PhD Program in Biomolecular Sciences
Department of Cellular, Computational and Integrative Biology – CIBIO
37th Cycle

Exploiting the potential of metagenomics to uncover novel and uncharacterized gut microbiome diversity

Tutor

Prof. Nicola Segata

Department CIBIO, University of Trento, Trento, Italy

Advisor

Prof. Mireia Valles-Colomer

Department CIBIO, University of Trento, Trento, Italy

MELIS Department, University Pompeu Fabra, Barcelona, Spain

Ph.D. Thesis of

Davide Golzato

Academic Year 2023-2024

Declaration

I, Davide Golzato, confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Abstract **1**

Chapter 1 - Introduction and aims of the thesis **3**
1.1 Introduction 3
1.1.1 Alterations in the composition of the gut microbiome and its consequences 4
1.1.2 The still unexplored fraction of the gut microbiome: the cultivation problem 5
1.1.3 Next-generation sequencing technologies to study the microbiome 5
1.1.4 Putting the genomic puzzle together: metagenomic assembly 6
1.1.5 How challenges in metagenomic assembly hinder the characterisation of undescribed microbial diversity 7
1.1.6 Multi-sample strategies to improve the metagenomic assembly 8
1.2 Aims of the thesis 9
1.3 Structure of the thesis 9
1.4 Contribution 10

Chapter 2 - Background **12**
2.1 Evolution of Gut Microbiome Research - From Early Discoveries to Shotgun Metagenomic Sequencing 12
2.1.1 Culture-based techniques and first-generation sequencing 12
2.1.2 16S rRNA gene amplicon sequencing 13
2.1.3 Shotgun metagenomics 13
2.2 The two main computational approaches to analyze shotgun metagenomics 15
2.2.1 Reference-based profiling 15
2.2.2 Assembly-based methods 17
2.2.2.1 Metagenomic assembly 17
2.2.2.2 Metagenomic binning 20
2.2.2.3 Multi-sample variants of metagenomic assembly and binning 22
2.2.2.4 Assessing the quality of assembled genomes 25
2.2.2.5 Taxonomic assignment of MAGs 26
2.2.2.6 Functional annotation of MAGs 27
2.3 Analysis of microbiome data (post-processing analysis) 28
2.3.1 Microbial phylogenetics, genomics, and ecology analyses 28
2.3.2 Microbial diversity analysis (alpha- and beta-diversity) 29
2.3.3 Discoveries and applications of MAGs 30
2.3.4 The BioBakery workflow and the underlying MetaRefSGB database 31

Chapter 3 - Performance Assessment of Metagenomic Assembly and Co-assembly Strategies for Gut Microbiome Taxonomic Analysis 33
3.1. Abstract 34
3.2. Introduction 35
3.3. Results and discussion 37
3.3.1. Metagenomic assembly complements reference-based profiling by uncovering novel gut microbiome taxa 37
3.3.2. Deep sequencing boosts MAG reconstruction but does not saturate metagenomic assembly capabilities 38
3.3.3. Metagenomic (co)assembly reconstructs more MAGs of putative new species 41
3.3.4. Underestimated alpha-diversity estimates linked to low sequencing depths 41
3.3.5. Sample handling protocols influence taxonomic composition and MAG reconstruction 42
3.3.6. Species relative abundance is associated with the likelihood of assembling a genome 44
3.3.7. Co-assembly of longitudinal samples improves recovery of low-abundance SGBs stably present in the community 47
3.3.8. Co-binning improves MAG quality by exploiting the covariance of the contig depth of technical variants 48
3.3.9. Co-assembly of samples from multiple subjects produces chimeric assemblies 50
3.4. Conclusions 54
3.5. Material and Methods 55
3.5.1. Metagenomic cohorts 55
3.5.2. Sample collection, storage, and DNA extraction 56
3.5.3. DNA Sequencing and read pre-processing 56
3.5.4. Metagenomic assembly and co-assembly 57
3.5.5. Metagenomic binning and co-binning of co-assembled contigs 58
3.5.6. Assessing completeness and contamination of MAGs 58
3.5.7. Taxonomic annotation of MAGs 58
3.5.8. Co-assembly of technical variant pairs 59
3.5.9. Reference-based taxonomic profiling 59
3.5.10. Comparison of reference- and assembly-based profiling 60
3.5.11. Alpha- and beta-diversity estimation of metagenomes 60
3.5.12. Statistical methods 60
3.5.13. Association of SGB genome coverage and chances of assembly 61
3.5.14. Comparison of binning and co-binning 61

3.5.15. Assessment of MAG chimericity 61
3.6. Supplementary Tables 62
3.6.1. Supplementary Figures 63

Chapter 4 - Other Contributions 73
4.1 Screening, isolation, and cultivation of unknown gut microbial lineages 73
4.1.1 Draft Genome Sequence of *Neopoerus faecalis* gen. nov., sp. nov., an *Oscillospiraceae* Strain Isolated from Human Feces 74
4.1.2 Draft genome sequence of a representative strain of the *Catenibacterium* genus isolated from human feces 76
4.1.3 Draft genome sequences of multiple bacterial strains isolated from human feces 78
4.2 Distinct Chromosomal Mutation Associated with Cefiderocol Resistance in *Acinetobacter baumannii*: A Combined Bioinformatics and Mass Spectrometry Approach to unveil and validate the in-vivo acquired Chemoresistance 79
4.3 Short-term cocoa supplementation influences microbiota composition and serum markers of lipid metabolism in elite male soccer players 81
4.4 Meta-analysis of 22,710 human metagenomes defines an index of oral to gut microbial introgression and associations with age, sex, BMI, and diseases 84

Chapter 5 - Conclusions and Future Perspectives 86
Limitations of this thesis and future perspectives 89

References 93
List of abbreviations 110

Abstract

Metagenomic sequencing has revolutionized gut microbiome research by providing comprehensive access to the entire genomic content of any biological sample, namely a metagenome. Thanks to the possibility of studying microbial ecosystems in-depth without requiring direct isolation or cultivation of their members, metagenomics has greatly expanded knowledge on the taxonomic and functional diversity of the human gut microbiome and how deeply it is involved in human physiology. Metagenomic assembly is a computational technique that enables the reconstruction of bacterial genomes, known as metagenome-assembled genomes (MAGs). Systematically recovering MAGs from gut metagenomes has allowed researchers to progressively unfold the complexity of the microbiome-host system by cataloging and characterizing the genomes of thousands of previously unknown bacterial lineages that comprise it. Despite its importance, this task faces computational limitations that complicate the recovery of microbial diversity associated with rare and low-abundance species, popularly known as the 'microbial dark matter'. Consequently, optimizing available metagenomic data to maximize observable diversity and genome reconstruction is crucial for comprehensive microbiome analysis.

In this doctoral thesis, I explore how the concurrent processing of multiple biologically similar metagenomes, when available, using reference- and assembly-based approaches can help in the identification of previously undetected bacterial species. More specifically, I performed metagenomic (co)assembly and (co)binning and applied it to a cohort of ultra-deep, redundantly sequenced gut metagenomes from a small number of individuals. I demonstrate that the careful application of this approach allows for the recovery of high-quality MAGs from novel and under-characterized bacterial species that would otherwise be missed with a single sample. This allowed for the reconstruction of genomes from 198 species lacking reference genomes and 39 completely novel microbial species from gut communities that should already be well represented, highlighting how a significant amount of phylogenetic diversity has remained hidden primarily due to the low sequencing depth of most studies, rather than an insufficient number of sampled individuals. Although multi-sample approaches have been applied in numerous studies for the aforementioned reasons, this work outlines the ideal conditions to apply them in cross-sectional and longitudinal contexts to minimize the occurrence of assembly errors. I show that (co)assembly is most effective with samples from the same subject, as combinations of samples from unrelated subjects generates strain-chimeric MAGs that do not represent actual strains populations. In parallel, I also provide estimates of the sequencing requirements needed to capture this diversity by complementing (co)assembly with reference-based methods. The findings in this thesis advance our understanding of metagenomic assembly

techniques and highlight the importance of optimizing data usage in microbiome studies. The recovery of high-quality MAGs empowers various applications, from surveying unknown species to guiding their experimental isolation and characterization. Furthermore, integrating these MAGs into reference-based approaches enables large-scale screening to draw associations with host-related variables, ultimately contributing to a more comprehensive understanding of the gut microbiome.

Chapter 1 - Introduction and aims of the thesis

1.1 Introduction

The human body is a complex biological system, home to trillions of microorganisms from various life kingdoms, constantly competing for energy resources and colonizing the distinct biological niches each body site represents. These microorganisms form unique communities across different body sites, each with its own composition and genetic repertoire (Curtis et al. 2012). These microbial communities and their genetic makeup are collectively defined as the "Human Microbiome" (Berg et al. 2020). The idea that human microbes actively contribute to our health and well-being is not new; however, discoveries made in the last two decades have led us to realize that the microbiome has such a profound and wide-ranging impact on all aspects of human physiology, from the digestion of fibers to complex social behaviors (Nagpal and Cryan 2021), that defining human health without considering the microbiome is overly simplistic.

Among the body microbiomes, the gut microbiome has certainly been in the spotlight of scientific research. Being the primary site for nutrient breakdown, the gastrointestinal tract (GI) is a vibrant and diverse environment regarding energy sources and metabolites, making it a highly sought-after niche for microbial colonization. As a result, the gut achieves the highest bacterial density (estimates are around 10^{14} microbes) in the body (Meštrović 2016; Ley, Peterson, and Gordon 2006; Luckey 1972) and an intra-individual species variety estimated to be in the order of 1,000 species (Guinane and Cotter 2013).

The gut microbiome is not a passive coexistent with the human host; instead, its strategic location at the interface between the external environment and internal body makes it a crucial mediator of host-environment interactions. Most host-microbe interactions are either commensal or mutualistic, establishing a dynamic equilibrium known as 'symbiosis' (Malard et al. 2021). This symbiotic relationship involves a two-way exchange of nutrients, inhibition of pathogen colonization through immunomodulation (Zheng, Liwinski, and Elinav 2020; Thaiss et al. 2016; C. A. Lozupone 2018), and the breakdown of complex dietary compounds into metabolites available for the host (Fu et al. 2022; Tremaroli and Bäckhed 2012; Rowland et al. 2018; Krishnan, Alden, and Lee 2015). The combined genetic repertoire of the gut microbiota is estimated to exceed that of the human host by at least a hundred times, and this vast array of microbial genes produces a multitude of metabolites that either supplement or complete various metabolic functions of the human body (Grice and Segre 2012). Consequently, the microbiome is increasingly recognized as a 'virtual organ' due to its

pivotal role in regulating fundamental aspects of human physiology, from basal metabolic processes to the onset of various pathologies.

The gut microbiome of each individual is unique in terms of taxonomic composition, with a substantial degree of genomic uniqueness even across members of the same microbial species, providing a biological signature that is more unique than that of the host genome. The gut microbiome is relatively stable throughout adulthood despite compositional fluctuations (Lianmin Chen et al. 2021). Due to this high inter-individual variability, defining the 'core' gut microbiome of healthy individuals is still an ongoing challenge, as the microbiome is influenced by host genetics (Goodrich et al. 2014) and factors such as age, geography, ethnicity, dietary habits or usage of prescription drugs and antibiotics (Gacesa et al. 2022) and cohabitation (Valles-Colomer et al. 2023).

1.1.1 Alterations in the composition of the gut microbiome and its consequences

Gut dysbiosis, which is the disruption of the host-microbe equilibrium, has been proven to contribute to the onset of diseases such as inflammatory bowel disease (IBD) (Halfvarson et al. 2017; Knox et al. 2019), obesity (Castaner et al. 2018; Bouter et al. 2017; Zhi et al. 2019), type-II diabetes (Das et al. 2021), cancer (Thomas et al. 2019; Sun, Chen, and Wu 2023; Fujita et al. 2022; Fattizzo et al. 2021), and even neurological disorders (Cryan et al. 2020). It is characterized by the blooming of inflammation-related taxa and a concomitant reduction in the abundance and diversity of health-promoting commensal members (McDonald et al. 2016). Reduced bacterial alpha diversity has been associated with many diseases (Kostic et al. 2015; Y. Wang et al. 2022; Zhuoxin Li et al. 2022; Le Chatelier et al. 2013; Manichanh et al. 2006), although the biological causes and whether it is the cause or consequence of the diseased state remain debated. Dysbiotic taxonomic shifts result in a change in the overall functional capabilities of the gut microbiome (Manor and Borenstein 2017; Turnbaugh et al. 2009; Zysset-Burri et al. 2019). Various factors can drive the disruption of host-microbe homeostasis, including host genetics favoring the growth of certain taxa, endogenous and exogenous compounds to which the host is exposed, such as dietary compounds or xenobiotics, inflammation, and the transmission and engraftment of microbial species from external sources into the host gut microbiome (Levy et al. 2017; Kelsen and Wu 2012).

1.1.2 The still unexplored fraction of the gut microbiome: the cultivation problem

The inputs and conditions that fuel the intricate network of metabolic pathways underlying a microbe's growth often remain unknown, making the traditional isolation- and cultivation-based methods inapt for the study of indigenous bacteria. Most gut microbial taxa are obligate anaerobes, and their growth is often bound to the unique environmental conditions present in different parts of the GI, which include the presence of particular energy sources or molecular co-factors, physical conditions, syntrophy with other bacteria or adherence to host epithelial cells. Despite technology also advancing in this direction (Xu et al. 2024; Hitch et al. 2021; Lagier et al. 2012) and many of the dominant members of the gut microbiome having been cultivated, a large fraction of gut microbial taxa remains uncultivated, unisolated or even undetected (Thomas and Segata 2019). Unexplored microbial biodiversity has many times been referred to as Microbial Dark Matter, also due to the unknown functional potential and applications its exploration might bear (Zha et al. 2022).

1.1.3 Next-generation sequencing technologies to study the microbiome

The advent of next-generation sequencing technologies (NGS) allowed the high-throughput and untargeted (“shotgun”) sequencing of the genetic material contained in any kind of environmental sample. Metagenomics concerns the study of the structure and functions of a metagenome, defined as the collection of genomes and genes of all the microorganisms (‘meta-’) present in a sample (J. Handelsman et al. 1998; Marchesi and Ravel 2015). Although 16S rRNA amplicon sequencing provided a first omic way to estimate, catalog, and quantify the microbial diversity of an environment, it does not provide access to genetic regions of the microorganisms other than that of a single marker gene. Metagenomics, in contrast, provides an isolation- and culture-free approach to reconstructing whole microbial genomes and annotating genes to characterize at high phylogenetic resolution the taxonomy and the potential functional roles of the microbial members of a community. Since the first large metagenomic research initiatives, such as the Human Microbiome Project (Human Microbiome Project Consortium 2012) and the MetaHIT consortium (Qin et al. 2010), the field of metagenomics has experienced significant growth. Today, more than 100,000 metagenomes are publicly available on NCBI alone (N. Kim et al. 2024). The scientific community has made a concerted effort to develop and maintain resources and tools to organize and utilize this wealth of information effectively (Beghini et al. 2021; Pasolli et al. 2017; Schmidt et al. 2024).

1.1.4 Putting the genomic puzzle together: metagenomic assembly

Unlike traditional genomics, where reference genomes for the organism of interest are available, metagenomic samples are mixtures of an unknown number of microbial lineages. When many reference genomes are available, detecting microbial taxa and estimating the overall composition of a microbial community with reference-based approaches is quick, sensitive and precise (Nelson et al. 2010; Blanco-Míguez et al. 2023; Almeida et al. 2021). However, this ideal scenario is rare in metagenomics, as many environmental microbes lack reference genomes, limiting the applicability of this approach.

Metagenomic assembly is the reconstruction of metagenome-assembled genomes (MAGs) from shotgun metagenomic reads, aiming to generate comprehensive collections of genomic fragments representing putative genomes of the microbial strains present in the sample. Thanks to metagenomic assembly, isolation-recalcitrant microbes from the 'uncultured majority' (Tyson et al. 2004; Hofer 2018) can be studied directly from their genomes. The systematic assembly of metagenomic datasets and efforts to catalog the resulting thousands of MAGs (C. Y. Kim et al. 2021; Almeida et al. 2021; Leviatan et al. 2022; Pasolli et al. 2019) into a coherent taxonomic framework (Parks et al. 2020) revealed the existence of thousands of novel microbial lineages, allowing to expand the prokaryotic tree of life massively (Mukherjee et al. 2017; Parks et al. 2017; Pasolli et al. 2019). Applications of MAGs range from comparative genomics studies of common gut commensal species, characterizing functional and genetic potential at a global scale (pangenome), studying their evolutionary trajectory in the context of human history (Tett et al. 2019; Karcher et al. 2021), to the mining of genes with bioactive potential, such as antimicrobial molecules or enzymes (Jia et al. 2022), or for studying the spread and development of new antibiotic resistance genes (Zhang et al. 2022). Genes and other prokaryotic genetic elements can be obtained from a MAG, and information about their function can be inferred from catalogs of experimentally validated taxonomic and functional gene annotations. However, most of the genes are functionally not annotated.

Once retrieved, MAGs and their genes can be included in reference databases, enabling their detection in metagenomic samples through profiling without the need for assembly (Quince, Walker, et al. 2017). Regular updates of reference databases with MAGs progressively enhance reference-based approaches, yielding increasingly accurate and more representative profiles of microbial gut communities. In this sense, MAG recovery has become a required step when approaching an understudied environment or reconstructing previously undetected microbial species. MAGs have become the basis for reference-based approaches that are commonly used in many

multi-cohort large-scale studies to determine bacterial signatures able to confidently predict multiple diseases such as NAFLD (Oh et al. 2020), colorectal cancer (Thomas et al. 2019) or Crohn's disease (Pascal et al. 2017).

1.1.5 How challenges in metagenomic assembly hinder the characterisation of undescribed microbial diversity

De novo genome assembly from short sequencing reads is a computationally intensive and mathematically intractable problem that requires heuristic algorithms (Medvedev et al. 2007), further complicated by sequence characteristics such as intra-genomic repetitions or tandem repeats, which pose challenges for reconstruction due to the limitations of sequencing technologies. These difficulties are amplified in metagenomic assembly, where the goal is to reconstruct multiple genomes from a complex microbial community with varying abundances.

The challenges of accurately and comprehensively reconstructing the real microbial community through metagenomic assembly can be attributed to both biological and technical factors. Biologically, redundant sequences across different species, such as conserved 16S rRNA gene regions or other phylogenetically conserved regions, complicate distinguishing and reconstructing individual genomes. Technical factors, including pre-sequencing sample handling (collection, DNA extraction, and storage) and sequencing depth, can alter the final microbial composition and impact each genomes' coverage, with higher sequencing depths increasing the likelihood of capturing low-abundance species and rare variants.

