. *Bioinformatics* **31**, 282–283 (2015).

392. Woloszynek, S. *et al.* Themetagenomics: Exploring Thematic Structure and Predicted Functionality of 16s rRNA Amplicon Data. *bioRxiv* 678110 (2019) doi:10.1101/678110.

393. Mahé, F., Rognes, T., Quince, C., de Vargas, C. & Dunthorn, M. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* **2**, e593 (2014).

394. Puente-Sánchez, F., Aguirre, J. & Parro, V. A novel conceptual approach to read-filtering in high-throughput amplicon sequencing studies. *Nucleic Acids Res.* **44**, e40–e40 (2015).

395. Guiasu, R. C. & Guiasu, S. Weighted Gini-Simpson quadratic index of biodiversity for interdependent species. *Nat. Sci. (Irvine)* **06**, 455–466 (2014).

396. Shannon, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal* vol. 27 623–656 Preprint at https://doi.org/10.1002/j.1538-7305.1948.tb00917.x (1948).

397. Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J. & Knight, R. UniFrac: an effective distance metric for microbial community comparison. *ISME J.* **5**, 169–172 (2011).

398. Haynes, W. Benjamini–Hochberg Method. in *Encyclopedia of Systems Biology* (eds. Dubitzky, W., Wolkenhauer, O., Cho, K.-H. & Yokota, H.) 78–78 (Springer New York, 2013).

399. Wirbel, J. *et al.* Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. *Genome Biol.* **22**, 93 (2021).

400. Clayden, A. Causal relationships in medicine: A practical system for critical appraisal. *Ann. Intern. Med.* **114**, 916 (1991).

401. Vujkovic-Cvijin, I. *et al.* Host variables confound gut microbiota studies of human disease. *Nature*

587, 448–454 (2020).

402. Asnicar, F. *et al.* Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat. Med.* **27**, 321–332 (2021).

403. Polster, S. P. *et al.* Permissive microbiome characterizes human subjects with a neurovascular disease cavernous angioma. *Nat. Commun.* **11**, 2659 (2020).

404. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software, Articles* **77**, 1–17 (2017).

405. Nearing, J. T. *et al.* Microbiome differential abundance methods produce disturbingly different results across 38 datasets. *bioRxiv* 2021.05.10.443486 (2021) doi:10.1101/2021.05.10.443486.

406. Hooper, L. V., Littman, D. R. & Macpherson, A. J. Interactions between the microbiota and the immune system. *Science* **336**, 1268–1273 (2012).

407. Peschel, S., Müller, C. L., von Mutius, E., Boulesteix, A.-L. & Depner, M. NetCoMi: network construction and comparison for microbiome data in R. *Brief. Bioinform.* (2020) doi:10.1093/bib/bbaa290.

408. Gao, R., Gao, Z., Huang, L. & Qin, H. Gut microbiota and colorectal cancer. *Eur. J. Clin. Microbiol. Infect. Dis.* **36**, 757–769 (2017).

409. Evans, S. E. & Wallenstein, M. D. Climate change alters ecological strategies of soil bacteria. *Ecol. Lett.* **17**, 155–164 (2014).

410. Weiss, S. *et al.* Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* **10**, 1669–1681 (2016).

411. Hirano, H. & Takemoto, K. Difficulty in inferring microbial community structure based on co-occurrence network approaches. *BMC Bioinformatics* **20**, 329 (2019).

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

413. Feizi, S., Marbach, D., Médard, M. & Kellis, M. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nat. Biotechnol.* **31**, 726–733 (2013).

414. Barzel, B. & Barabási, A.-L. Network link prediction by global silencing of indirect correlations. *Nat. Biotechnol.* **31**, 720–725 (2013).

415. Birt, H. W. G. & Dennis, P. G. Inference and Analysis of SPIEC-EASI Microbiome Networks. *Methods Mol. Biol.* **2232**, 155–171 (2021).

416. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441 (2008).