This raises the question: why is it important to assemble low-relative abundance and novel gut microbial species? While the dominant members of the gut microbiome have been successfully assembled, the microbial dark matter, comprising low-abundance species and rare variants of prevalent species, remains challenging to assemble. Despite their low abundance, these microbes can have a profound impact on the ecology of their community (Saw 2021; Han and Vaishnava 2023), or the health of the host, as exemplified by *Fusobacterium nucleatum*, a low-abundance oral commensal, which has been linked to colorectal cancer (Zhu et al. 2024; Zepeda-Rivera et al. 2024), or *Bacteroides fragilis (Boleij et al. 2015)*, another low-abundance gut microbe, has been associated with inflammatory bowel disease and colorectal cancer (Hajishengallis et al. 2011; Bhute, Ghaskadbi, and Shouche 2017; Pust and Tümmler 2022; Rinke et al. 2013; Han, Luong, and Vaishnava 2022). Assembling a sufficient number of genomes of these species has proven crucial for understanding their roles in health.

Furthermore, variants of well-known species can exhibit functional characteristics that deviate substantially from their phylogenetic relatives (Van Rossum et al. 2020). Reconstructing this strain-level variation into distinct MAGs is particularly challenging in

taxonomically complex environments, where even species-level genome recovery is difficult due to factors affecting assembly quality. Successfully preserving strain-level information during metagenome reconstruction is thus essential not only for accurate species characterization but also for developing a complete picture of microbiome diversity and its functional potential.

With this purpose, the scientific community made great efforts to devise and implement optimal strategies to obtain a complete and accurate representation of the human gut microbiome by pushing shotgun metagenomic datasets to the technical limit. Many benchmarks have been done to understand which approaches and tools are suited for this task in different contexts (Meyer et al. 2022; Sczyrba et al. 2017).

1.1.6 Multi-sample strategies to improve the metagenomic assembly

Reconstructing all genomes of species present in environments inhabited by complex microbial communities is problematic. In large-scale metagenomic studies, multiple samples involving the same biological condition are collected and sequenced at sequencing depths that, taken individually, are not enough to gather sufficient reads from low-abundance and rare species to be assembled. Additionally, different samples from the same environment might present differences in the final observed microbial composition due to how the sample was handled before sequencing. For this reason, approaches to exploit information from as many samples as possible have been applied to maximize the retrieval of MAGs from a given environment.

Metagenomic co-assembly is a post-sequencing strategy in which multiple metagenomes are combined so to increase the genomic coverage of microbial species present in the sample environment and improve the chances of their successful assembly (Stewart et al. 2019; Pasolli et al. 2019; Kogawa et al. 2018).

Although abundant microbiome members have been assembled thousands of times, a substantial fraction of the gut microbial diversity remains undiscovered, even in well-studied environments like the human gut. Multi-sample assembly strategies are pivotal for recovering this diversity, but guidelines and benchmarks for these strategies on real gut metagenomic data remain scarce.

1.2 Aims of the thesis

This thesis aims to maximize the potential to assemble rare and low-abundant species in microbiomes by defining the parameters and conditions to enable so.

In particular, my specific aims to reach the overall goal are:

1. To evaluate the proportion of known and unknown microbial diversity that is missed by reference-based profiling, even in well-characterized microbial ecosystems, and to assess how much of this missed diversity can be reconstructed using metagenomic assembly.
2. To assess the impact that sequencing characteristics (e.g. depth, coverage) and sample processing protocols have on metagenomic assembly and its potential to reconstruct so-far undescribed microbial diversity
3. To evaluate the combination of different assembly and binning strategies to improve the performances also considering longitudinal sampling

1.3 Structure of the thesis

This thesis is divided into five chapters, structured around the work conducted during my PhD, including the manuscript of my first research article (currently under submission to a peer-reviewed journal) and other collaborative projects.

Chapter 1 serves as an introduction to the thesis, providing an overview of the topic, its significance, and the main research questions addressed in this work.

Chapter 2 provides a background of the techniques used in and related to the main topics of my thesis, ranging from experimental approaches for generating microbiome research data to computational analyses employed in shotgun metagenomics.

Chapter 3 presents the main manuscript, "Performance Assessment of Metagenomic Assembly and Co-assembly Strategies for Gut Microbiome Taxonomic Analysis." This study evaluates multi-sample assembly methods for recovering hidden gut microbiome diversity often missed due to technical constraints. It also provides guidelines for applying these strategies in various experimental settings.

Chapter 4 describes other works to which I contributed during my PhD, and contextualizes them with the work presented in Chapter 3.

Chapter 5 summarizes the key findings of Chapters 3 and 4, speculates on their implications for gut microbiome analysis, addresses limitations, and proposes strategies for overcoming these in future investigations.

1.4 Contribution

My contribution to the first chapter involved reviewing general-scoped literature on microbiome research and reworking it into a discursive context for the scientific questions presented. For the second chapter, I reviewed technical literature on the state-of-the-art tools and techniques that are presented as background.

As for the third chapter, my contribution includes the ideation of the overall strategy for the computational and statistical analysis of the cohort samples. Specifically, I've coded the pipeline to perform metagenomic co-assembly and co-binning used to retrieve MAGs from the sample combinations, and wrote most of the scripts for post-processing data analysis and visualization, and the manuscript.

My contribution to the fourth chapter, mainly involved taking care of assembly-based and reference-based profiling tasks in collaborative projects, along with genomic, phylogenetic and post-processing analyses.

In the fifth and final chapter, I re-elaborated the results from the previous two chapters in light of emerging sequencing technologies, contextualized the findings, and proposed future research directions.

Chapter 2 - Background

2.1 Evolution of Gut Microbiome Research - From Early Discoveries to Shotgun Metagenomic Sequencing

Research surrounding gut microbes, and the impact they can have on human health can be traced well before the advent of high-throughput DNA technologies, back to 1885, when Theodor Escherich first isolated, cultivated, and described the phenotype of a ‘bacterium coli commune’ isolated from the lower gut, which today we know as *Escherichia coli (Hacker and Blum-Oehler 2007)*. Empirical evidence for the contribution of gut microbiota to host physiology and health was obtained throughout the 20th century, with seminal experiments such as microbiota transplantation to successfully treat enterocolitis (Eiseman et al. 1958), and the demonstration that gut microbes are directly involved in host drug metabolism, as evidenced by germ-free mice's inability to degrade salicylazosulfapyridine - a capacity readily restored when these mice were colonized with human microbes (Peppercorn and Goldman 1972). Despite the groundbreaking observations made during this time, the lack of technology to observe the complex ecological reality of the gut microbiome at the genetic level limited researchers' ability to interpret them and even propose mechanistic explanations.

2.1.1 Culture-based techniques and first-generation sequencing

Bacterial species classification historically relied on phenotypic traits such as colony morphology, nutritional requirements, chemotype, and pathogenicity, all characteristics that in order to be observed required the ability to isolate and cultivate them in vitro. However, different microbes have different growth requirements and conditions, requiring the development of ad-hoc cultivation media and protocols to reproduce the right environmental conditions. As exemplified by the 'great plate count anomaly’ (Razumov 1932; Staley and Konopka 1985), there is a vast discrepancy between directly observed microbial cells in a sample and the number of colonies that can be cultivated from it, highlighting how most species of the microbiome are not easily culturable. Additionally, even assuming that microbial species could be cultivated, observation of phenotypic features lacked sufficient discriminatory power to accurately categorize microbial diversity, as phylogenetically distinct microorganisms may exhibit analogous phenotypes due to convergent or horizontal evolution. First-generation DNA sequencing technologies (Heather and Chain 2016) finally provided limited, but direct access to the genes of cultured bacteria, allowing to experimentally link genotype with the observed phenotypes, and enable genetic-based classification (Woese and Fox

1977). Moreover, another limitation of cultivation-based studies is that isolating species from their natural habitats precludes the study of their interactions within the ecological community context and with other external factors. Despite these limitations, cultivating a microbial species is still a required step to perform *in vitro* validation of what is observed with modern technologies, and advancements in cultivation-based methods are being made to increase their throughput (Ha and Devkota 2020).

2.1.2 16S rRNA gene amplicon sequencing

The first culture-free technique to obtain a comprehensive gut microbiome profile was high-throughput sequencing of the 16S rRNA gene, encoding for the small subunit of the prokaryotic ribosome. Being a fundamental component of the translational machinery, the 16S rRNA gene is an optimal phylogenetic marker, as it is present across all prokaryotes, and it includes different regions that accumulate mutations over time at different mutation rates. NGS allows the high-throughput amplification of 16S rRNA gene copies present in a sample by using universal primers that bind on the conserved regions of the gene to sequence some of the nine hypervariable regions (V1-V9). The resulting sequence data can be used to detect known prokaryotic taxa by comparison to curated 16S rRNA gene databases (Cole et al. 2014; Quast et al. 2013), and obtain the relative abundances profiles of operational taxonomic units (OTUs). 16S rRNA gene amplicon sequencing, while pioneering ecological modeling of microbial communities and enabling quantitative associations with metadata, presents several methodological limitations. These constraints include insufficient resolution for closely related species, inability to assess intra-species diversity, and lack of functional gene information for individual community members. Additionally, the 'universal' primers employed in this technique may fail to amplify a significant portion of microbial diversity due to sequence mismatches (Hong et al. 2009). Moreover, PCR biases and the presence of a variable number of 16S rRNA gene copies across different taxa can distort the observed community composition and abundance estimates.

2.1.3 Shotgun metagenomics

As sequencing throughputs increased and the cost per kilobase sequenced decreased, it became possible to perform untargeted sequencing of the metagenome, i.e the collective DNA content of a sample (J. Handelsman et al. 1998). Metagenomics allows for the sequencing of individual genes present in a microbial community. By eliminating the biases arising from universal primers, metagenomics also provides a better depiction of microbial taxonomic and functional diversity, allowing also for the appreciation of intra-species diversity.

Regardless of the specific experimental design, a typical shotgun metagenomics study [Fig. I] begins with the collection of samples from the environment of interest (e.g.: stool

samples for gut metagenomics). Samples are then stored appropriately, especially if processing is not performed immediately, and the conditions are crucial to maintain the integrity of the microbial community and prevent alterations of the microbial composition. The genetic material of the microorganisms present in the sample is then extracted from the microorganisms, usually by mediating the lysis of the cellular membranes or walls through enzymatic (e.g: lysozyme) or physical (e.g: sonication, bead beating) protocols, with the choice largely depending on the type of sample and the target microorganisms.

Isolated DNA is then sequenced to generate metagenomic data. The majority of metagenomic sequencing is currently performed using short-read sequencing-by-synthesis technologies, which produce millions of short DNA sequences.

Sequencing data then undergoes a pre-processing step to ensure data quality, which usually involves the removal of adapter and barcoding sequences, and low-quality bases, typically located at the ends of the reads, and are either trimmed off or masked to minimize their impact on downstream analyses. Reads from contaminating sources, such as the host (e.g., human DNA in the case of human microbiome studies) or other organisms, are identified by mapping against reference genomes and removed.

The subsequent step is sequence analysis, crucial for converting the bulk of reads into biologically interpretable information (features) to answer fundamental questions such as "Who is present?" and "What are they capable of?" about the sampled community. There are two approaches for sequence analysis: reference-based (or assembly-free) and assembly-based (Quince, Walker, et al. 2017). Reference-based methods categorize reads by comparison to reference databases, while assembly-based approaches reconstruct DNA fragments from sample genomes agnostically. These two approaches are complementary, and it is often crucial to employ both.

Once the structure and gene repertoire of the microbial community in the sample is known, it becomes possible to use statistical methods to uncover and visualize associations with clinical, phenotypic, and experimental metadata. This is the post-processing step, which comprise all downstream analyses, ranging from testing a hypothesis in a clinical context (Davar et al. 2021), exploring multi-cohort data to associate microbial features with variables of interest through machine learning approaches (Asnicar et al. 2021; Thomas et al. 2021), inferring biomarkers for microbiome-based disease detection (Thomas et al. 2019), or to ecological modeling (Pedron et al. 2019) and microbial population genomics (Karcher et al. 2021; Tett et al. 2019)

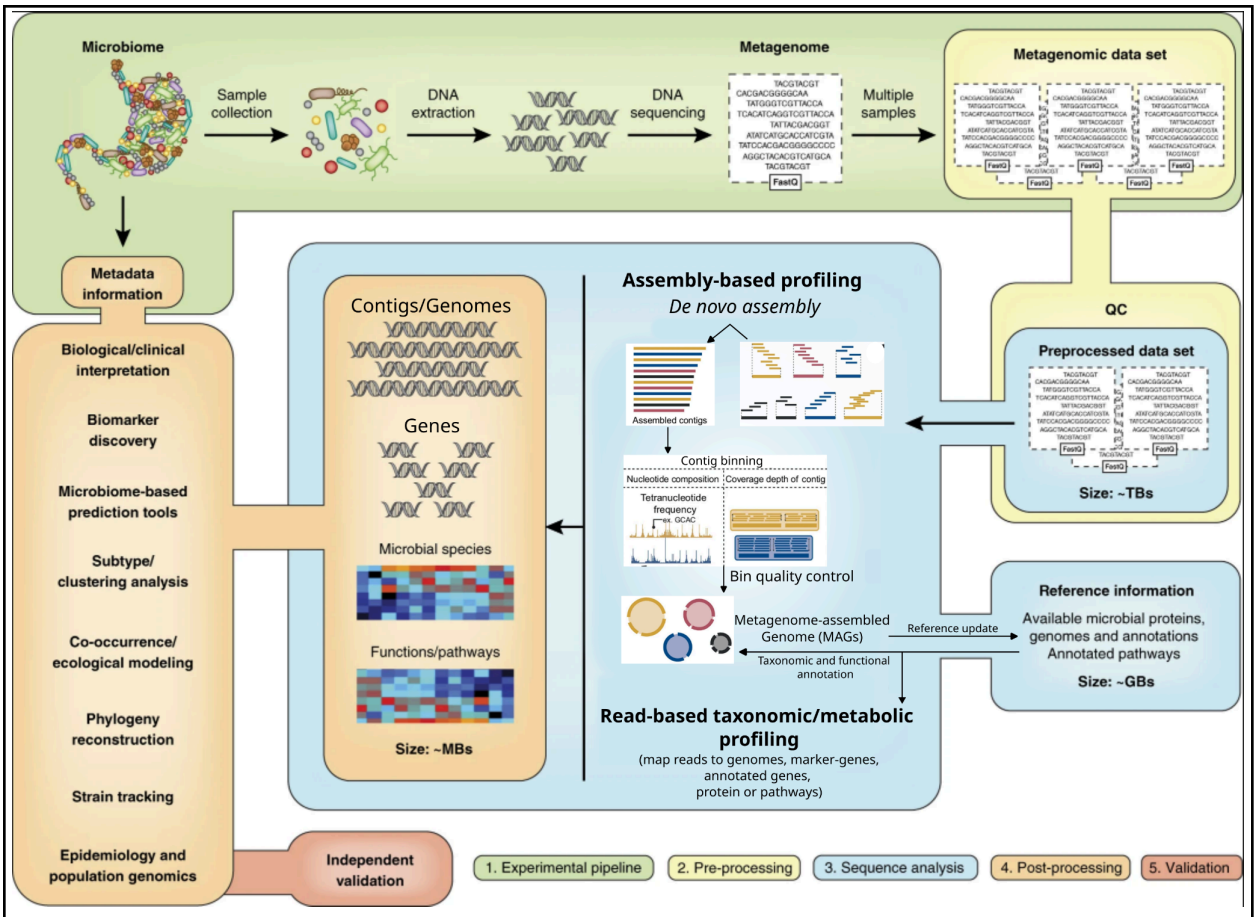

Fig. I: Schematic workflow of the phases comprised in a typical shotgun metagenomics experiment. Readapted from figures taken from Quince, Walker, et al. 2017, and Kim et al. 2024.

2.2 The two main computational approaches to analyze shotgun metagenomics

2.2.1 Reference-based profiling

Reference-based (or assembly-free) methods allow to infer taxonomic and functional composition of a microbial community by direct classification of metagenomic reads using external resources, such as databases of taxonomically annotated reference genomes or curated catalogs of genes and functions. Assembly-free approaches are generally less demanding than metagenomic assembly in terms of computational resources, making them suitable for performing large-scale analyses and/or accurately detecting low-abundance microbial features, such as the presence of specific species or genes if they are represented in reference databases (Quince, Walker, et al. 2017).

Profiling of a microbial community with this approach is accurate as long as the reference databases represent well the taxonomic and functional diversity of the sampled environment. This is because reference-based profiling alone cannot detect nor quantify microbial species for which no reference is available. Hence, the fraction of metagenomic reads that can be classified with reference databases is usually lower for underrepresented environments or body sites (e.g.: soil, freshwaters, stomach). The gut microbiome's biodiversity is now sufficiently covered by a large number of MAGs and reference genomes, making these approaches highly effective for analyzing gut metagenomes (Blanco-Míguez et al. 2023).

Among taxonomic profilers, Kraken is a popular software suite (Lu et al. 2022) that classifies metagenomic reads by first splitting them into subsequences of fixed size (i.e: *k-mers*) that are queried (by exact k-mer matching) against a reference database. The reference database contains k-mers extracted from RefSeq genomes, annotated with the lowest common ancestor (LCA) taxonomy of their source genomes (O'Leary et al. 2016). While this approach is fast and sensitive, since k-mers in the database condense information from whole genomes, it has high memory requirements (Meyer et al. 2022).

An approach employed by some taxonomic profilers to reduce the size of reference databases is the selection of specific gene markers from whole genomes. These markers are chosen to be maximally and uniquely informative about the presence of microbial taxa in a sample. Reads are mapped against sets of clade-specific markers, which should be representative of the intra-clade genomic diversity and allow for unambiguous detection. Two prime examples of software implementing this approach are mOTUs (Milanese et al. 2019) and MetaPhlAn4 (Beghini et al. 2021; Truong et al. 2015; Blanco-Míguez et al. 2023). MetaPhlAn4 utilizes a vast database of over one million reference genomes and high-quality MAGs, clustered into 26,970 species-level genome bins (SGBs). For each SGB, it selects a set of markers that are prevalent within the SGB and do not cross-map with genomes from other SGBs. Metagenomic reads are then aligned against these markers to compute the relative abundances of microbial taxa. By incorporating MAGs, MetaPhlAn4 extends its profiling capabilities to include species that lack taxonomically well-described genomes, thus providing a more comprehensive analysis of microbial communities.

Functional profiles of a metagenome can also be obtained by directly mapping reads against databases of functionally annotated genes and proteins. HUMAnN (Franzosa et al. 2018; Beghini et al. 2021) processes metagenomic and meta-trascriptomics datasets to generate compositional profiles with functions and metabolic pathways of relative abundances stratified by taxonomy. This is achieved by first detecting microbial species present in the sample by running MetaPhlAn and then mapping sample reads against pre-built species pangenomes. Reads that are left unmapped for the pangenomes are

translated at protein-level and mapped against more comprehensive and general protein databases (e.g.: UniRef90/50).

2.2.2 Assembly-based methods

2.2.2.1 Metagenomic assembly

One of the main steps in a shotgun metagenomic workflow is metagenomic assembly. Metagenomic assembly is a computational procedure that aims at reconstructing whole bacterial genomes directly from metagenomic reads produced by shotgun sequencing. As the term suggests, it is mainly based on the algorithmic background of *de novo* whole genome assembly, already used for isolate sequencing, with the difference that is is able to operate under the assumption that the reads originate from a mixture of multiple organisms that are unevenly represented, rather than from a single organism. The main idea behind assembly is that reads whose subsequences are overlapping were sequenced from neighboring regions of the genome, and hence can be matched and ordered into an extended alignment to reconstruct contiguous sequences that are longer than the original reads, namely *contigs*.

Despite this simple idea, *de novo* genome assembly of short reads is a mathematically intractable problem, and this complexity stems mainly from the presence of DNA sequences that are repeated within the same genome or different genomes. In fact, the short length of the reads does not provide enough context specificity to link them to their actual sequence of origin. For this reason, heuristics and assumptions are required to make assembly computationally tractable. The two main strategies for de novo genome assembly are the overlap-layout-consensus (OLC) and De Bruijn graphs (DBG) [Fig. II] (Zhenyu Li et al. 2012).

The OLC approach works by firstly identifying overlapping reads to build an overlap graph, in which reads are nodes and overlaps are edges; the algorithm then finds the path that traverses all graph nodes exactly once (a Hamilthonian path) to derive a sequential layout and use it to derive the final consensus sequence. It is an intuitive algorithm, but it requires all-against-all pairwise alignment of reads, which makes it computationally impractical for assembling libraries with millions of short reads.

De Bruijn-based methods decompose short-read libraries into a non-redundant catalog of short, overlapping subsequences of a fixed size *k*, known as *k-mers*. The set of kmers is then used to build a graph where nodes are all possible (k-1)-mers and edges are drawn between nodes if the suffix of the first node matches the prefix of the second node. This approach has the advantage of using an efficient data structure to represent short-read libraries, and computation time and resources don’t scale up with sequencing depth, making it the method of choice for assembly of short-read libraries.

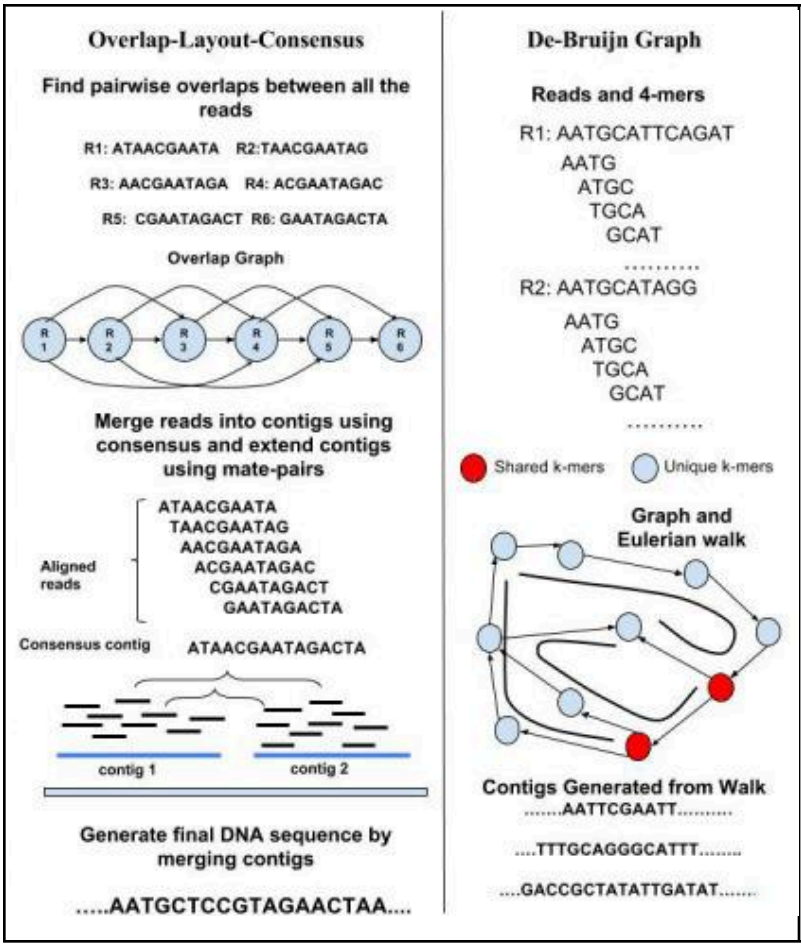

Fig. II: Comparison of OLC and DeBruijn Graphs, the two main algorithmic strategies for de novo genome assembly

Left:The Overlap-Layout-Consensus (OLC) method first identifies overlaps between read pairs to construct an overlap graph, where nodes represent reads and edges represent overlaps. Overlapping reads are then merged to generate consensus contigs, which are finally combined into the complete sequence. Right: In the De Bruijn Graph (DBG) approach, reads are split into k-mers and used to build a graph where nodes are (k-1)-mers and connecting edges represent k-mers. Contigs are generated by finding Eulerian paths that traverse each edge once. Readapted from (Ghurye, Cepeda-Espinoza, and Pop 2016).

Once a DBG graph is constructed, the original sequence can be reconstructed by tracing paths that traverse all edges exactly once, i.e Eulerian paths¸ which are mathematically guaranteed to exist in DBGs. However, the DBG structures generated from real sequencing data are often convoluted and to find the ‘correct’ path multiple heuristics are needed.

Technical and biological factors can complicate the De Bruijn Graph structure. Sequencing errors introduce spurious k-mers, complicating the graph and making it harder to identify correct paths. Highly repetitive DNA regions (e.g. tandem repeats) pose another challenge. Regardless of their length, these regions get collapsed and represented by a limited set of k-mers in the DBG, making full reconstruction of their sequence impossible. Another factor is the presence of intra-genomic homologous sequences located far apart, which get collapsed into identical k-mers and appear to be erroneously closer in the graph, inducing the assembler to erroneously link two regions and generate misassemblies (Olson et al. 2017). Despite these limitations, the low error rates of modern sequencing technologies, combined with strategies to address errors

and ambiguities, have made the DBG paradigm the most widely implemented in metagenomic assemblers for short-reads.

Metagenomic assembly is more complicated than single-genome assembly, given that many of the assumptions of traditional genome assembly are not respected. Many of the ambiguities that in single-genome assembly are treated as technical artifacts can be biologically meaningful in a metagenome and need to be addressed differently. In metagenomes, adding to the previously mentioned issue of intra-genomic repetitions, another source of redundancy are those genes or whole regions that are phylogenetically conserved (e.g: 16s rRNA gene) across different taxa and appear multiple times as variants. For instance, regions that are largely conserved between same-species strains that differ by single nucleotide polymorphisms (SNPs), insertions and deletions (InDels), and structural rearrangements, should ideally be assembled as distinct contigs. However, from the perspective of metagenomic assemblers, it is challenging to discern whether these differences are of technical origin (e.g., sequencing errors or artifacts) or truly represent biological variation. Even more problematic is the assembly of genes or genomic regions that display sequence conservation across phylogenetically distinct bacterial taxa within the same microbial community, as this can result in the generation of chimeric contigs where sequences of disparate taxonomic origins are artificially concatenated (Olson et al. 2017).

For these reasons, strain diversity is a problem in metagenomic assembly of short-read libraries, and variants can introduce branching paths in the DBG graphs much like sequencing errors (Morowitz et al. 2011). The ability to solve these ambiguities is limited to the size of the chosen k-mer length which cannot exceed the read length. Often, assemblers cannot determine the correct path through the DBG due to insufficient genomic context, resulting in either fragmented or chimeric sequences. A notable example is the 16S rRNA gene, which, due to its multiple copies and sequence variants within single and multiple microbial genomes, is frequently reconstructed incompletely or incorrectly (Yuan et al. 2015). Its presence typically fragments genome assembly into multiple contigs, explaining its frequent absence in MAGs (Pasolli et al. 2019). In general, the assembly of metagenomes from environments characterized by high ecological complexity and strain heterogeneity, such as those from soil, wastewater, or ocean samples is more difficult and tends to be more fragmented (Meyer et al. 2022; Fierer 2017; Luo et al. 2012).

Another fundamental variable affecting the quality and sequence representativity of metagenomic assembly is sequencing depth. Sequencing depth is crucial as the reads available for a species genome are proportional to its relative abundance, and for this reason, those from low abundant species are often under-sequenced. Insufficient coverage leads to gaps in the genome coverage profile of a species, resulting in the absence of key k-mers that are essential for forming edges between vertices in the

assembly graph. This ultimately yields incomplete and discontiguous genome reconstructions (Morowitz et al. 2011). Moreover, since strain variants represent only fractions of a species' total abundance, capturing strain-level variation also requires increased sequencing depth.

These aspects have been addressed with various implementations of different metagenomic assembly software, but there is no consensus on which one is the best and none performs well across all metrics, as surveyed by the Critical Assessment of Metagenome Interpretation (CAMI), a community initiative aimed at benchmarking tools for interpretation of metagenomic data (Meyer et al. 2022; Sczyrba et al. 2017; Greenwald et al. 2017). The choice of a de novo assembler is heavily influenced by several factors such as available computational power and the taxonomic diversity of the microbial community to assemble (Mendes et al. 2022).

MEGAHIT (D. Li et al. 2015) and metaSpades (Nurk et al. 2017) are two popular metagenomic assemblers that iteratively assemble the same sample using different k-mer sizes. Larger k-mers improve the resolution of repetitive regions, producing more contiguous sequences, but may introduce errors, while smaller k-mers improve assembly of low-coverage regions but may struggle with repetitive regions (Olson et al. 2017). MEGAHIT is currently the most efficient metagenomic assembler, making it particularly useful for assembly of large or multi-sample datasets (see background section 2.2.2.3, “Multi-sample variants of metagenomic assembly and binning”).

More recently, short-reads strain-aware assemblers such as StrainXpress have become available, which differently from most metagenomic assemblers uses an approach to make the overlap-graph (OG) based approach computationally feasible by first performing local assemblies on priorly clustered reads that are likely to originate from the same species to obtain strain-specific contigs; these local assemblies are then used as seeding to extend them across the clusters (X. Kang, Luo, and Schönhuth 2022). MetaCORTEX (Martin et al. 2023) is a DBG-based assembler that implements a Subtractive Walk algorithm which, instead of removing nodes after path traversal of the graph, reduces their coverage values iteratively, allowing shared k-mers to be incorporated into multiple variant paths and thus capturing strain diversity.

2.2.2.2 Metagenomic binning

The output of short-read metagenomic assembly is a fragmentary collection of contigs of varying lengths, each representing a genomic segment from a microbial species. Metagenomic binning consists of sorting taxonomically unassigned contigs obtained from metagenomic assembly into discrete groups, i.e. *bins* representing putative draft

genomes of individual microbial members in the community. Binning approaches can be divided into two categories, supervised and unsupervised.

Supervised methods assign a putative taxonomical label by aligning contigs, usually with common tools such as BLAST (Altschul et al. 1990), Bowtie2 (Langmead et al. 2009), or HMMER (Finn, Clements, and Eddy 2011), against a database of taxonomically annotated sequences. However, this approach is limited since microbial diversity is vast and reference sequences are not available for the majority of species.

Unsupervised metagenomic binners use machine learning to group contigs based on sequence composition (like GC-content and k-mer frequencies) and read coverage of contigs, or genes they contain, under the assumption that each member of a metagenome will have distinctive overall compositional features, characteristic of that species, and also be present at a specific relative abundance. The rationale is that contigs with similarity in such characteristics are more likely to originate from the genome of the same microbial species.

Combining both features improves contig clustering accuracy, and most current binners use an hybrid approach with various machine learning strategies, and can also use coverage information from multiple samples (see background section 2.2.2.3, “Multi-sample variants of metagenomic assembly and binning”).

CONCOCT (Alneberg et al. 2014) is a hybrid binner that leverages on coverage and composition information across samples by combining a coverage matrix and a tetranucleotide frequency (TNF) matrix, applies PCA to reduce dimensionality, and then uses a Gaussian mixture model (GMM) to cluster the contigs into distinct genomic bins.

MetaBAT2 (D. D. Kang et al. 2019) is a widely used metagenomic binner, known for its low running times and memory requirements (Meyer et al. 2022), that employs a graph-based approach for clustering contigs into putative genome bins. It first constructs an initial graph based on the tetranucleotide frequency (TNF) distance between contigs, then iteratively adds edges weighted by a composite similarity score (S) that integrates TNF, abundance, and coverage correlation information across samples. The final clustering is performed using a modified label propagation algorithm that operates on the constructed graph, considering the edge weights to refine the clusters.

Recent tools have incorporated de Bruijn graph connectivity information to refine metagenomic binning results (Mallawaarachchi, Wickramarachchi, and Lin 2021; Ochkalova et al. 2023). This approach addresses the limitations of many metagenomic binning tools, which typically do not allow contigs to be assigned to multiple bins. Such multi-bin assignment may be biologically relevant for genomic regions shared between

species or strains due to phylogenetic conservation (e.g. 16S rRNA genes) or horizontal gene transfer.

Deep learning methods, such as neural networks, have been implemented in binners such as AAMB (Líndez et al. 2023) and VAMB (Nissen et al. 2021), which use adversarial and variational autoencoders to represent compositional features of contigs into latent spaces and are then clustered using an iterative medoid algorithm.

Similarly to metagenomic assemblers, no single binning tool excels in all aspects. Ensemble binners like MetaWRAP (Uritskiy, DiRuggiero, and Taylor 2018) and DAStool (Sieber et al. 2018) have been developed, combining results from multiple binning algorithms to leverage their complementary strengths and improve overall performance.

2.2.2.3 Multi-sample variants of metagenomic assembly and binning

The quality of MAGs can be comparable to those derived from the assembly of isolate sequencing (Pasolli et al. 2019). However, single metagenomes typically yield MAGs only from the most abundant species, representing a minor fraction of the community. High- and medium-quality MAG recovery necessitates a genomic coverage of 10-20X (Nayfach et al. 2019; Royalty and Steen 2019), but it can vary based on metagenome complexity. Standard single-sample sequencing depths (0-7 Gbp) are generally insufficient to achieve adequate coverage for low-abundance species (Tremblay, Schreiber, and Greer 2022).

Microbiome studies frequently encompass multiple samples from identical environments or subjects to study specific biological conditions. Pooling metagenomes increases the chances of gathering enough reads from those low-abundance species that are shared across samples and increase their genomic coverage. Hence, multi-sample metagenomic assembly and binning strategies for maximizing MAG retrieval have been used in many studies.

Metagenomic co-assembly, i.e assembly of pooled metagenomes, has been useful in capturing more microbial diversity into MAGs when applied to longitudinal and cross-sectional metagenomes from environmental (Delmont et al. 2018; Jégousse et al. 2021; Delgado and Andersson 2022; Haryono et al. 2022) or host-associated biomes, including the human gut microbiome (Stewart et al. 2018; Pasolli et al. 2019). However, pooling samples increases the size of the resulting metagenome, making co-assembly computationally intensive and time-consuming. Highly efficient metagenomic assemblers, such as MEGAHIT or MetaSpades are usually preferred for co-assembly,

and ad-hoc assemblers like MetaHipMer were specifically developed for large-scale co-assemblies.

Pooling samples can exacerbate the challenge of assembling genomically complex metagenomes, as it can introduce reads with sequence variants from additional species or strains that result in branching in the DBG, causing highly fragmented assemblies. Co-assembly performances are influenced by the level of phylogenetic divergence of the microbial community. When fragmentation does not occur, strain diversity of a sample can result in MAGs that are population-level representation of the different strains, with contigs that are likely to retain the haplotypes of the dominant strain (Vosloo et al. 2021). Tools to deconvolve strain-chimeric MAGs into the genomes representing the strain haplotypes, and their accessory genes, such as DESMAN (Quince, Delmont, et al. 2017) and STRONG (Quince et al. 2021) have been developed. DESMAN detects variants across lineage-specific single-copy core genes (SCGs) to identify MAGs that might have originated from multiple strains, and uses co-occurrence of these variants, across multiple samples, to infer strain haplotypes. STRONG is a fully-fledged co-assembly pipeline that performs strain haplotype deconvolution by extracting from the non-simplified de Bruijn graph (HRG, high-resolution graph), initially generated during metaSPAdes co-assembly, the subgraphs of single-copy core genes of each MAG; reads from the samples are threaded onto these subgraphs to obtain per-sample unitig coverages, and this information, together with the graph structure constraints, is used by BayesPaths, a variational Bayesian algorithm, to simultaneously determine the number of strains present, their haplotype sequences across the linked SCGs, and their abundances in each sample. A limitation of these approaches is that the strain resolution phase can be particularly computationally intensive.

When assembling environments where only a dominant strain for each species is typically present, such as in the human gut, strain deconvolution is not strictly necessary and could be avoided.

Merging biologically unrelated samples can degrade co-assembly performances and increase assembly chimeric MAGs that are unrepresentative of any constituent sample's bacterial populations. Choice of samples to co-assemble can be evaluated by metagenome metadata, similarity in metagenome sequence (Churcheward et al. 2022) or phylogenetic composition.

In the context of human gut studies, it was observed that co-assembly of long time-series metagenomes can lead to a sensible increase in the quality and the number of assembled MAGs from the same environment, whereas the gain was limited when only few longitudinal samples or cross-sectional were available (Pasolli et al. 2019). The current view is that while co-assembly is useful for recovering MAGs from

low-abundance species (Stewart et al. 2018; Sczyrba et al. 2017), the quality for the abundant ones is superior when obtained with single-sample assembly (Hofmeyr et al. 2020).

Metagenomic co-binning leverages contig co-abundance profiles across multiple samples to improve the clustering process. By analyzing how contigs, or their genes, co-vary in abundance across different samples, this approach provides additional evidence for grouping sequences that are likely originate from the same genome, enhancing the quality of the resulting bins. Tools like CONCOCT and MetaBAT2 support co-binning natively.

A clear example of how co-variation of genomic features across multiple samples can be used to infer the presence of unknown microbial species is that of the canopy clustering algorithm, originally implemented in MGS-Canopy (Nielsen et al. 2014) and refined in MSPminer (Plaza Oñate et al. 2019). Although not directly focused on binning contigs into MAGs, this alternative method demonstrates how co-abundance patterns across multiple samples can provide biologically meaningful signals supporting the unsupervised detection of known and unknown microbial species. MGS-Canopy generates a catalog of non-redundant genes from the unbinned assemblies obtained from multiple samples and computes the coverage for each gene across all the samples to obtain a gene-depth matrix. Genes presenting similar trends of co-variation across the samples are grouped into canopies, or co-abundant gene groups (CAGs). CAGs that contain a minimum number of genes (> 700), compatible with that of a microbial species are referred to as Metagenomic Species (MGS). Once an MGS has been identified, it is possible to use its co-abundant genes to recruit the respective contigs and use the reads mapping on these to perform de novo assembly to reconstruct the respective genome. This approach has been instrumental in the discovery and study of unknown microbial species (those whose CAGs lack taxonomic annotation) and their relation to host metadata (Minot and Willis 2019; Flemer et al. 2017; Dhakan et al. 2019), especially in cohorts comprising a large number of metagenomic samples.