417. Meinshausen, N. & Bühlmann, P. High-dimensional graphs and variable selection with the Lasso. *aos* **34**, 1436–1462 (2006).

418. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.* **58**, 267–288 (1996).

419. Liu, H., Roeder, K. & Wasserman, L. Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. *Adv. Neural Inf. Process. Syst.* **24**, 1432–1440 (2010).

420. Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S. & Bähler, J. Proportionality: a valid alternative to correlation for relative data. *PLoS Comput. Biol.* **11**, e1004075 (2015).

421. Kurtz, Z. D. *et al.* Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11**, e1004226 (2015).

422. Yoon, G., Gaynanova, I. & Müller, C. L. Microbial Networks in SPRING - Semi-parametric Rank-Based Correlation and Partial Correlation Estimation for Quantitative Microbiome Data. *Front. Genet.* **10**, 516 (2019).

423. Fisher, R. A. Statistical Methods for Research Workers. in *Breakthroughs in Statistics: Methodology and Distribution* (eds. Kotz, S. & Johnson, N. L.) 66–70 (Springer New York, 1992).

424. Gill, R., Datta, S. & Datta, S. A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics* **11**, 95 (2010).

425. Siska, C., Bowler, R. & Kechris, K. The discordant method: a novel approach for differential correlation. *Bioinformatics* **33**, 150 (2017).

426. Strimmer, K. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics* **24**, 1461–1462 (2008).

427. Blei, D. M. & Lafferty, J. D. A correlated topic model of Science. *aoas* **1**, 17–35 (2007).

428. Sankaran, K. & Holmes, S. P. Latent variable modeling for the microbiome. *Biostatistics* **20**, 599–614 (2019).

429. Woloszynek, S. *et al.* Exploring thematic structure and predicted functionality of 16S rRNA

amplicon data. *PLoS One* **14**, e0219235 (2019).

430. Brandl, B. *et al.* A Phenotyping Platform to Characterize Healthy Individuals Across Four Stages of Life - The Enable Study. *Front Nutr* **7**, 582387 (2020).

431. Rennekamp, R., Brandl, B., Giesbertz, P., Skurk, T. & Hauner, H. Metabolic and satiating effects and consumer acceptance of a fibre-enriched Leberkas meal: a randomized cross-over trial. *Eur. J. Nutr.* **60**, 3203–3210 (2021).

432. McRae, M. P. Dietary Fiber Is Beneficial for the Prevention of Cardiovascular Disease: An Umbrella Review of Meta-analyses. *J. Chiropr. Med.* **16**, 289–299 (2017).

433. Anderson, J. W. *et al.* Health benefits of dietary fiber. *Nutr. Rev.* **67**, 188–205 (2009).

434. Aune, D. *et al.* Dietary fibre, whole grains, and risk of colorectal cancer: systematic review and dose-response meta-analysis of prospective studies. *BMJ* **343**, d6617 (2011).

435. Xu, X., Zhu, Y., Li, J. & Wang, S. Dietary fiber, glycemic index, glycemic load and renal cell carcinoma risk. *Carcinogenesis* **40**, 441–447 (2019).

436. Reitmeier, S. *et al.* Handling of spurious sequences affects the outcome of high-throughput 16S rRNA gene amplicon profiling. *ISME Communications* **1**, 1–12 (2021).

437. da Cruz, A. G., Senaka Ranadheera, C., Nazzaro, F. & Mortazavian, A. *Probiotics and Prebiotics in Foods: Challenges, Innovations, and Advances*. (Academic Press, 2021).

438. Bailén, M. *et al.* Microbiota features associated with a high-fat/low-fiber diet in healthy adults. *Front. Nutr.* **7**, 583608 (2020).

439. Aoe, S., Nakamura, F. & Fujiwara, S. Effect of Wheat Bran on Fecal Butyrate-Producing Bacteria and Wheat Bran Combined with Barley on Bacteroides Abundance in Japanese Healthy Adults. *Nutrients* **10**, (2018).

440. Van den Abbeele, P. *et al.* Butyrate-producing Clostridium cluster XIVa species specifically colonize mucins in an in vitro gut model. *ISME J.* **7**, 949–961 (2013).

441. Menni, C. *et al.* Gut microbiome diversity and high-fibre intake are related to lower long-term weight gain. *Int. J. Obes.* **41**, 1099–1105 (2017).

442. Higgins, J. A. *et al.* Resistant starch consumption promotes lipid oxidation. *Nutr. Metab.* **1**, 8 (2004).

443. Kverka, M. *et al.* Oral administration of Parabacteroides distasonis antigens attenuates experimental murine colitis through modulation of immunity and microbiota composition. *Clin. Exp. Immunol.* **163**, 250–259 (2011).

444. Wang, K. *et al.* Parabacteroides distasonis Alleviates Obesity and Metabolic Dysfunctions via Production of Succinate and Secondary Bile Acids. *Cell Rep.* **26**, 222–235.e5 (2019).

445. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).

446. Cuffaro, B. *et al.* In Vitro Characterization of Gut Microbiota-Derived Commensal Strains: Selection of Parabacteroides distasonis Strains Alleviating TNBS-Induced Colitis in Mice. *Cells* **9**, (2020).

447. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).

448. Ezeji, J. C. *et al.* Parabacteroides distasonis: intriguing aerotolerant gut anaerobe with emerging antimicrobial resistance and pathogenic and probiotic roles in human health. *Gut Microbes* **13**, 1922241 (2021).

449. Tian, T. *et al.* Effects of Short-Term Dietary Fiber Intervention on Gut Microbiota in Young Healthy People. *Diabetes Metab. Syndr. Obes.* **14**, 3507–3516 (2021).

450. Shortt, C. *et al.* Systematic review of the effects of the intestinal microbiota on selected nutrients and non-nutrients. *Eur. J. Nutr.* **57**, 25–49 (2018).

451. Sun, B., Hou, L. & Yang, Y. Effects of altered dietary fiber on the gut Microbiota, short-chain fatty acids and cecum of chickens during different growth periods. *Preprints* (2020) doi:10.20944/preprints202002.0109.v1.

452. Corrêa-Oliveira, R., Fachi, J. L., Vieira, A., Sato, F. T. & Vinolo, M. A. R. Regulation of immune cell function by short-chain fatty acids. *Clin. Transl. Immunology* **5**, e73 (2016).

453. Donohoe, D. R. *et al.* The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon. *Cell Metab.* **13**, 517–526 (2011).

454. Wang, G. Human antimicrobial peptides and proteins. *Pharmaceuticals* **7**, 545–594 (2014).

455. Fukagawa, N. K., Anderson, J. W., Hageman, G., Young, V. R. & Minaker, K. L. High-

carbohydrate, high-fiber diets increase peripheral insulin sensitivity in healthy young and old adults. *Am. J. Clin. Nutr.* **52**, 524–528 (1990).

456. Soriano-Lerma, A. *et al.* Influence of 16S rRNA target region on the outcome of microbiome studies in soil and saliva samples. *Sci. Rep.* **10**, 13637 (2020).

457. Matsuo, Y. *et al.* Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinION^{TM} nanopore sequencing confers species-level resolution. *BMC Microbiol.* **21**, 35 (2021).

458. Quince, C., Lanzen, A., Davenport, R. J. & Turnbaugh, P. J. Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* **12**, 38 (2011).

459. Huse, S. M., Welch, D. M., Morrison, H. G. & Sogin, M. L. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.* **12**, 1889–1898 (2010).

460. Rosselló-Mora, R. & Amann, R. The species concept for prokaryotes. *FEMS Microbiol. Rev.* **25**, 39–67 (2001).

461. Haft, D. H. *et al.* RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.* **46**, D851–D860 (2018).

462. Portik, D. M., Brown, C. T. & Pierce-Ward, N. T. Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets. *BMC Bioinformatics* **23**, 541 (2022).