While co-binning is generally preferred when multiple samples are available, few studies have comprehensively evaluated its performance compared to single-sample binning in terms of MAG quality (Haryono et al. 2022; Salazar et al. 2022; Churcheward et al. 2022).

2.2.2.4 Assessing the quality of assembled genomes

In the context of isolate single genome sequencing, the quality of an assembled genome usually coincides with simple summary statistics assessing the contiguity of the assembly, or in other words, how close the contigs are to the ideal aim of assembling a contig per each chromosome. These metrics include the number of assembled contigs, their average length or their overall sum, or more robust indexes such as the N50 (i.e: the weighted median contig size) or L50 (i.e: the length-based rank of the N50 contig). Contiguity measures are not informative about the quality of the information contained in the contigs and can be used to evaluate how representative a genome assembly is only when a reference genome is available.

Gene-based metrics, based on open reading frames (ORFs) or specific genes, are more useful for evaluating the quality of metagenomic bins that represent fragmented bacterial genomes for which no reference information (e.g.: genome length, taxonomy) is available. The main metrics estimated with these approaches are genome completeness, which estimates how representative a MAG is in terms of gene content with respect to the genome it should represent, and contamination, which estimates whether it contains contigs that should belong to different bins (mis-binning) or that contain sequences from multiple sources (mis-assembly) and should be labeled as chimeric.

CheckM (Parks et al. 2015) is the most popular tool in metagenomics to evaluate the completeness and contamination of prokaryotic bins, and it uses a database of lineage-specific single-copy gene (SCG) markers. SCG are genes that are conserved across a phylogenetic lineage and appear only once in the genome. CheckM computes the completeness of a bin by computing the fraction of how many SCGs, out of those available for the inferred lineage, appear in the bin contigs. The presence of SCGs from discordant lineages, or redundancy of SCGs are used to estimate the contamination of the bin. Bins with a contamination above the 5% and/or a completeness below the 50% are usually discarded, and those that also exceed a completeness of 50% or 90% are respectively categorized as medium- (MQ) or high-quality (HQ) (Bowers et al. 2017). BUSCO is another SCG-based tool that provides markers to estimate also the quality of eukaryotic MAGs (Simão et al. 2015).

When the taxonomy of a MAG can be reliably inferred, and a reference genome is available, MetaQUAST (Mikheenko, Saveliev, and Gurevich 2016) can estimate MAG completeness by aligning contigs to the reference to compute the percentage of bases that align to it.

Methods implementing alternatives to the SCG-based approach to detect contamination and chimerism have been developed. GUNC (Orakov et al. 2021) uses gene-wise

phylogenetic annotation consistency across the full gene complement of its contigs to detect chimeric MAGs. A more straightforward approach to detect where assembly errors occur in the contigs is to scan for inconsistencies in the read pileup (repeat collapse, drops in coverage, inconsistency in paired-end read direction), as implemented in REAPR (Hunt et al. 2013).

More recently, the second version of CheckM (Chklovski et al. 2023) has been developed and uses neural networks and gradient-boosted decision trees, trained on a set of complete and curated reference genomes to avoid relying on SCG, as these might may not fully represent the different source lineages.

2.2.2.5 Taxonomic assignment of MAGs

Taxonomic classification of a MAG is a necessary step to know whether it belongs to a known, referenced taxon or to a potentially new one. In the latter case, it might be useful to infer its higher level of taxonomy, by inferring the last common ancestor (LCA).

The simplest approach to assign taxonomy to a MAG is to align its contigs or subsequences, against a database of taxonomically-annotated reference sequences so to propagate to it the taxonomy of the best-hit. Given that many sequences can appear across multiple genomes, even of highly divergent species, such in the case of horizontal gene transfer or for highly conserved genes, this strategy can produce ambiguous taxonomic annotations, and alignment-based tools often implement the LCA algorithm and majority-voting scores to collapse multiple annotations into a consensus taxonomy. This approach is implemented in tools such as the Bin Annotation Tool (BAT) (von Meijenfeldt et al. 2019), which detects and translates ORFs within the contigs, assign LCA taxonomy to each using a reference database, and compute a majority score to assign an overall taxonomic label to the MAG. Also tools developed for generic metagenome classification, such as the already mentioned Kraken, or Kaiju (Menzel, Ng, and Krogh 2016), can also be used to classify MAGs. A drawback of this approach is its reliance on sequence homology with reference sequences, which in the case of novel genomes are not available. Phylogenetic-based approaches (see background section 2.3.1, “Microbial phylogenetics, genomics, and ecology analyses”) are thus more suitable to infer taxonomy of novel MAGs. These approaches infer MAG taxonomy by estimating its evolutionary divergence from genomes of known bacterial taxa, usually by computing sequence distances based on sets of lineage-conserved gene markers (phylogenetic approach) or whole-genome sequences (phylogenomic approach).

The phylogenetic approach is exemplified in softwares like PhyloSift (Darling et al. 2014). It identifies homologs to 37 conserved gene families in input contigs, aligns them to comprehensive reference MSAs, and places them on corresponding phylogenetic

trees. Using pre-computed mappings, these phylogenetic placements are converted to taxonomic labels, which are then combined for final classification.

The phylogenomic approach is also implemented in the *phylophlan_metagenomic* subroutine from the phylogenetic analysis suite PhyloPhlAn 3.0 (Asnicar et al. 2020). This tool is able to infer the taxonomy of multiple input MAGs, by calculating an all-vs-all whole-genome distance (MASH distance, (Ondov et al. 2016) with a database of > 16,000 SGBs genome representatives that recapitulate more than 10^6 reference genomes and MAGs (see background section 2.3.3, 'The BioBakery workflow and the underlying MetaRefSGB database').

A hybrid approach is implemented in GTDB-Tk (Chaumeil et al. 2019, 2022), which uses 120 ubiquitous prokaryotic marker genes and a pre-computed backbone tree of life (Parks et al. 2018), to infer the taxonomy of the query MAG up to the class level, and species level by integrating ANI with genome representative of that class' species. Ambiguous taxonomic assignments for lower taxonomic levels, for family and below, are resolved with measures of evolutionary divergence. This approach allows to detect whether a MAG belongs to a new microbial taxa or not.

2.2.2.6 Functional annotation of MAGs

Translating nucleotide sequences of reconstructed contigs into meaningful biological knowledge requires parsing sequences to detect the specific genetic elements that define a microbial cell functionalities and metabolism. This task is achieved during the genomic and functional annotation step of microbial genomes and MAGs. Annotation pipelines typically begin by identifying different types of genetic elements in the contigs, such as open read frames (ORFs) and respective coding sequences (CDS), RNA genes (e.g: rRNA, tRNA, sRNA, ncRNA), structural DNA elements (e.g: origins of replication), regulatory regions (e.g: transcription binding sites) and mobile genetic elements (e.g: transposon, prophages). Genes, and genetic elements in general, can be detected through pattern-matching of distinctive sequence motifs. Different types of signatures can be used - for instance, ORF detection may involve recognition of Shine-Dalgarno motifs, regulatory element identification can use Pribnow box sequences, and RNA gene detection often relies on sequence or structural homology to already characterized RNA molecules. Popular tools for gene calling, like Genemark-HM (Lomsadze et al. 2021) or Prodigal (Hyatt et al. 2010) embed complex models to predict non-spurious ORFs and avoid false positives. Although not in the scope of this thesis, a variety of other tools for detection of other types of genetic elements have been developed. Once genes have been detected from a MAG, it becomes possible to infer their putative function or product by finding homology to sequences whose products or functions have already been experimentally or computationally validated. A large number of databases collect and organize different aspects of biological knowledge related to genes. These

range from general-scope databases of genes and proteins (e.g: RefSeq, UniProt, UniRef), and their functionality and metabolism (e.g: KEGG, COG, EggNOG), to more specialized ones describing antimicrobial resistance genes and virulence factors (CARD, VFDB), or specific specific enzymatic activities, such as carbohydrate utilization enzymes (CAZymes). Several user-friendly pipelines that integrate multiple tools for genetic feature detection and annotation have been developed, such as Prokka (Seemann 2014), Bakta (Schwengers et al. 2021), and PGAP (Tatusova et al. 2016). These require only genome files in fasta format as input to produce annotation files by searching multiple reference databases at once.

2.3 Analysis of microbiome data (post-processing analysis)

Post-processing analysis involves the biological interpretation of the output of primary metagenome processing, obtained through reference- or assembly-based methods through statistical tools. The input to post-processing analyses typically consists of matrices of inferred microbial features linked to the respective sample metadata when employing reference-based profiling, or annotated sequence data when utilizing assembly-based approaches. Microbiome data is complex and multidimensional, and multivariate statistical tools are essential for inference and hypothesis testing, as well as supervised or unsupervised machine learning techniques for classification or prediction tasks, or to uncover latent data patterns emerging from underlying biological factors (e.g: clustering). In conjunction with these, dimensionality reduction techniques (e.g., PCoA, UMAP, nMDS, tSNE) and other graphical representations (e.g., heatmaps, ordination plots) are employed to integrate different data types to help in the identification of visually intuitive patterns within the data. Examples of post-processing analyses include identifying microbial taxa differentially abundant between different conditions or associated with continuous parameters, discovering microbiome-based biomarkers with diagnostic potential for diseases, and reconstructing the phylogeny of microbial taxa (Quince, Walker, et al. 2017).

2.3.1 Microbial phylogenetics, genomics, and ecology analyses

Microbial phylogenetics investigates the evolutionary relationships between microorganisms by analyzing similarities and differences in their genetic content. Phylogenetic relationships are usually visualized through phylogenetic trees in which organisms are shown as tips (or leaves), joined by pairwise connections forming a branching structure, in which internal nodes represent the most recent common ancestor of the two descendant lineages, and the sum their branch lengths connecting

two tips indicates their sequence divergence (a proxy for evolutionary distance). Integration of host or species metadata variables into phylogenetic trees enables the assessment of their influence on microbial evolution. This approach has been employed to study biologically significant events, such as person-to-person microbial transmission (Valles-Colomer et al. 2023), tracking of multi-drug resistant species outbreaks in healthcare settings (Manara et al. 2018), and evaluating sequence consistency of contigs assembled from putative subject-specific microbial species, as I will show in this thesis.

Constructing a phylogeny usually begins with the selection of a single (e.g., 16S rRNA gene) or multiple orthologous genes, whose evolutionary conservation level is determined by the tree's scope. Ancient and slowly evolving genes are utilized for studying highly divergent species, while highly variable genes are employed for closely related species. Sequences are compared through multiple sequence alignment (MSA), which is subsequently used to infer the phylogenetic tree using tools like RAxML-ng (Kozlov et al. 2019) or FastTree (Price, Dehal, and Arkin 2009, 2010).

PhyloPhlAn3.0 is a pipeline that wraps this entire process by directly taking as input sets of microbial genomes (or their annotation-derived proteomes) to generate phylogenies for microbial lineages across different scales of genetic divergence. For each step of the phylogeny building, PhyloPhlAn3.0 allows users to choose between multiple softwares and tune specific parameters. Construction of genome phylogenies of highly divergent microbial lineages relies on the identification, selection and MSA of 400 protein prokaryotic universal markers across the genomes in input; single-gene MSAs are then concatenated into a single and wide alignment, which is further refined (e.g: trimming, removal of local misalignments) to improve phylogeny reconstruction. When the task is to reconstruct phylogenies at strain-level resolution, the same process is applied by using SGB-specific marker genes markers. In this way, even fine changes in the phylogenetic structure of a microbial species can be tracked to study emerging associations with available metadata.

2.3.2 Microbial diversity analysis (alpha- and beta-diversity)

Ecological aspects of microbial communities are a crucial part of metagenomics. Microbial diversity analysis aims to describe the taxonomic composition and the relative abundances of microbial lineages within individual communities (alpha-diversity), as well as compare and identify relationships among them (beta-diversity).

Alpha-diversity consists in estimating how many microbial taxa (e.g: species, genera, phyla) are present in a community (*richness*), and the relative abundance proportion of

each (*relative abundance*). Apart from simply using richness, alpha-diversity can be evaluated by calculating indexes that also take into account the relative abundance of each species. Common abundance-aware alpha-diversity indexes include the Gini-Simpson, a numerical value ranging from 0 to 1 that can be interpreted as the probability that two bacterial cells, randomly drawn from the community, will belong to different species, or the Shannon-Wiener, which measures diversity as the uncertainty of correctly guessing the taxa of a random bacteria drawn from the community. Different alpha-diversity indexes can give more or less weight to dominant or rare taxa.

Beta-diversity analysis compares taxonomic and compositional diversity occurring between the profiles of two or more distinct microbial communities. It consists in calculating similarity (or dissimilarity) indexes between numerical vectors describing the taxonomic composition of a sample's microbial community. Among the mostly used, the Jaccard index is used to compute beta-diversity between binary vectors encoding for simple presence-absence of microbial taxa in a profile, whereas other indexes, such as the Bray-Curtis (Bray and Curtis 1957) or Aitchison's (Aitchison 1982), also account for their specific relative abundances. Other widely used beta-diversity metrics worth mentioning are the UniFrac and weighted UniFrac indexes, which incorporates information on phylogenetic relatedness between taxa of the compared community profiles (C. Lozupone et al. 2011).

Beta diversity of multiple metagenomes is usually summarized as a matrix of pairwise distances, which can then be usually visualized with ordination techniques in a lower dimensional space with techniques like Multidimensional Scaling (MDS) or Principal Component Analysis (PCoA), and then annotated with sample metadata to track community structural changes across spatial and temporal gradients, sample conditions (e.g: subject, disease) or experimental factors (e.g: sequencing depth).

2.3.3 Discoveries and applications of MAGs

Catalogs organizing thousands of MAGs assembled from gut metagenomes within a coherent taxonomic framework and their metadata enable different applications (see background section 2.3.3, "The BioBakery workflow and the underlying MetaRefSGB database").

One of these is expanding the known microbial diversity of the gut microbiome at the global level. In 2017, Parks et al. collected over 8,000 MAGs from undersampled environments and used a genome-based approach to expand the phylogenetic tree of life, discovering that many MAGs were the first representatives of new bacterial phyla. In 2019, Pasolli et al. took a similar approach on a larger scale, focusing on the human microbiome, and recovered more than 150,000 MAGs from the assembly of 9,428 metagenomes spanning different body sites, ages, geographies, and lifestyles. MAGs

were dereplicated into 4,930 SGBs, 77% of which were uncharacterized (uSGB), having no reference genome available in public databases. Interestingly, most uSGBs are harbored mostly by populations underrepresented in gut metagenomic studies.

MAGs allow to study of microbial species at a genomic population scale. In the same study, a set of uSGBs, belonging to a neglected genus-level lineage of the *Clostridiales* order*,* was discovered to be globally widespread in humans, often at high relative abundances (1.14% on average). In particular, one of these was found to be the first representative of a new candidate genus *Cibionibacter*, and analysis of its pangenome revealed that strains obtained from non-westernized populations harbored specific metabolic functionalities, such as the ability to metabolize tryptophan or synthesize vitamin B12, that reflect their adaptation to the different lifestyles of their hosts (Pasolli et al. 2019).

The availability of multiple MAGs or a single clade facilitates high-resolution comparative genomics analyses, enabling the exploration of genetic and functional intra-clade diversity. De Filippis et al. performed a genome-wide analysis of more than 3,000 *Faecalibacterium* MAGs and reference genomes to show that this genus comprises 22 different SGBs, some of which are human-specific and enriched for specific diseases, age categories, or host geography (De Filippis, Pasolli, and Ercolini 2020).

Pangenomic analysis of same-species MAGs allows tracking differences in the functional and metabolic potential between different strains down to gene-level resolution, which can be used to guide experimental phenotyping of microbes. Phylogenetic and pangenomic analysis of over 1,300 MAGs of *Eubacterium rectale*, a highly prevalent species of the human gut, revealed genomic divergence in European strains, which lack the motility operons present in their Eurasian, African, and Asian counterparts, despite overall genetic closeness. Immotility of the European strains was further confirmed *in vitro*. Conversely, the European lineage exhibits an expanded genomic repertoire for carbohydrate metabolism, encoding a significantly broader array of genes involved in the catabolism of diverse polysaccharides, suggesting a potential metabolic adaptation to compensate for the loss of motility (Karcher et al. 2020)

Databases of MAGs are a useful resource for identifying microbial genes implicated in human health through machine learning or deep learning methods, especially in the field of antibiotic resistance (AR) and how to counteract it (Lund et al. 2023; Khedher et al. 2020; Ma et al. 2022), or encoding for molecules with clinical or industrial applications.

2.3.4 The BioBakery workflow and the underlying MetaRefSGB database

MetaRefSGB is a comprehensive genomic database that currently includes over 10^6 genomes, combining reference genomes from NCBI with MAGs assembled from publicly available metagenomic datasets (Blanco-Míguez et al. 2023). The database organizes genomes into SGBs based on hierarchical clustering of all-versus-all MASH genetic distances, using a 5% intra-cluster cutoff (Jain et al. 2018). SGBs are further grouped into genus-level (GGB) and family-level (FGB) bins. Each SGB is assigned a unique numerical ID and categorized as either known (kSGB) if it contains NCBI reference genomes with propagatable taxonomy, or unknown (uSGB) if it doesn't.

The MetaRefSGB database serves as the foundation for the BioBakery workflows, a comprehensive suite for metagenomic analysis (Beghini et al. 2021). Through systematic functional genome annotation, the database facilitates the definition of SGB-specific pangenomic catalogs, thereby enabling the integration of MAGs into reference-based approaches. This integration empowers various metagenomic analyses: MetaPhlAn utilizes SGB-specific markers for detection and quantification of SGBs, including unknown species; StrainPhlAn (Truong et al. 2017; Blanco-Míguez et al. 2023) and PhyloPhlAn (Asnicar et al. 2020) employ these markers, which also represent core genes phylogenetically conserved at the SGB level, to reconstruct phylogenies respectively from metagenomes, or already available MAGs; PanPhlAn (Scholz et al. 2016) performs strain-level metagenomic profiling by identifying which genes from a species pangenome are present; lastly, HUManN (Beghini et al. 2021; Franzosa et al. 2018) performs reference-based functional profiling of entire metagenomes. Thus, the continued increase in MAG availability in MetaRefSGB significantly augments the capacity in reference-based metagenomic analyses.