463. Gehrig, J. L. *et al.* Finding the right fit: evaluation of short-read and long-read sequencing approaches to maximize the utility of clinical microbiome data. *Microb Genom* **8**, (2022).

464. Ash, C., Farrow, J. A., Dorsch, M., Stackebrandt, E. & Collins, M. D. Comparative analysis of Bacillus anthracis, Bacillus cereus, and related species on the basis of reverse transcriptase sequencing of 16S rRNA. *Int. J. Syst. Bacteriol.* **41**, 343–346 (1991).

465. Bavykin, S. G. *et al.* Use of 16S rRNA, 23S rRNA, and gyrB gene sequence analysis to determine phylogenetic relationships of Bacillus cereus group microorganisms. *J. Clin. Microbiol.* **42**, 3711–3730 (2004).

466. Yan, Y., Nguyen, L. H., Franzosa, E. A. & Huttenhower, C. Strain-level epidemiology of microbial communities and the human microbiome. *Genome Med.* **12**, 71 (2020).

467. von Mentzer, A. *et al.* Long-read-sequenced reference genomes of the seven major lineages of

enterotoxigenic Escherichia coli (ETEC) circulating in modern time. *Sci. Rep.* **11**, 9256 (2021).

468. Chin, C.-S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).

469. Xu, W. *et al.* Characterization of Shallow Whole-Metagenome Shotgun Sequencing as a High-Accuracy and Low-Cost Method by Complicated Mock Microbiomes. *Front. Microbiol.* **12**, 678319 (2021).

470. Lugli, G. A. & Ventura, M. A breath of fresh air in microbiome science: shallow shotgun metagenomics for a reliable disentangling of microbial ecosystems. *Microbiome Res. Rep.* (2022) doi:10.20517/mrr.2021.07.

471. Jansson, J. K. & Hofmockel, K. S. The soil microbiome—from metagenomics to metaphenomics. *Curr. Opin. Microbiol.* **43**, 162–168 (2018).

472. Abu-Ali, G. S. *et al.* Metatranscriptome of human faecal microbial communities in a cohort of adult men. *Nat Microbiol* **3**, 356–366 (2018).

473. Jiang, Y., Xiong, X., Danska, J. & Parkinson, J. Metatranscriptomic analysis of diverse microbial communities reveals core metabolic pathways and microbiome-specific functionality. *Microbiome* **4**, 2 (2016).

474. Yan, X. *et al.* Integrated multi-omics of the gastrointestinal microbiome and ruminant host reveals metabolic adaptation underlying early life development. *Microbiome* **10**, 222 (2022).

475. Rebollar, E. A. *et al.* Using 'Omics' and Integrated Multi-Omics Approaches to Guide Probiotic Selection to Mitigate Chytridiomycosis and Other Emerging Infectious Diseases. *Front. Microbiol.* **7**, (2016).

476. Ferrocino, I. *et al.* The need for an integrated multi-OMICs approach in microbiome science in the food system. *Compr. Rev. Food Sci. Food Saf.* (2023) doi:10.1111/1541-4337.13103.

477. Park, C. H., Hong, C., Lee, A.-R., Sung, J. & Hwang, T. H. Multi-omics reveals microbiome, host gene expression, and immune landscape in gastric carcinogenesis. *iScience* **25**, 103956 (2022).

478. Jagtap, P. D. *et al.* Metaproteomic analysis using the Galaxy framework. *Proteomics* **15**, 3553–3565 (2015).

479. Muth, T. *et al.* MPA Portable: A Stand-Alone Software Package for Analyzing Metaproteome

Samples on the Go. *Anal. Chem.* **90**, 685–689 (2018).

480. Cheng, K. *et al.* MetaLab: an automated pipeline for metaproteomic data analysis. *Microbiome* **5**, 157 (2017).

481. Starr, A. E. *et al.* Proteomic and Metaproteomic Approaches to Understand Host–Microbe Interactions. *Anal. Chem.* **90**, 86–109 (2018).

482. Zhang, X. *et al.* Deep Metaproteomics Approach for the Study of Human Microbiomes. *Anal. Chem.* **89**, 9407–9415 (2017).

483. Maier, T. V. *et al.* Impact of Dietary Resistant Starch on the Human Gut Microbiome, Metaproteome, and Metabolome. *MBio* **8**, (2017).

484. Gavin, P. G. *et al.* Intestinal Metaproteomics Reveals Host-Microbiota Interactions in Subjects at Risk for Type 1 Diabetes. *Diabetes Care* **41**, 2178–2186 (2018).

485. Zoppi, J., Guillaume, J.-F., Neunlist, M. & Chaffron, S. MiBiOmics: an interactive web application for multi-omics data exploration and integration. *BMC Bioinformatics* **22**, 6 (2021).

486. Tewari, A. *et al.* MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. in *2017 IEEE International Conference on Computer Vision (ICCV)* 1274–1283 (IEEE, 2017).

487. Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).

488. Singh, A. *et al.* DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* **35**, 3055–3062 (2019).

489. Muñoz-Benavent, M. *et al.* gNOMO: a multi-omics pipeline for integrated host and microbiome analysis of non-model organisms. *NAR Genom. Bioinform.* **2**, lqaa058 (2020).

490. Ali, R. O. *et al.* Longitudinal multi-omics analyses of the gut–liver axis reveals metabolic dysregulation in hepatitis C infection and cirrhosis. *Nature Microbiology* **8**, 12–27 (2022).

491. Yu, Z. *et al.* Gut microbiome in PCOS associates to serum metabolomics: a cross-sectional study. *Sci. Rep.* **12**, 22184 (2022).

492. Miller, D., Stern, A. & Burstein, D. Deciphering microbial gene function using natural language processing. *Nat. Commun.* **13**, 5731 (2022).

493. Tataru, C. *et al.* Topic modeling for multi-omic integration in the human gut microbiome and implications for Autism. *bioRxiv* 2022.09.30.509056 (2022) doi:10.1101/2022.09.30.509056.

494. Chen, X. *et al.* Estimating functional groups in human gut microbiome with probabilistic topic models. *IEEE Trans. Nanobioscience* **11**, 203–215 (2012).

495. Movassagh, M. *et al.* Vaginal microbiome topic modeling of laboring Ugandan women with and without fever. *NPJ Biofilms Microbiomes* **7**, 75 (2021).

496. Xiong, X. *et al.* A new method for mining information of gut microbiome with probabilistic topic models. *Multimed. Tools Appl.* (2022) doi:10.1007/s11042-022-13916-7.

497. Parfrey, L. W., Walters, W. A. & Knight, R. Microbial eukaryotes in the human microbiome: ecology, evolution, and future directions. *Front. Microbiol.* **2**, 153 (2011).

498. Huffnagle, G. B. & Noverr, M. C. The emerging world of the fungal microbiome. *Trends Microbiol.* **21**, 334–341 (2013).

499. Kayama, K. *et al.* Prediction of PCR amplification from primer and template sequences using recurrent neural network. *Sci. Rep.* **11**, 7493 (2021).

Publications

1. Dietrich, A., **Matchado, M.S.**, Zwiebel, M., Ölke, B., Lauber, M., Lagkouvardos, I., Baumbach, J., Haller, D., Brandl, B., Skurk, T. and Hauner, H., 2021. Namco: A microbiome explorer. Microb Genom . 2022 Aug;8(8). doi: 10.1099/mgen.0.000852 (shared first authors)
2. **Matchado, MS,** Michael Lauber, Sandra Reitmeier, Tim Kacprowski, Jan Baumbach, Dirk Haller, and Markus List. "Network analysis methods for studying microbial communities: A mini review." Computational and structural biotechnology journal (2021).
3. Abellan-Schneyder, I., **Matchado, MS**., Reitmeier, S., Sommer, A., Sewald, Z., Baumbach, J., List, M. and Neuhaus, K., 2021. Primer, Pipelines, Parameters: Issues in 16S rRNA Gene Sequencing. Msphere, 6(1).