Chapter 3 - Performance Assessment of Metagenomic Assembly and Co-assembly Strategies for Gut Microbiome Taxonomic Analysis

I report here the manuscript of the paper entitled “Performance Assessment of Metagenomic Assembly and Co-assembly Strategies for Gut Microbiome Taxonomic Analysis” submitted to a peer-reviewed journal. My contribution to this work, as already mentioned, involved the writing of the whole manuscript and most of the code used for the analyses presented here.

Davide Golzato[1], Camila Alvarez-Silva[2], Gillian Donachie[4], Moreno Zolfo[1], Federica Armanini[1], Francesco Asnicar[1], Mani Arumugam[2], Alan Walker[3], Nicola Segata[1,4]*, Mireia Valles-Colomer[1,5]*

1. *Department CIBIO, University of Trento, 38123 Trento, Italy.*
2. *Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 2200, Denmark*
3. *Rowett Institute, University of Aberdeen, Aberdeen, Scotland, AB25 2ZD.*
4. *Department of Experimental Oncology, IEO European Institute of Oncology IRCCS, 20141 Milan, Italy.*
5. *Department of Medicine and Life Sciences, Universitat Pompeu Fabra, 08002 Barcelona, Spain.*

3.1. Abstract

Metagenomic assembly is used to reconstruct genomes of known and unknown members of a microbiome sample from metagenomic shotgun sequencing libraries. While powerful, the quality of metagenomic assembly depends on sequencing depth and on several other intrinsic characteristics of the microbiome sample and of the sequencing technology. Variants such as co-assembly that exploits multiple metagenomic samples for the same task have been proposed to overcome some of these limitations. However, comprehensive evaluations of (co)assembly performance as a function of ecological and technical features have not been performed so far. Here, we assess the performance of metagenomic (co)assembly and binning on a newly generated set of closely related gut metagenomic samples that were pooled into combinations of increasing sequencing depths and microbiome complexities. We show that metagenomic co-assembly procedures can improve chances of reconstructing metagenome-assembled genomes (MAGs) of low-abundant species over standard metagenomic assembly by increasing genomic coverage to an empirical minimum of 11x. Despite this improvement, even at the highest sequencing depth tested here (75 Gbp) a substantial portion of detectable species is not retrieved, and we show that pre-sequencing steps can also influence the assembled biological diversity. Notably, co-assembling samples from different individuals can lead to the formation of chimeric MAGs. Co-assembly is instead more effective on longitudinal samples when paired with multi-sample binning (co-binning). Our work provides a guide for users to evaluate which metagenomic assembly approach to use for specific tasks and to set informed expectations on the performance of the approach under a set of microbiological and technical variables.

Importance. Characterizing the genomes of microbiome species in the human gut microbiome is key to enabling downstream analyses to investigate their role in human health. Metagenomic assembly enables reconstructing the genomes of multiple microorganisms but while it can retrieve high-quality genomes, large parts of the biodiversity in a metagenomic sample are typically left unexplored and several reconstructed genomes might be of insufficient quality. Identifying the optimal settings for assembly and co-assembly strategies will first inform further improvement of both techniques and ultimately facilitate the characterization of microbiome species with influence on human health and disease. In this study, we found that only when closely related samples are available, co-assembly can be exploited to improve genome reconstruction, as otherwise may lead to chimeric MAGs.

3.2. Introduction

Characterizing the human gut microbiome has been the focus of many efforts in modern microbiology research, due to the tight association between microbiome composition and human health (Human Microbiome Project Consortium 2012; Thomas et al. 2019; Ghoshal et al. 2012; Palmu et al. 2020; Ghensi et al. 2020; Menni et al. 2020). However, the high biodiversity in the intestinal tract, together with the specific and strict metabolic growth requirements of many gut bacteria, limit the success of isolation and cultivation approaches, and a majority of them remain uncultivated (Pasolli et al. 2019; Diakite et al. 2020). Metagenomics applies high-throughput DNA sequencing directly to the entire genomic content of an environmental sample, and can characterize many gut microbial species simultaneously while bypassing isolation and cultivation steps (Jo Handelsman 2004; Quince, Walker, et al. 2017; Oulas et al. 2015).

Implementations of *de novo* genome assembly algorithms in a metagenomic setting have made it possible to reconstruct metagenome-assembled genomes (MAGs), i.e, genomes of individual microbes in a microbial community (D. Li et al. 2015; Nurk et al. 2017; Peng et al. 2012). To reconstruct MAGs, metagenomic short reads are first subjected to standard assembly approaches to obtain longer sequences, i.e. *contigs*. Contigs are then sorted into *bins* (binning step) based on sequence composition and coverage patterns that are supposed to represent full microbial genomes. Finally, methods to assess the completeness and potential contamination of MAGs are applied to perform quality control and retain only MAGs of sufficient quality. Under many circumstances, the quality of MAGs can be similar to that of genomes originating from isolate sequencing (Bowers et al. 2017; Pasolli et al. 2019), allowing new species discovery and *in silico* genomic and functional characterization (Hall et al. 2017; Anyansi et al. 2020; Karcher et al. 2020). The increasing availability of public catalogs containing MAGs reconstructed from metagenomes obtained from various sources, including humans, animals, and different environments, has greatly improved reference databases and increased the mappability of the microbiome (Pasolli et al. 2019; Blanco-Míguez et al. 2023; Almeida et al. 2021).

The number and the quality of MAGs that commonly employed pipelines retrieved from the assembly of complex metagenomic libraries are affected by technical and biological factors in ways that are not entirely predictable. First, assembly requires higher sequencing depth than that employed in some studies (Meyer et al. 2021). Next, for many species, coverage can be too low or too uneven along a genome, biasing the portion of the microbiome from which quality MAGs are reconstructed. Additionally, pre-sequencing steps, such as sample collection, storage, and DNA extraction methods, can selectively enrich or deplete certain microbial taxa or genomic regions and influence their genomic coverage (Forry et al. 2024; Szóstak et al. 2022; Bowers et

al. 2015). As a result, some species may remain systematically undetected in metagenomic assembly efforts (Hallmaier-Wacker et al. 2018; Probst et al. 2015; Wesolowska-Andersen et al. 2014).

Strategies to use metagenomic reads from multiple samples to increase the genomic coverage of microbial species and improve MAG retrieval, i.e metagenomic co-assembly, have been proposed as an approach to maximize MAG reconstruction and quality (Qin et al. 2010; Curtis et al. 2012). Especially for rare and low abundant species, co-assembly thus holds the potential to increase the chances of obtaining high-quality MAGs. However, the effect of introducing samples from different sources increases metagenome complexity, adding to the already poorly understood set of potential factors influencing the quality and the number of retrieved MAGs. In addition, using multiple samples can be also used to improve the step of contig binning by incorporating co-abundance information (co-binning) (Churcheward et al. 2022; D. D. Kang et al. 2019). However, co-assembly and co-binning can also result in chimeric contigs and MAGs, ignoring the strain variation between hosts (Ayling, Clark, and Leggett 2020; Nayfach et al. 2019). To date, the impact of these aspects when applying metagenomic co-assembly in the context of human gut studies has not been investigated.

Here, we investigate under what conditions co-assembly maximizes the reconstruction of quality MAGs of the species in a community when compared to single-sample assembly, by exploring the co-assembly of combinations of cross-sectional and longitudinal samples from five individuals. We find that MAG reconstruction scales with sequencing depth and thus improves with co-assembly, although even at sequencing depths well above the average, MAG reconstruction still fails for some species in the community. However, we caution against co-assembling samples from different subjects, as this produces chimeric MAGs that appear high-quality but are not biologically meaningful.

3.3. Results and discussion

3.3.1. Metagenomic assembly complements reference-based profiling by uncovering novel gut microbiome taxa

We first evaluated the performance of metagenomic assembly in detecting the microbial genomes present in a metagenome. To do so, we assessed how the number of reconstructed MAGs scales with increasing sequencing depths. We used a cross-sectional dataset of metagenomes (N=61) obtained from stool specimens of five different subjects, each sequenced up to 12 times after using variations in sample preparation protocols (i.e. *technical variants,* Methods) [Fig. S1; Table S1]. We combined and merged metagenomes of technical variants into combinations of different sizes (k = 1,2,4,6,9,12; total n = 376 combinations), resulting in a gradient of sequencing depths (see Methods). Read depth of metagenomes ranged from a median of 4.99 Gbp (IQR = 2.87 Gbp) of pre-processed and quality-controlled reads for single metagenomes which is the typical depth of most metagenomic cohorts, to a median of 67.4 Gbp (IQR = 1.61 Gbp) reads for sets of 12 technical variants [Table S2].

We used an assembly-free reference-based strategy to estimate the total number of species present in a metagenome and thus the upper limit of MAGs that could potentially be assembled in the metagenome (considering that only one MAG for each species is typically successfully assembled) (Ruscheweyh et al. 2021; Wood, Lu, and Langmead 2019; Nielsen et al. 2014). Reference-based microbiome profiling typically allows the detection of microbial species with lower coverages than those required to assemble their genome as long as the biodiversity in that environment is well described and represented by the reference database (Pasolli et al. 2019). Since the gut microbiome has been extensively studied, it is an optimal environment to assess the performance of metagenomic assembly with reference-based profiling as a standard, considering large databases of also uncultivated microbes are available (Nelson et al. 2010; Almeida et al. 2019; Nayfach et al. 2019; Stewart et al. 2019; C. Y. Kim et al. 2021). Reference-based profiling with MetaPhlAn 4 (Methods) (Blanco-Míguez et al. 2023; Segata et al. 2012; Beghini et al. 2021) of all the 376 technical variant combinations of the 5 individual gut environments identified a total of 974 distinct species-level genome bins (SGBs, (Pasolli et al. 2019) (Methods). In contrast, co-assembly of these 376 combinations produced 109,913,928 contigs, which were sorted into 55,565 genomic bins, resulting in 27,786 high (HQ) and medium (MQ) quality MAGs after filtering for completeness and contamination (see Methods, section "Assessing completeness and contamination of MAGs"). We assessed the taxonomy of

these MAGs by computing their genomic distance to genomes in MetaRefSGB, a database that clusters at a 5% genomic distance more than 10^6 genomes and MAGs (Blanco-Míguez et al. 2023). SGBs are classified into known and unknown SGBs (kSGB and uSGB respectively) according to whether they contain at least a reference genome taxonomically annotated NCBI or not (Pasolli et al. 2019). MAGs that could not be assigned to any of the kSGB or uSGBs (genomic distance > 5%) were re-clustered into distant SGBs (dSGB), being putative so far undescribed species (Methods). Overall, the 27,786 MAGs assembled in this study belonged to 425 unique SGBs, 397 of which were also detected by reference-based profiling [Fig. 2B]. Of these, 168 (44.23%) were kSGBs, 198 (46.59%) uSGBs and 39 (9.18%) dSGBs [Fig. 2C]. Although the true number of unique SGBs present across each subject's communities is unknown, metagenomic (co)assembly and binning allowed to cover only the 41% of the SGBs that were detected by reference-based profiling across all subject (co)assemblies. Additionally, 28 SGBs were detected with assembly but not with the reference-based approach. This is because they contained an insufficient number of high-quality genomes for markers to be inferred (n = 26), with only some exceptions that failed to be detected despite the specific markers being present in the MetaPhlAn4 reference database (n = 2).

3.3.2. Deep sequencing boosts MAG reconstruction but does not saturate metagenomic assembly capabilities

Next, we assessed how species detection and assembly scales with sequencing depth. With reference-based profiling we detected a median of 322 SGBs (IQR= 64) for single samples and 490 SGBs (IQR= 86) in sets of 12 technical variants, meaning that a 12.2-fold increase in read depth resulted in a 1.5 average fold increase in richness [Table S3]. Single-sample species richness estimates were thus, on average, 66% of the SGBs richness detected in 12-sample same-individual metagenomes. Expectedly, the difference was even more striking with metagenomic assembly, which yielded an average of 30.72 ($CI_{95\%}$ = [28.69, 32.74]) HQ and MQ MAGs on single samples, while co-assembly of 12 technical variants yielded an average of 123.8 ($CI_{95\%}$ = [110,1.37]), leading to a 4-fold average increase [Fig. 1A]. At equal sequencing depths, reference-based profiling detects a median of 5.6 times (IQR = 2.9) more SGBs than those that can be assembled [Fig. 1C]. Although assembly-based profiling detects fewer species compared to reference-based profiling, the former scales better with increasing sequencing depth, showing a more significant improvement in the number of assembled species.

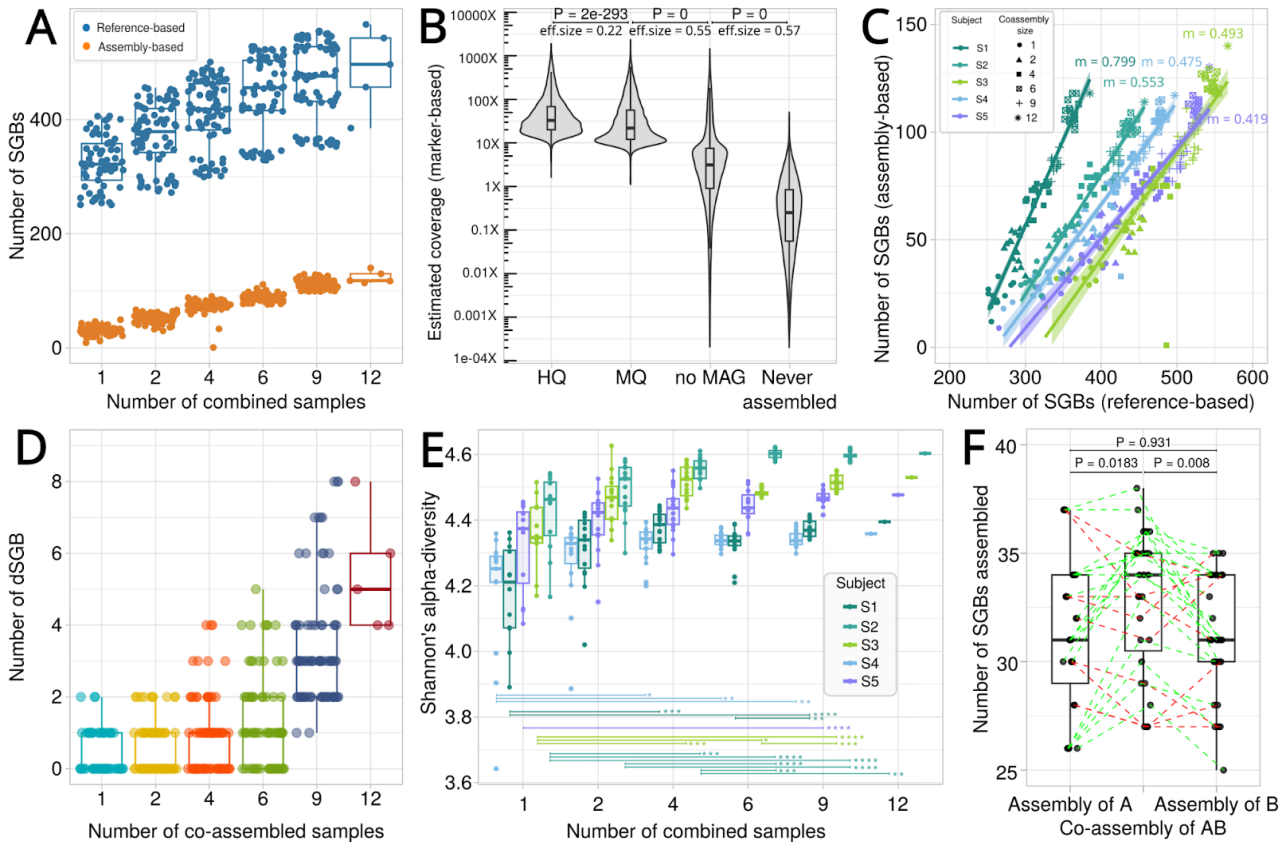

Fig. 1 (A) The number of species-level genome bins (SGBs) detected, using a reference-based approach (upper boxplots, in blue), and reconstructed with the assembly-based approach (lower boxplots, in orange) across combinations of different sizes (Statistics, Table S5A). (B) Distributions of the estimated genome coverages for SGBs that were assembled into HQ and MQ MAGs, those that were not assembled but detected in the respective sample (no MAG), and SGBs that were detected by reference-based profiling but never assembled (Wilcoxon rank-sum test, P < 2.2e-16, Methods). (C) Comparison between the number of SGBs assembled into MQ/HQ MAGs (y-axis) and the number of SGBs detected by reference-based profiling (x-axis), across the 376 technical variant combinations, stratified by subject. Linear regressions for each subject (color-coded) represent the relationship between SGBs detected and assembled across combinations of different sizes (shape-coded) (Statistics, Table S5C) (D) Number of dSGBs detected across co-assemblies of different sample sizes (group means and CI_{95}0.44±0.152, 0.547±0.166, 0.89±0.234, 1.342±0.307, 3.413±0.379, 5.4±2.078) (Table S5D, Methods). (E) Variation of community alpha-diversity (Shannon diversity) across co-assemblies of different sizes and subjects (Wilcoxon rank-sum, *: P.adj < 0.1, **: P.adj < 0.01, ***: P.adj < 0.001, ****: P.adj < 0.0001; Table S5E, Methods). (F) Boxplots comparing the number of MAGs retrieved from the assembly of individual technical variant pairs (Assembly of A and B) to those obtained from the co-assembly of their combinations (Co-assembly of AB), all rarefied to the same sequencing depths. Dots in A and B are matched to the co-assembly of their combination by dashed lines, and the color indicates whether the co-assembled combination yields a greater (green, n = 17), lesser (red, n = 9), or equal number (black, n = 1) of MAGs compared to the two individual technical variants used to form it. At similar read depth, co-assembling combinations of two technical variants (A and B, i.e. AB) allows the reconstruction of more SGBs in 17 out of the 27 combinations when compared to single-sample assembly (Wilcoxon signed-rank test, median difference = 3, Statistics at Table S5F).

Reference-based richness curves appear to saturate earlier than the assembly-based ones, with the latter starting to show separation among the five subjects only at high sequencing depths, similarly to what we observe in reference-based profiling at low sequencing depth [Fig. S2]. This observation suggests that a significantly greater sequencing effort, beyond the maximum depth of 75 Gbp used in this study, may be necessary to reach a plateau in assembly-based richness. Supporting this observation, the fraction of SGBs that are detected through reference-based profiling, represented by a MQ or HQ MAG obtained by (co)assembly and binning of the co-assembled combinations, ranges from a median of only 4.86% (IQR = 1.848) in single technical variants, to 17.12% (IQR= 2.13) in full combinations, showing that MQ and HQ MAGs recovered from metagenomic (co)assembly and binning represent a very limited fraction of the gut microbiome. These findings indicate that even high sequencing depths are insufficient to reconstruct genomes of adequate quality for the majority of the microbial diversity present in a gut metagenome [Fig. S3].