In addition to the first author publications, I also published as a contributing author in peer-reviewed journals (not part of this dissertation):

1. D Häcker, K Siebert, A Metwaly, H Hölz, F De Zen, N Köhler, J K Pauling, **M Matchado**, M List, T Schwerd, D Haller, P366 Exclusive enteral nutrition drives protective microbiome modulation in paediatric Crohn’s Disease, Journal of Crohn's and Colitis, Volume 16, Issue Supplement_1, January 2022, Pages i371–i372, https://doi.org/10.1093/ecco-jcc/jjab232.493

Manuscript in preparation:

1. **Matchado MS**, Malte Rühlemann [2], Sandra Reimeiter[3], Tim Kacprowski[4], Fabian Frost[5], Dirk Haller[3,6], Jan Baumbach[7,8], Markus List. On the limits of 16S-based metagenome prediction and functional profiling

List of figures

Figure 1: Overview of the thesis 7

Figure 2: Structural diagram of 16S rRNA 17

Figure 3: Flowchart of the major steps involved in 16S rRNA gene analysis 21

Figure 4:Overview of microbial data analysis pipeline and biases 25

Figure 5: Summary of a metagenomics workflow 30

Figure 6: Different types of clustering methods utilized in 16S rRNA gene analysis 42

Figure 7: Overview of the analysis strategies used in this study 48

Figure 8: MDS plots illustrating microbiome's distinct communities in human samples across different primer pairs 49

Figure 9: : Comparison of the impact of the clustering method on taxonomic categorization for the ZIEL-I mock community 51

Figure 10: Genus level comparison of mock communities using different primer regions and different databases as references 52

Figure 11: Studying the impacts specific pipeline settings have on taxonomic classification 55

Figure 12: Recommendation for a validation process prior to initiating new microbiome investigations, particularly in rare habitats 59

Figure 13: Workflow of two main methods in PICRUSt 69

Figure 14: Simplified workflow of PICRUSt2 71

Figure 15: Entire workflow of Tax4Fun2 73

Figure 16: Flow diagram of PanFP 75

Figure 17: A schematic workflow of a development MetGEMs toolbox 76

Figure 18: A breakdown of the various 16S rRNA gene copy levels found in bacterial population 78

Figure 19: Overall benchmarking workflow to compare and evaluate the performance of functional inference tools from 16S rRNA gene sequences to metagenomics 81

Figure 20: Spearman correlations plot between metagenome predictions and shotgun metagenome sequencing in unpermuted and permuted datasets 85

Figure 21: Comparison of detected KEGG orthologs between predicted metagenomes and MGS. Precision, F1, and Recall are reported for each category compared to the MGS data 87

Figure 22: The overlap of significant KO terms between different functional inferred tools and MGS results in the CRC cohort..88

Figure 23: Comparison of significantly differentially abundant KO terms between predicted metagenomes and MGS 89

Figure 24: Spearman correlations between p-values obtained from prediction tools and metagenome sequencing for level 1 KEGG functional categories including cellular process, genetic information processing for CRC cohort..91

Figure 25: Comparison of relative abundance distributions among major KEGG gene categories such as (a) glycan biosynthesis metabolism and (b) lipid metabolism between inferred tools and MGS functional profiles in CRC cohort 93

Figure 26: Comparison of relative abundance distributions among major KEGG gene categories such as (a) carbohydrate metabolism and (b) amino acid metabolism between inferred tools and MGS functional profiles in the KORA cohort.......................................95