Assembly statistics show that increasing the number of co-assembled samples improves assembly contiguity, with median N50 values showing a 3-fold increase from single technical variants to the largest combinations (P.adj = 5.11e-06, effect size = 0.51). While the total assembly length grows sublinearly with sequencing depth (increasing from a median of 0.22 Gbp to 0.79 Gbp), the size of the largest contigs plateaus, indicating that increased sequencing depth alone cannot resolve the inherent assembly fragmentation of short-read data, probably stemming from low-complexity and redundant regions in the microbial genomes [Fig S4 A].

For the 27,786 high- and medium-quality (HQ/MQ) MAGs, we observe improved contiguity with co-assembly, as evidenced by an increase in median N50 from 14,765 bp to 30,772 bp (Wilcoxon signed-rank test, P.adj < 2e-16). However, the median cumulative contig length within MAGs remains stable regardless of co-assembly size, suggesting that increased sequencing depth does not necessarily yield more complete MAGs [Fig S4 B].

The percentage of input reads mapping to assembled contigs quickly reaches a plateau near 100%, indicating efficient read incorporation into assemblies. However, only a small fraction of the assembled sequences are successfully binned into MAGs. Specifically, less than 50% (IQR = 5.02%) of the assembled sequence length eligible for binning (contigs > 1,500 bp) is incorporated into medium/high-quality MAGs [Fig. S5]. This observation highlights that contig binning, rather than assembly itself, represents a major bottleneck in recovering HQ/MQ MAGs from metagenomic data.

Although the sequencing depth used in our cohort far exceeds that of typical metagenomic studies [Fig. S6], our findings suggest that depths greater than 75 Gbp

are still necessary to assemble under-characterized components of the gut microbiome, in line with another study from (Tremblay, Schreiber, and Greer 2022).

3.3.3. Metagenomic (co)assembly reconstructs more MAGs of putative new species

We then assessed if co-assembling metagenomic combinations with increasing sequencing depth improves the discovery and genome reconstruction of novel species. Assembly of genomes of species that are not yet present in databases and thus cannot be detected with reference-based profiling remains difficult, especially for species that are usually present at low relative abundances. Our analysis shows that the assembly of novel species in our cohort requires exceptionally high sequencing depths. In fact, we observe that almost no dSGBs are assembled in single-sample assemblies (median: 0 dSGB) [Fig. 1D], while many are assembled from metagenomic combinations of 9 and 12 samples (median: 3 and 5 dSGBs). Indeed, 21 of the dSGBs were exclusively reconstructed in co-assemblies composed of 9 or 12 technical variants. A set of 175 SGBs is ubiquitously assembled across co-assemblies of all sizes, among which 105 are kSGBs, 68 uSGBs, and only 2 are dSGBs. No SGBs were found to be uniquely assembled only in single technical variants, meaning that all are always included in at least a co-assembly combination of > 2 samples, and the 51.7% of SGB genomes assembled in this study could be retrieved only by co-assemblies [Fig. 2E; Table S3A-C]. The categories of co-assemblies merging 9 and 12 technical variants account for most of the SGBs reconstructed across the whole cohort (91.3% and 84.47%).

3.3.4. Underestimated alpha-diversity estimates linked to low sequencing depths

Sequencing depth can influence microbial diversity estimates that incorporate taxa relative abundances (Ramakodi 2021; C. A. Lozupone et al. 2013). To understand the impact of these two aspects, we evaluated how alpha-diversity of the same community changes across metagenomic combinations with sequencing depth. For each of the 376 metagenomic combinations, we computed alpha-diversity indexes for both SGB relative abundance profiles, obtained with reference-based profiling (Shannon, Simpson's and Inverse Simpson indexes, see Methods), and presence-absence profiles, for both the assembly- and also the reference-based profiling (Jaccard, see Methods).

When examining alpha-diversity of single technical variants, estimates did not change drastically (CV of Shannon's intra-subject alpha-diversity = 0.034), indicating that the

observed overall composition of the communities were robust across the different pre-sequencing protocols and sequencing depths conditions. Although small, we report that these differences were attributable significantly to the DNA extraction protocol (ANOVA, $F(3,57) = 6.684$ $P = 0.0006$) rather then how they were stored (ANOVA, $F(3,57) = 0.4$, $P = 0.808$) or the sequencing depth at which the technical variants were sequenced (Spearman's correlation coefficient, $P = 0.18$). When considering the full range of sequencing depths in larger combinations, the impact of sequencing depth on alpha-diversity became statistically significant (Spearman's correlations, subject average $rho = 0.55$, $P < 6.83e-05$). While Shannon's diversity was only slightly higher in larger combinations compared to single variants, sequencing depth proved more crucial for detecting low-abundance species.

Full combinations detected an average of 93 additional SGBs per subject that were missed in single variants. Although these additional SGBs cumulatively represented only 0.826% of the total abundance, they led to a 33.38% increase in observed richness, compared to just 4.06% increase in Shannon's diversity. These findings suggest that a sequencing depth of ~5Gbp, as previously reported (Zaheer et al. 2018), is sufficient for capturing the dominant members of gut microbial communities, while deeper sequencing is needed to detect rare species [Fig. 1E].

3.3.5. Sample handling protocols influence taxonomic composition and MAG reconstruction

We next evaluated beta-diversity to understand whether technical variants from the same subject, while showing consistent overall composition, maintained similar taxonomic profiles. In fact, while two samples could have a similar alpha-diversity, the actual species that contribute to it might be different, or present at different relative abundances (Jost 2007). Different DNA extraction methods and storage protocols might introduce biases towards certain microbial taxa, altering their observed relative abundance despite originating from the same microbial community (Hallmaier-Wacker et al. 2018; Poulsen et al. 2021).

We assessed beta-diversity using both relative abundance-based approaches (Aitchison's and Jaccard distances on SGB profiles) and presence-absence profiles of assembled SGBs (Jaccard distances). Analysis of reference-based profiles from single technical variants revealed clear subject-specific clustering, indicating distinct taxonomic signatures for each individual's community. After controlling for subject effects, we evaluated how sequencing depth and pre-sequencing protocols influenced beta-diversity variability. While sequencing depth showed no significant impact

(PERMANOVA, $R^2_{sequencing\_depth}$ = 0.031, P = 0.217), effect of sample handling protocols on intra-subject variation was statistically significant (PERMANOVA, P < 0.001, $R^2_{Storage\_protocol}$ = 0.046, $R^2_{DNA\_extraction\_kit}$ = 0.039) [Fig. S7]. This suggests that within the typical sequencing depths of our technical variants (~5Gbp, range: 1.45-10.8 Gbp), differences in taxonomic composition are primarily driven by pre-sequencing protocols rather than sequencing depth.

This observation is supported by the detection patterns of SGBs across technical variants. In 706 cases, SGBs were inconsistently detected across technical variants of the same sample. Notably, most of these inconsistencies (682/706) were attributed to pre-sequencing protocols rather than sequencing depth, as evidenced by detection in lower-depth samples but absence in higher-depth ones. Additionally, some of these cases showed substantial variation, such as *E. hormaechei* (SGB10130) in subject S1, with up to 11% difference in relative abundance between samples where it was detected versus where it was not. Even in SGBs that are consistently detected, there was substantial variability in their relative abundance between technical variant profiles, with a median coefficient of variation of 0.48 (IQR = 0.22).

Extending the analysis to all 376 metagenomic combinations, we computed beta-diversity using both abundance-based (Aitchison's and Jaccard) and presence-absence approaches for assembled SGBs [Fig. S8]. As expected, subject of origin explained most variability (PERMANOVA, $R^2_{subject}$ > 0.88, P < 0.01), though this was less pronounced in assembly-based profiles ($R^2_{subject}$ = 0.53, P < 0.01). Sequencing depth explained a limited but statistically significant fraction of beta-diversity variability, with stronger impact on presence-absence based analyses due to the progressive detection of low-abundance species at higher depths.

A large fraction of variability in the assembly-based beta-diversity is left unaddressed by variables such as subject and sequencing depth (PERMANOVA, Residual R^2 = 0.39), meaning that taxonomic profiling performed exclusively looking at which SGBs are recovered as MQ/HQ MAGs is not consistent, and other variables come into play. In fact, although the number of MAGs correlates with sequencing depth (Pearsons' correlation = 0.78, P = 1.02e-13), some SGBs reconstructed from lower-depth technical variants are not consistently reconstructed in higher-depth variants. Still, clustering by subject in the assembly-based beta diversity is less clear than in reference-based beta diversity, especially for the single technical variant profiles, and the impact of sequencing depth appears to be progressively stronger in driving subject separation. However, This observation underscores the need for higher sequencing depths to effectively capture the full metagenomic diversity specific to each individual.

Given the significant effect of pre-sequencing protocols, aside from sequencing depth, on the SGBs we can reconstruct as MQ/HQ MAGs, we wanted to specifically discuss its impact on metagenomic assembly. To do so, we compared the number of MAGs retrieved by the assembly of single technical variant pairs and co-assembly of their combination, both rarefied to the same sequencing depth (see Co-assembly of technical variant pairs, Methods). Co-assembly of two technical variants generally yielded more or the same MAGs compared to sequencing a single sample at double the depth in the majority of comparisons (63% of cases, 17 out of 27 pairs) [Fig. 1F]. This highlights the potential of co-assembly of technical variants not only to increase the genomic coverage of the microbial species, but also to mitigate pre-sequencing biases arising from the use of a single protocol for the handling of the samples.

3.3.6. Species relative abundance is associated with the likelihood of assembling a genome

While the overall fraction of microbial diversity uncovered is highly dependent on sequencing depth, we expected some microbial genomes to be assembled more easily than others. By comparing the fractions retrieved with assembly to those with the reference-based approach, we found that species for which a MAG could be assembled in a given sample had significantly higher relative abundances (average = 0.877 %, n = 27,531, S.E = 0.008%) than those for which no MAG was not assembled despite being detected (average = 0.219 %, n = 57,413, S.E = 0.002%) (Wilcoxon signed-rank test, eff.size = 0.57, P < 2.2e-16). In addition, the SGBs that were detected but never assembled in the whole cohort had an even lower average relative abundance (average = 0.0111 %, n = 69,251) [Fig. 2D; Fig. S10]. In this way, using the average sequencing depth of a metagenome (curatedMetagenomicData, Fig. S6) would allow successful assembling genomes of species that have a median relative abundance of 1.49% (median absolute deviation = 1.62%). Increasing sequencing depths allows successfully assembling lower-abundance species, ranging from an average relative abundance of 0.77% (S.E = 0.022 %) with 30-35 Gbp to 0.48 % at 70-75 Gbp (S.E = 0.065 %) [Fig. S11; Table S5].

The positive relationship between relative abundance and chances of retrieving a MAG can be observed in species like *Coprococcus eutactus* and *Bifidobacterium adolescentis*, which can be easily assembled across single technical variants and co-assemblies in those subjects where it is present at high relative abundance, but never in those where it is present at low abundances, even in co-assemblies of bigger sizes [Fig. 2A; Fig. S9].

Besides the relative abundance with which the species is present in a metagenome, we expected its sequencing coverage would also determine whether it is successfully assembled. We estimated the coverage of the 974 SGBs (see Methods) that were detected in this cohort, including that of the 394 SGBs that were also assembled, and detected a significant association between coverage and the reconstruction of an HQ/MQ MAG (Wilcoxon rank-sum, P < 2e-16, eff.size = 0.22). On average, the minimum coverage for a species being successfully assembled is 11.17X (S.E = 0.241X) [Fig. 1B; Fig. S12], although the median coverage for MQ MAGs was 58.1X (IQR= 44.3X), and for HQ MAGs, of 32.9X (IQR= 48.20X). Higher sequencing coverage is thus associated with the quality of the reconstructed MAGs. Our estimates on the minimum coverage required to assemble a MAG are consistent with those already observed by (Nayfach et al. 2019) from the assembly of 3,810 cross-sectional metagenomes.

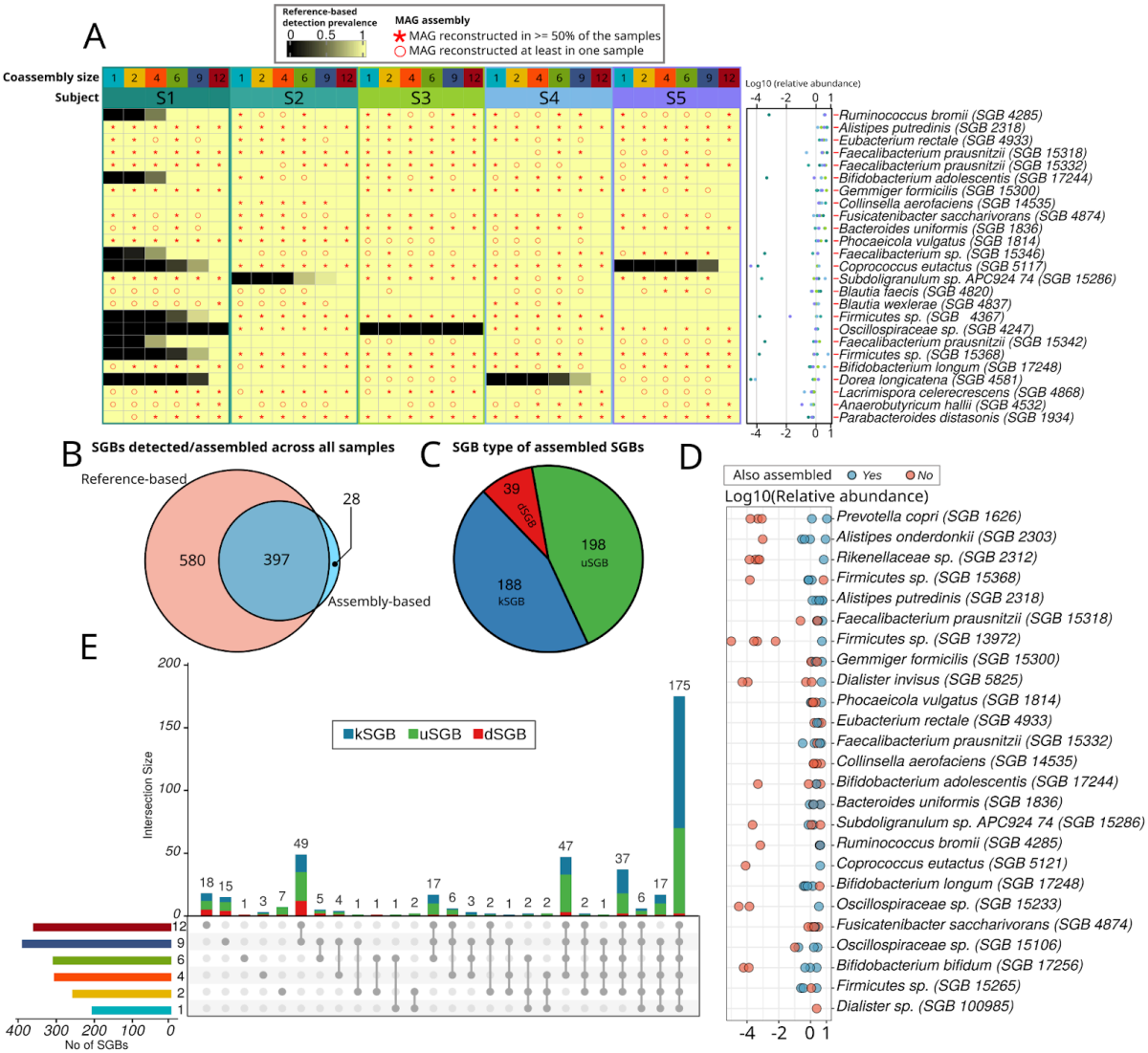

Fig. 2 (A) Heatmap comparing consistency of detection and assembly for the top 25 SGBs (ordered by median relative abundance in the 12-sample profiles, shown in the scatterplot panel on the left) across co-assemblies of different category sizes and subjects. The background color (black-yellow color scale) of each cell indicates the percentage of metagenomic combinations per subject and size category where MetaPhlAn4 detected an SGB. Instead, the symbols within each cell report the prevalence of assembly for an SGB (∗: the SGB was reconstructed in more than 50% of metagenomic combinations for that subject and category size; ○: SGB assembled in at least one sample of that subject and category size; None: the SGB was never assembled in that category). A heatmap with all the SGBs is included in Fig. S9. (B) Venn diagram showing how many of the SGBs detectable by reference-based profiling are detected exclusively by MetaPhlAn4 (pink set), by metagenomic (co-)assembly (cyan set) or by both methods (intersection set). (C) Piechart showing the percentage of the SGB type recovered across all the 376 metagenomic combinations (red: dSGB, cyan: kSGB, green: uSGB). (D) Plot showing relative abundances and HQ/MQ MAG reconstruction for the top 25 SGBs, ordered by decreasing maximum of relative abundance across 12-sample profiles across subjects. (E) Upset plot showing the number of SGBs uniquely detected for each subset of the co-assembly size category, colored by type of SGB (cyan: kSGB, green: uSGB, red: dSGB).

To test the intuitive idea that higher genome coverage leads to improved SGB MAG recovery, we compared coverage distributions between samples where MAGs were successfully recovered and those where recovery failed. The comparison included all SGBs that were detected by MetaPhlAn in at least one co-assembled combination for each of the five subjects (n = 107). As expected, for most of these SGBs (n = 93, log2(F.C.) > 0, P.adj < 0.05), estimates of genome coverage is positively associated with a higher number of recovered MAGs. However, in a few cases (n = 2, log2(F.C.) < 0, P.adj < 0.05), such as for *Fusicatenibacter saccharivorans* (SGB 4874), we observed that at high genomic coverages is less likely to assemble it. A potential explanation for this is that, at higher sequencing depths, non-dominant strains of the same species emerge in their detection. Increased strain variability is known to be problematic for metagenomic assembly, as reads that differ only by a few SNPs associated with strain diversity can introduce ambiguities in the DeBruijn graph structure, leading to fragmented contigs (Meyer et al. 2022) that are more difficult to correctly bin [Table S7].

We measured the overall consistency of SGB retrieval by measuring the prevalence of assembly across co-assembly size categories for each subject (see Methods), considering only those SGBs that were assembled in at least one technical variant combination of a subject. The SGB assembly rate of single sample technical variants is 19.1% ($CI_{95\%}$ [17%, 21.2%]), meaning that in 10 out of 12 technical variants of a subject, it is not possible to assemble the genome of an SGB that is actually present in a subject's community due to low coverage of the single sample. In contrast, the average rate of assembly increases to 87.5% ($CI_{95\%}$ [67.7%, 72.2%]) when co-assembling up to 9 technical variants of the same subject. This shows that increasing the number of co-assembled samples, which in turn increases genome coverage and sequencing depth, leads to a higher consistency in SGB retrieval [Fig. S13].

3.3.7. Co-assembly of longitudinal samples improves recovery of low-abundance SGBs stably present in the community

(Co)assembly of longitudinal datasets has been shown to be beneficial for the recovery of more MAGs. We wanted to test whether the inclusion of longitudinal samples to the full cross-sectional co-assemblies of 12 samples used in the previous paragraphs, could further increase the number of MAGs recovered from the same microbial communities. For each of subjects S2, S4 and S5, we had access to two pairs of 1-year follow-up samples collected two weeks apart (T1, T2). We used T1 and T2 longitudinal sample pairs to extend the full cross-sectional combinations (**T0,** 12-samples combinations) to obtain three longitudinal co-assemblies (**T0+T1**: 14 samples, **T0+T2:** 14 samples, **T0+T1+T2:** 16 samples) for each of the three subjects.

The inclusion of longitudinal sample pairs allowed to increase metagenomic read depths of the three subjects' metagenomes respectively up to 111.9, 134 and 95 Gbp. The majority of SGBs (81%) that were recovered across T0 (co)assemblies were also reconstructed in at least one of the longitudinal combinations of the respective subject [Fig. S14].

Increased metagenomic depth allowed the recovery of MAGs from 53, 46 and 10 SGBs that could not be retrieved from any of the T0 (co)assemblies of the respective subjects. Although few of these SGBs (20/120) were reconstructed because they were newly detected in the longitudinal samples, most of these SGBs (89/120) were already detected by MetaPhlAn in T0 samples at low relative abundances (median: 0.034%, IQR = 0.03), suggesting that the addition of longitudinal samples allowed to gather enough genomic coverage for those species that are stably present after 1 year. 11 of these SGBs, despite being reconstructed, remain undetectable by MetaPhlAn due to insufficient MAG representation in the database for marker gene inference. The remaining 19% of the SGBs were not reconstructed in any of the longitudinal combinations.

Combining longitudinal samples can help increase the genomic coverage of resident species that are present at low abundances but are stable throughout time in gut microbial community of a subject.

3.3.8. Co-binning improves MAG quality by exploiting the covariance of the contig depth of technical variants

Technical variants (or samples that are biologically related) could also improve metagenomic binning besides metagenomic assembly. We next assessed whether metagenomic co-binning (i.e: using inter-sample covariation of contigs coverage) of co-assemblies can increase the number and the quality of recovered MAGs. Co-binning, paired with co-assembly, has been successfully applied in other studies on longitudinal time series and co-environmental samples (Pasolli et al. 2019; Ye 2017) to recover bins of improved quality. We performed co-binning with MetaBAT2 (D. D. Kang et al. 2019), which allows using multiple samples to compute covariation of coverage using multiple mapping files (Seemann 2014).

We compared binning to co-binning on the full cross-sectional co-assemblies of each subject (**T0**: co-assembly of cross-sectional 12 technical variants), and additionally on longitudinal co-assemblies for subjects S2, S4, and S5. For these subjects, we had access to two pairs of follow-up samples after a year that were collected two weeks apart (T1, T2). We used T1 and T2 longitudinal samples to extend the full cross-sectional combinations to obtain longitudinal co-assemblies (**T0+T1**, **T0+T2**, **T0+T1+T2**).

By binning and co-binning the contigs generated from the co-assembly of cross-sectional and longitudinal combinations, we reconstructed 434 (746 MAGs) and 1,595 (2,667 MAGs) SGBs respectively. Among the 434 SGBs obtained from cross-sectional co-assemblies, the 71.9% of the SGBs (N = 312/434) were assembled by both binning techniques, but MAGs had higher overall quality when retrieved with co-binning (Wilcoxon signed-rank test, P < 1.5e-09, eff.size = 0.37), mostly due to increase in completeness (co-binning: 90.3% ± 1.14; binning: 84.6% ± 1.48, Wilcoxon signed-rank test: P < 2.2e-16) rather than to a decrease in contamination, which was slightly higher in co-binning (co-binning: 1.28% ± 0.142; binning: 0.897% ± 0.115, Wilcoxon signed-rank test: 2.457e-11) [Fig. 3A; Fig. S15]. Indeed, completeness and contamination across MAGs from cross-sectional co-assemblies were weakly but positively correlated (Spearman's correlation, rho = 0.046, P = 8.19e-10). Out of the 312 shared SGBs, 62 increased their quality from MQ to HQ when retrieved with co-binning, 2 decreased to MQ, and the remaining 248 SGBs did not change the quality category. The 28.1% (N = 122/434) of SGBs were recovered exclusively by one technique, with co-binning recovering more than binning (78 vs 44) [Fig. 3B; Fig. S18].

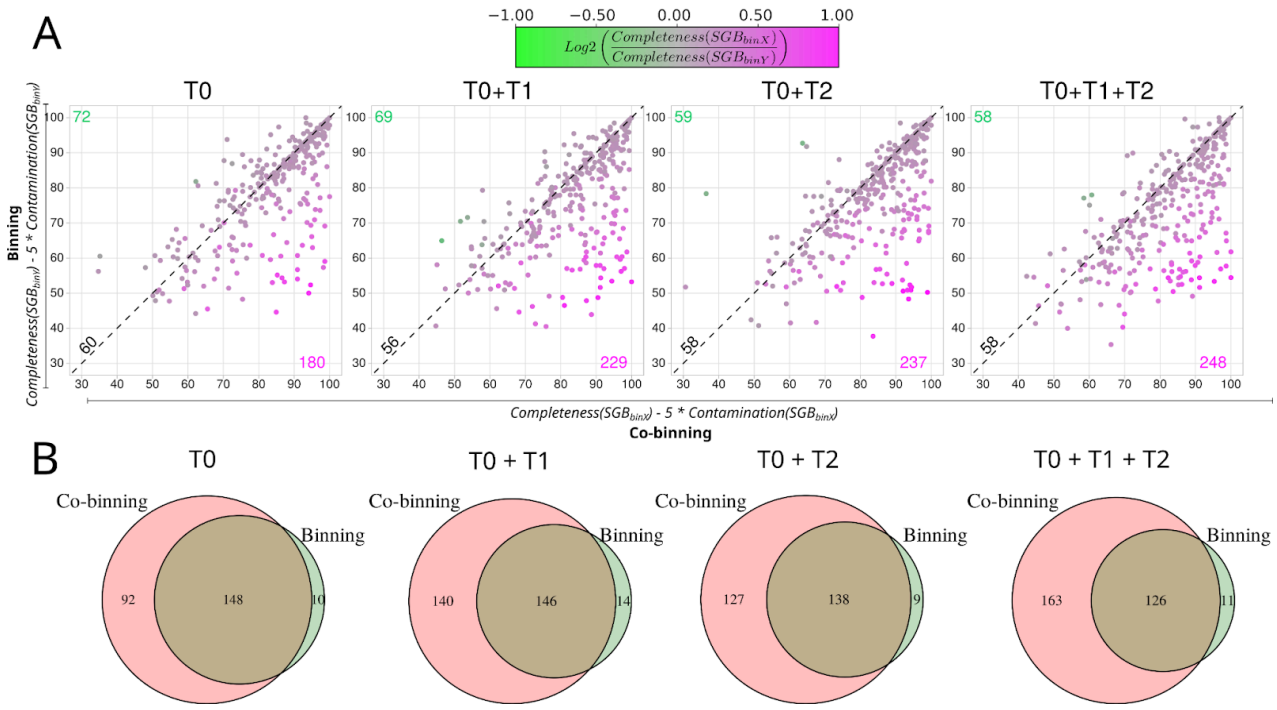

Fig. 3 (A) Scatterplots comparing the MAG quality scores of SGBs that were binned both by binning and co-binning in the cross-sectional (T0) and longitudinal co-assemblies (T0+T1, T0+T2, T0+T1+T2) of the three subjects (S2, S4, and S5) for which longitudinal samples were available. The two axes report a quality score combining completeness and contamination for MAGs obtained with co-binning (x-axis) or binning (y-axis) (see Methods, Comparison of binning and co-binning). The color scale for individual points represents the log2 ratio of the completeness of same-SGB MAGs retrieved by both binning (green: more complete in binning) and co-binning (purple: more complete in co-binning). The numbers at the edges of the plots indicate how many SGBs have better quality in co-binning (purple), binning (green) or share the same overall quality (gray). (B) Venn diagrams showing the number of SGBs with HQ MAGs retrieved by binning, co-binning, or both in co-assemblies incorporating longitudinal samples. The count shown within co-binning and binning sets includes the number of MAGs that shifted from MQ to HQ category, along with the MAGs that are uniquely binned by the respective technique as HQ MAGs. The intersection includes how many MAGs were binned as HQ by both techniques.

On longitudinal co-assemblies, improved binning performances of co-binning are confirmed, possibly with larger effect sizes. Co-binning improved the overall quality for the 66.7% (N = 714/1072) of SGBs that were retrieved also by binning (Wilcoxon signed-rank test, P < 2.22e-16 for all comparisons) [Fig. 3A; Fig. S16]. The 16% of shared SGBs (N= 172/1072) contamination and completeness were not affected by the binning method, and only a minority of the SGBs (N = 186/1072, 17.4%) had higher quality when obtained with binning. 73.54% of MAGs obtained with both methods had higher completeness with co-binning (average increase in completeness = 10.52%). Also in this case, the increase in MAG quality is driven by an increase in completeness (7.69% median difference, P < 10^{-16}, Wilcoxon signed-rank test), with contamination

increasing only slightly (0.285% median difference, P < 10-6, Wilcoxon signed-rank test).

Inclusion of longitudinal samples to the co-assembly lowers the median of the overall quality score distribution of the binned MAGs retrieved by simple binning (T0 vs T0+T1+T2, Wilcoxon signed-rank test, median difference = 5.48%, P = 0.00053), mainly due to SGBs being newly assembled, thanks to the increase in sequencing depth, and exceeding the minimum completeness and contamination thresholds for being considered. Interestingly, co-binning appears to be particularly useful when co-assembling longitudinal samples, as it rescues the completeness for most of the SGBs retrieved from longitudinal co-assemblies [Fig. S17]. Additionally, the number of MAGs exclusively recovered by one of the two techniques in the T0+T1+T2 co-assembly saw an increase of 84.6% for co-binning, whereas traditional binning saw only an 18.2% increase [Fig. S18].

In conclusion, the co-assembly of longitudinal samples has proven to be an effective strategy to assemble a higher fraction of the metagenomic diversity of the gut community of a subject by increasing the genome coverage of species within the gut community. Although the addition of longitudinal samples may decrease the completeness of binned MAGs, this can be mitigated through the application of co-binning. Still, a small fraction of MAGs are exclusively or better recovered through traditional binning. Therefore, to maximize the recovery of MAGs a combined approach utilizing both binning strategies is recommended.

3.3.9. Co-assembly of samples from multiple subjects produces chimeric assemblies

A critical factor to keep into account when doing metagenomic assembly is the potential presence of multiple strains sharing a high fraction of homologous but not identical genomic sequences, which might lead to the generation of chimeric contigs, i.e. contigs that span genetic sequences originating from different biological sources (L. Chen et al. 2020). However, co-assembling samples from different sources or subjects could help increase MAG retrieval, especially when the sequencing depth of the single samples is low.

In this section, we investigated whether combining stool metagenomes from multiple subjects during assembly produces MAGs that differ genomically from those obtained when assembling each subject's samples individually. To observe this, we selected technical variants from the five subjects of the cross-sectional cohort and created metagenomic combinations with varying proportions of samples from different subjects.

In total, we (co)assembled 23 combinations of same-subject metagenomes (SSM) and 42 combinations of mixed-subject metagenomes (MSM). Due to the large number of possible sample combinations that can be made from 61 technical variants, and the available computational resources at the time, for most MSM (n = 40) combinations we chose to combine samples from S1 and S2 and limit to a maximum of 4 combined samples. However, for each subject, we still co-assembled at least one SSM, and additionally, two combinations include samples from all five subjects. We refer to different combinations of different sample sizes and subject compositionality with the notation N_{S1}:N_{S2}:N_{Ot}, where N_{S1} and N_{S2} refer respectively to the number of technical variants included from subjects S1 and S2, and N_{ot} the number of samples from the other three subjects. [Table S8].

Examination of the phylogenetic trees of those SGBs that contain MAGs from all five subjects, it is possible to appreciate that MAGs from single samples and SSM of the same subject are phylogenetically the closest and are co-located on the same clade [Fig. 4A]. As for MAGs recovered from MSM combinations, made from subjects S1 and S2, it is possible to appreciate how the phylogenetic placement between the subject-specific clades reflects the subject composition of the combinations from which the MAGs were obtained. MAGs obtained from MSM with an unbalanced subject-composition (3_{S1}:1_{S2}:0_{Ot} or 1_{S1}:3_{S2}:0_{Ot}) will be phylogenetically closer to the clade of the dominant subject. In contrast, MAGs from balanced MSM combinations (2_{S1}:2_{S2}:0_{Ot} or 1_{S1}:1_{S2}:0_{Ot}) occupied intermediate positions between the two subject-specific clades. Similarly, MAGs reconstructed from combinations including samples from all five subjects (1_{S1}:1_{S2}:3_{Ot} or 3_{S1}:3_{S2}:9_{Ot}) did not associate with any subject-specific clade but rather showed intermediate phylogenetic placement. These patterns suggest that the phylogenetic signal incorporated in MSM-derived MAGs is proportional to the contribution of subject-specific samples in the assembled combination.

To quantify and summarize these phylogenetic relationships between single sample MAG and SSM and MSM-derived MAGs observed from the phylogenetic trees, for all SGBs, we evaluated the distributions of mutation rates for the pairwise categories of interest. For each SGB, pairwise mutation rates were classified into 10 categories based on two factors: the number of samples from the original subject present in the combination, and the total number of samples in the combination [Fig. 4B]. Within each SGB phylogeny, if multiple MAGs were reconstructed from individual samples of the same subject, the MAG with the highest quality metrics was selected as the representative genome for the strain of the corresponding subject. We excluded from the analysis SGB phylogenies that did not include MAGs from single samples, ending with 147 SGB phylogenies.

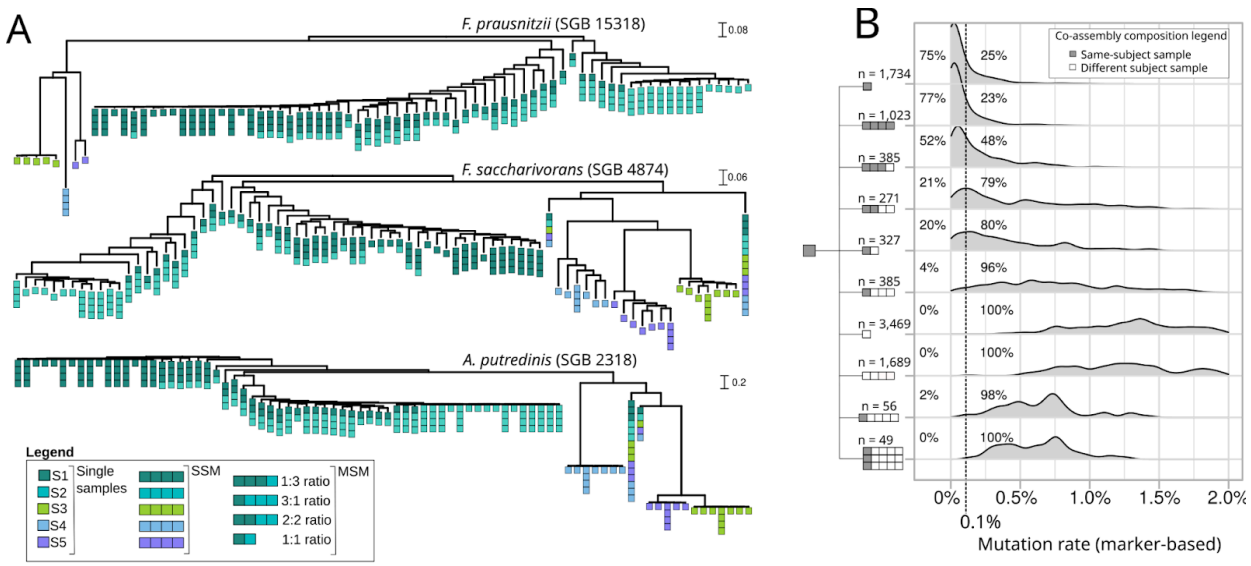

Fig 4: (A) Phylogenetic trees of SGBs with MAGs reconstructed across co-assemblies with different compositions of samples from S1, S2, and the other individuals. Tree tips show the sample composition of the co-assembly from which the respective MAG was obtained (S1 sample: cyan, S2 sample: dark green, Other subject: green, light blue and purple); (B) Ridge-plot showing the density distribution of the pairwise mutation rates between MAGs from single samples, MSM and SSMs (see Methods). From top to bottom, comparisons ordered by source similarity of the MTM to the reference subject SSMs (top: same subject SSM, bottom: MSM with a majority of single-sample from a subject different than that of SSM). The color of singular tiles that compose SSM and MSM combinations indicates the relation of the single sample with the subject of the MAG considered as reference in the pairwise mutation rate being considered (grey: the sample is from the same subject, white: the sample is from a different subject as the reference. Numbers (n) above each tile combination indicates the number of pairwise mutation rates considered to populate the respective distribution. The dashed line highlights the average intra-subject mutation rate between MAGs from single-samples of the same subject. Percentages indicate how many pairwise mutation rates of the respective comparison distribution fall below or above the intra-subject average mutation rate between single sample MAGs.

Mutation rates for those categories that involve comparisons between MAGs obtained from combinations that include samples from a single subject (1st and 2nd rows of Fig. 4B), which include comparisons between single-sample MAGs of the same subject (1st row, median mutation rate: 0.03%, IQR: 0.103%, min-max: 0-21.67%), or between representative genomes and SSM of the same subject (2nd row, median mutation rate: 0.03%, IQR: 0.103%) they are the lowest, reflecting the phylogenetic closeness that was observed in the respective SGB trees. As expected, pairs of MAGs from different subjects, whether they were recovered from single samples or SSM combinations (7th and 8th rows), were the farthest (median: 1.38%, IQR: 0.698%, min-max: 0.13%-4.47%).

Mutation rate distributions of comparison categories in which the composition of MSM is progressively enriched for samples of a different subject (4th, 5th, 6th, and 7th rows) progressively get closer to being similar to the distribution of mutation rates between MAGs of different subjects (mutation rate medians shifting from 0.103% to 0.711%).

Accordingly, MAGs from MSM that include few samples from all of the five subjects have mutation rates higher than those obtained from SSM but lower than those from SSMs of references of a completely different subject (9th row: median = 0.659%, IQR = 0.310%, 10th row: median = 0.514, IQR = 0.366).

Although the MAGs obtained from MSM were categorized as HQ/MQ, they contained strain-chimeric contigs that incorporated genomic differences from multiple strains present in the co-assembled samples from different subjects, reflecting the proportional contribution of each subject's samples in the (co)assembly. This suggests that (co)assembly should be limited to samples from the same subject, as combinations of samples from unrelated subjects generate strain-chimeric MAGs that do not represent actual strain populations.