Figure 27: Comparison of relative abundance distributions among major KEGG gene categories such as (a) carbohydrate metabolism and (b) amino acid metabolism between inferred tools and MGS functional profiles in the Popgen cohort 97

Figure 28: Comparison of relative abundance distributions among major KEGG gene categories such as (a) carbohydrate metabolism and (b) amino acid metabolism between inferred tools and MGS functional profiles in the FoCus cohort..99

Figure 29:(a) Functional beta diversity of the simulated metagenome-amplicon sample pairs between PC1 and PC2. (b) Functional beta diversity of the simulated metagenome-amplicon sample pairs between PC2 and PC3..100

Figure 30: Pairwise differential analysis was performed between simulated GI vs Oral and Skin vs airway 101

Figure 31: Namco's comparison with other web-based tools for the analysis of 16S rRNA gene data...110

Figure 32: Welcome page of Namco...111

Figure 33: Overview of Namco features 111

Figure 34: Workflow of upstream analysis and data overview page of Namco................113

Figure 35: Workflow indicating the suitable network approaches depending on different challenges...119

Figure 36: The basic concept of the topic modelling algorithm 120

Figure 37: Design of meals for fibre-enriched Intervention (enriched) and the placebo (standard) meatloaf and salami pizza meal..121

Figure 38: Microbial Diversity between intervention groups 125

Figure 39:Taxonomic composition across intervention groups..127

Figure 40: Differential abundance of species between intervention groups 127

Figure 41: Corrplot explains the positive and negative correlation between gut microbes and selected clinical variables at the phyla (a) and (b) genus level....................................129

Figure 42: Bar plot represents the abundance difference in the KEGG pathways between intervention and non-intervention groups...131

Figure 43: Bar plot depicts the variation in relative abundance of ABC transporters and glucokinase between the IM and M groups...132

Figure 44: Differential network of genus between intervention groups............................133

List of tables

Table 1: Advantages and disadvantages of using the 16S rRNA gene as a marker gene in microbial analysis 19

Table 2: Software tools and packages used for 16S rRNA gene analysis 23

Table 3: Overview of disease cohorts 82

Abbreviations

ABC	ATP-Binding Cassette
AST	Aspartate AminoTransferase
ASV	Amplicon Sequence Variant
BH	Benjamini–Hochberg
BMI	Body Mass Index
CLR	Centred Log Ratio
FASYN-ELONG-PWY	Fatty Acid Elongation Saturated Pathway
GDPR	General Data Protection Regulation
GG	GreenGenes
GI	GastroIntestinal
GLCMANNANAUT-PWY	Super Pathway Of N-Acetylglucosamine, N-Acetylmannosamine And N-Acetylneuraminate Degradation
GALT	Gut-Associated Lymphoid Tissue
HMP	Human Microbiome Project
IBD	Inflammatory Bowel Disease
IBS	Irritable Bowel Syndrome
iHMP	Integrative Human Microbiome Project
KEGG	Kyoto Encyclopedia of Genes and Genomes
INSDC	The International Nucleotide Sequence Database Collaboration
LTP	The All-Species Living Tree
MetaHIT	Metagenomes of the Human Intestinal Tract
MGS	Metagenomics shotgun metagenomics
NAGLIPASYN-PWY	lipid IVA biosynthesis
NGS	Next Generation Sequencing
NMDS	Non-metric MultiDimensional Scaling
OANTIGEN-PWY	O-antigen building blocks biosynthesis (E. scherichia coli)
OTU	Operational Taxonomic Unit
PCoA	Principal Coordinates Analysis
POLYISOPRENSYN-PWY	Polyisoprenoid biosynthesis (E. coli)

P161-PWY	Acetylene degradation
ROC	Receiver Operator Characteristic
sFL16S	synthetic long-read sequencing technology
SCFA	Short-chain fatty acid
zOTU	Zero radius Operational taxonomic unit