We then explored the impact of longitudinal co-assemblies of the same subject on MAGs that should represent the same strain by assembling mixed-timepoint metagenomes (MTM, n = 12) and single-time point metagenomes (STM, n = 9). MAGs of the same SGB, assembled from STM at different time points have mutation rates that are comparable to those observed between single-sample SSMs (T0 vs T1: median = 0.069%, IQR = 0.417%, n = 47; T0 vs T2: median = 0.083%, IQR = 0.426%, n = 36), and mutation rates between STM and MSM with T0 and T1 metagenomes did not differ significantly (T0 vs T1+T2: median = 0.075%, IQR = 0.479%, n = 38). Instead, MSM including all 3 STM for each subject produced MAGs that have mutation rates lower than those observed across MAGs from cross-sectional technical variants of the same subject (T0 vs T0+T1+T2: median = 0.022%, IQR = 0.101%, n = 218) [Fig. S19]. This highlights the person-specificity and temporal stability of the gut microbiomes (Lianmin Chen et al. 2021). While in the current work we access samples that were taken more than one year apart, these results suggest that longitudinal samples from the same subject can be combined to increase the read depth of the metagenome to maximize MAG reconstruction while avoiding chimeric MAGs. Overall, these results underscore the importance of carefully selecting the samples to co-assemble, as it is a potentially powerful approach that can also lead to misleading results when applied to samples from different subjects.

3.4. Conclusions

Here we assessed the performance of metagenomic co-assembly in improving the recovery of MAGs. With a dedicated study design including both technical variants (microbiome samples from the same fecal community that differ only in the pre-sequencing protocols) and longitudinal samples, we first confirmed sequencing depth as a critical factor to comprehensively describe the microbial diversity of a gut microbial community. Given that databases describing microbial diversity of the gut microbiome are comprehensive enough, with more than 1,000,0000 reference genomes, reference-based profiling has become a sensitive and reliable way to profile a metagenomic community. Hence, sequencing depths in a range of 2 to 5 Gbp commonly employed in metagenomic studies are sufficient to describe the main composition and dominant microbial members of a stool sample, obtaining alpha- and beta-diversity estimates that are only slightly underestimated compared to those obtained at higher sequencing depths of metagenomic combinations. Most of the differences in the observed communities of the same subject single technical variants can be addressed mostly to the slight differences in the chosen sequencing protocol of the technical variants. However, detection of low-abundance species is greatly impacted by sequencing depth, as we observe that 33% of species whose relative abundance sums up to 1% can be detected only at higher sequencing depths. Additionally, the possibility of detecting these low-abundance species relies on the fact that a number of MAGs, already assembled from other studies, is sufficient to infer clade-specific markers to reliably detect them.

Reference-based profiling detects more than five times the number of SGBs assembled from the same sample, showing how much more requiring is the assembly of a species genome in terms of sequencing depth and genome coverage, which we estimated the minimum to be 11X on average, a threshold that is hard to achieve for the majority of SGBs in a gut community. Given the scarcity of reads originating from low abundance species, and the technical biases arising from using specific pre-sequencing protocols, we show that co-assembly of technical variants is a valuable post-sequencing approach to capture a larger fraction of biodiversity of a microbial community and maximize retrieval of MAGs of under characterized (uSGBs) or putatively new species (dSGBs). In this study, the genomes of 39 potentially novel species have been assembled, including 11 SGBs classified at their highest known taxonomic level within the phylum Firmicutes. Although producing technical variants is expensive and time-consuming, it might be a convenient approach to maximize the recovery of MAGs where the process of sample collection is difficult. We show that co-assembly can be further improved when performing metagenomic co-binning (also known as multi-sample binning) which outperforms single-sample binning, increasing the number and the quality of MAGs, especially from co-assemblies that include the same subject longitudinal samples.

However, we suggest the application of both co-binning and binning strategies for optimal results, as in a few cases some MAGs are recovered exclusively or with improved quality with single sample binning. Co-assembling combinations of metagenomes from mixed sources (i.e. different subjects) produced MAGs that are genomic entities halfway to those found in the subjects and do not thus constitute real strains. In contrast, and thanks to the longitudinal stability of the microbiome, co-assembly of longitudinal samples from the same subject improved MAG retrieval while limiting the risk of producing chimeric MAGs (Linxing Chen et al. 2020; Hofmeyr et al. 2020). Overall, our study shows that, even in well-studied environments, sequencing depth is key to both reference- and assembly-based approaches. Co-assembly and co-binning in metagenomics are promising tools to elucidate the microbial diversity and can be applied to publicly available metagenomic datasets with lower sequencing depths, but are only advised for application within an individual metagenome.

3.5. Material and Methods

3.5.1. Metagenomic cohorts

A total of 73 stool samples were obtained from 5 healthy donors from Scotland, aged between 18 and 65 years, recruited in the context of a metagenomic benchmarking study, were included. The study was approved by the Rowett Institute Ethical Review Committee under study number 5946 (“Metabolic roles of anaerobic bacteria in the human large intestine”). For each subject, 12 samples were collected from the very same single stool specimen (except for one subject, for which 13 samples were available) (*cross-sectional cohort,* N=61 samples). We refer to cross-sectional samples as T0. A second cohort (*longitudinal cohort,* N=49 samples) included 12 samples from a follow-up of three of the five participants (subjects S2, S4, and S5), together with the respective 37 samples from the cross-sectional cohort. The follow-up samples were collected during two visits after a one-year follow-up, one week apart, which we refer to as T1 and T2. At each time point, a stool sample was subsampled to obtain 2 technical variants (N=12). Available sample metadata included information on the stool sample collection procedure, storage protocol, the protocol for metagenomic DNA extraction, and statistics on the NGS libraries after quality control [Table S1].

3.5.2. Sample collection, storage, and DNA extraction

Metagenomic samples from the cross-sectional cohort were obtained by re-sampling each of the volunteer's stool specimens to produce up to 13 *technical variants*. We use the term *technical variant* to refer to metagenomic samples from the same biological source (i.e. the microbial community of a single stool specimen) that differ in the technical protocols used for sample processing before sequencing (sample collection, stool sample storage, and DNA extraction. Each technical variant from the cross-sectional subcohort was pre-processed with a unique combination of sample storage and DNA extraction protocols [Table S1]. Of the 12 samples per subject, 9 were analyzed shortly after DNA extraction, whereas 3 were stored at -80 °C for one year prior to sequencing. Four different protocols were used for DNA extraction: ProtocolQ (IHMS_SOP 06 V2, as from www.microbiome-standards.org), ZymoBIOMICS ™ DNA Mini Kit (ZymoResearch, CA, United States), DNeasy® PowerSoil® Pro Kit (QIAGEN, Germany) and FastDNA™ SPIN Kit (MP Biomedicals, France).

Next, 14 samples were sequenced right after sample collection and DNA extraction, 13 samples were left at room temperature for 2 days without additional buffers, and 19 samples were conserved at room temperature for 4 days with the addition of Zymo RNA/DNA shield buffer. Samples sequenced one year later were stored at -70° (5 samples stored without buffer and 10 samples stored with Zymo RNA/DNA shield buffer (ZymoResearch, CA, United States).

The 12 samples from the longitudinal cohort were stored at room temperature with Zymo RNA/DNA shield buffer, and the genetic content was extracted using the DNeasy® PowerSoil® Pro Kit. For each volunteer's visit, one replicate was collected using a scoop and one with a swab.

3.5.3. DNA Sequencing and read pre-processing

Sequencing libraries for samples from the cross-sectional cohort that were sequenced shortly after DNA extraction were obtained with the NexteraXT DNA Library Preparation Kit (Illumina, US) and were sequenced with Illumina HiSeq 2500 (2x101, paired-end reads). Sequencing libraries for the cross-sectional samples were sequenced one year later after DNA isolation and all samples from the longitudinal cohort were prepared with NEXTFLEX® Rapid DNA-Seq Library (Illumina, US) and sequenced with Illumina NovaSeq (2x151 bp, paired-end reads).

Raw reads of all metagenomic libraries were uniformly pre-processed to keep only high-quality reads. Quality control was conducted using Trim Galore to remove short

and low-quality reads (with parameters: "--stringency 5 --length 75 --quality 20 --max_n 2 --trim-n", https://github.com/FelixKrueger/TrimGalore). Illumina PhiX adapters were removed from reads and human DNA was discarded by using Bowtie2 (Langmead and Salzberg 2012) to map the reads against the reference PhiX 174 genome (NCBI accession ID 10847), and the hg19 Human Genome, respectively. Custom Python scripts were used to further process and sort the raw reads to produce three .fastq files for each sample (two for paired-end reads, R1 and R2, and one for unpaired reads) (https://github.com/SegataLab/preprocessing). In total, the 61 samples from the cross-sectional cohort were successfully sequenced and produced a total of 3061 x 10^6 reads (median: 47.02 x 10^6, IQR = 25.55 x 10^6), and the 12 samples from the longitudinal cohort produced 914 x 10^6 reads (median: 69.9 x 10^6, IQR = 4.33 x 10^6).

The average sequencing depth of samples in publicly available datasets was obtained from curatedMetagenomicData3 (version 3.5.3, (Pasolli et al. 2017) by filtering the metadata table by keeping samples with attribute "body_site" with value "stool" and computing statistics on the 'number_bases' and 'number_reads' variables (n = 21,030).

3.5.4. Metagenomic assembly and co-assembly

Metagenomic assembly was performed using MEGAHIT (D. Li et al. 2015) with default parameters (k-mer sizes 21, 29, 39, 59, 79, 99, 119). Contigs shorter than 1,000 nt were removed from assemblies. Co-assembly was performed by assembling the combined reads from multiple metagenomic samples. We produced metagenomic combinations of different size categories (*k*-technical variants combinations, $k = 2, 4, 6, 9, 12$) by pooling together *k* random technical variants from those available for the same subject. For each subject we assembled all the available single samples (*1-sample* combination), and co-assembled at least 15 *k*-technical variants combinations for sizes $2 \leq k < 12$, plus the combination including the full set of technical variants that were available for each subject (12-technical variants combination). While the full combination of subject S4 comprised 13 technical variants instead of 12, as the sequencing depth of all its samples combined was comparable to that of the other subjects' full co-assemblies we still referred to it as a 12-technical variants combination for simplicity. Indeed, subject S5 had a higher sequencing depth than subject S4 (74.65 Gbp vs 74.48 Gbp). Paired-end and unpaired read libraries of *k*-sample combinations were given as input for MEGAHIT. The same approach was used to co-assemble the 3 longitudinal combinations (T0+T1, T0+T2, T0+T1+T2) for subjects S2, S4, and S5 (for a total of 9 co-assemblies of sample sizes 14 and 16 samples), and the 21 inter-subject co-assemblies (see Chimericity of MAGs obtained across multi-subject and single-subject metagenomes, Methods)

3.5.5. Metagenomic binning and co-binning of co-assembled contigs

We use the term "co-binning" to refer to technical variants from the same subject being mapped subsequently to the same set of assembled contigs, which differs from "binning" as technical variants are mapped at once to the assembled contigs. Short-read mapping was performed using the aligner Bowtie2 (parameters: --no-unal *--a --sensitive-local*, version 2.3.5.1, (Langmead and Salzberg 2012). Binning was performed by mapping each co-assembled combination as a single mapping file (.bam). To perform co-binning, we mapped single technical variant reads against the same set of co-assembled contigs to obtain as many BAM files as the number of co-assembled technical variants. We sorted read alignments in BAM files by left-most coordinates using *Samtools* version 1.7 (H. Li et al. 2009). We used the *jgi_summarize_bam_contig_depth* module from MetaBAT2 (which accepts a comma-separated list of BAM files) to produce a tabular file (.depth file) containing the average genomic coverage depth and variance for the co-assembled contigs in the .fasta files. The .depth files were then used to perform the actual metagenomic binning with MetaBAT2 (version 2.12.1, (D. D. Kang et al. 2019), using default parameters.

3.5.6. Assessing completeness and contamination of MAGs

Measures of completeness and contamination of the MAGs obtained from all analyses were computed using CheckM (Parks et al. 2015), using the lineage-based workflow *lineage_wf* with default parameters. We then defined MAGs with contamination ≤ 5% and completeness ≥ 90% as high quality (HQ), whereas those with contamination ≤ 5% and contamination 50% ≤ x < 90% were defined as medium quality (MQ). MAGs with contamination > 5% and/or completeness < 50% were discarded from further analyses.

3.5.7. Taxonomic annotation of MAGs

Taxonomy of MQ and HQ MAGs assembled in this study was predicted using a subroutine script from PhyloPhlAn 3.0 version 3.0.39 (Asnicar et al. 2020), https://github.com/biobakery/phylophlan), *phylophlan_metagenomic.py*, which computes the average MASH distance of a query MAG against all genomes in the MetaRefSGB database (release: full SGB.Oct22 database) to find the putative SGB to which it belongs. Phylophlan_metagenomic was run using `--add_ggb --add_fgb` flags to report taxonomical membership of MAGs to their closest Genus-level (GGB) and Family-level genome (FGB) bins. MAGs were assigned to an SGB if its average MASH

distance to the SGB genomes was below 0.05. SGBs are categorized into known (kSGB) or unknown (uSGB), according to whether the SGB contains a reference genome from NCBI or not. MAGs whose average MASH distance from any SGB in MetaRefSGB was greater than 0.05 were reclustered to define new SGBs, which should represent putatively new species. We refer to putative new SGBs as dSGBs. In this study, all MAGs were assembled before 2020 and added to MetaRefSGB, which were classified as uSGBs and used to generate successive versions of MetaPhlAn4. Consequently, we retroactively labeled as dSGBs those uSGBs that were newly discovered at the time and were being defined for the first time in our cohort.

3.5.8. Co-assembly of technical variant pairs

First, we selected the 18 technical variants with the highest sequencing depth (> 58,000,000 reads) from the cross-sectional samples that were sequenced with Illumina HiSeq2500 (see DNA Sequencing and read pre-processing, Methods). Fastq libraries were subsampled down 27,370,864 reads using *seqtk* software (version 1.3-r115, https://github.com/lh3/seqtk/) and assembled as single samples (assembly of A and B). Subsampled technical variants were further subsampled to half of their depth, and were merged to compose paired metagenomes (matched by subject) of 27,370,864 reads and then co-assembled (co-assembly of AB). Assembly and co-assembly were performed with the same pipeline used throughout this work.

3.5.9. Reference-based taxonomic profiling

We generated taxonomic relative abundance profiles of single technical variants with MetaPhlAn4 (version 4.1.0, (Blanco-Míguez et al. 2023), using database version mpa_vOct22_CHOCOPhlAnSGB_202212 [Table S9] and default parameters. Relative abundance profiles were then merged into a single table using MetaPhlAn4 utility script *merge_metaphlan_table.py*, which we then filtered to keep relative abundances of SGB-level entries only. We estimated the coverage of SGBs across all the 376 technical variant combinations by running MetaPhlAn4 on the previously generated bowtie2 outputs with options `--input_type bowtie2 -t clade_profiles --stat avg_g` to obtain clade profiles. Each clade profile contains the list of SGB-specific markers, along with their coverage (RPK). For each SGB and its respective markers, the robust average of RPK values was computed and used to estimate the SGB coverage with the formula $Marker\text{–}based\ coverage = \frac{Robust\ Avg.\ RPK_{SGB} * Avg.\ read\ length)}{1000}$

3.5.10. Comparison of reference- and assembly-based profiling

We assessed the overlap of the species detectable by the reference- and assembly-based approaches by checking which SGBs were detectable by the MetaPhlAn4 database mpa_vOct22_CHOCOPhlAnSGB_202212 and those that were assembled into MQ or HQ MAGs. The prevalence of SGB detection and assembly of an SGB was computed for each subject and size category by checking how many times an SGB was detected across the combinations available for that category. Heatmaps were then plotted with the R package *ComplexHeatmap* version 2.12.0 (Gu, Eils, and Schlesner 2016).

3.5.11. Alpha- and beta-diversity estimation of metagenomes

We estimated Shannon's alpha-diversity and SGB richness (observed) of all technical variants combinations with *vegan* R package version 2.6-4 (Dixon 2003). SGB richness for assembly-based profile was computed by summing presence-absence values (assembled/not assembled) of SGBs and further categorized the count into kSGB, uSGB, or dSGB. Beta-diversity between metagenome combinations was estimated with Aitchison distance (Aitchison 1982), by computing the matrix of pairwise Euclidean distances with *vegan* package version 2.6-4 (Dixon 2003) after CLR-transforming the data using the *compositions* package (van den Boogaart and Tolosana-Delgado 2013) and imputing zeroes by using the minimum proportional abundance detected for each SGB. Principal Coordinate Analysis (PCoA) was then applied using the *ape* package version 5.7-1 (Paradis, Claude, and Strimmer 2004) to visualize the distance between samples.

3.5.12. Statistical methods

All signed-rank and rank-sum Wilcoxon tests, along with their corresponding effect sizes, were computed using the 'wilcox_test' and 'wilcox_effsize' functions from the R package rstatix version 0.7.2 (https://rpkgs.datanovia.com/rstatix/), with two-sided tests conducted unless specified. PERMANOVA tests were conducted with *adonis2* function from the R package *vegan* version 2.6-4. All other statistical methods were applied using functions from the base R package *stats* (version 4.3.1).

3.5.13. Association of SGB genome coverage and chances of assembly

From the MetaPhlAn relative abundance matrix (Table S9) and the presence-absence of SGBs assembled across the metagenomic combinations (Table S2), we selected SGBs (n = 107) that met the following criteria: (1) showed non-zero relative abundance in at least one metagenomic combination per subject, and (2) this condition was satisfied across all five subjects. Additionally, these SGBs must have been successfully assembled as medium- or high-quality MAGs in at least one metagenomic combination. Then, for each selected SGB, we calculated the Log2 fold change between two conditions: the median marker-based coverage when the SGB was detected but failed assembly, and the median marker-based coverage when the SGB was both detected and successfully assembled. Only non-zero coverage values were included in calculations of medians.

3.5.14. Comparison of binning and co-binning

To assess whether co-binning improved the quality of MAGs when compared to binning, we selected MAGs obtained from cross-sectional and longitudinal co-assemblies, restricting our analysis to co-assemblies that include all the technical variants available at each time point. For each co-assembly combination, we paired same-SGB MAGs binned using both single- and co-binning, using the subject of origin and the co-assembly ID as matching criteria. To compare the quality of MAGs obtained by both methodologies, we summarized the measures of completeness and contamination obtained with CheckM (Parks et al. 2015) by defining an overall quality score:

$$Quality\ score_{MAG_X} = Completeness_{MAG_X} - 5 * (Contamination_{MAG_X})$$

We compared the effect size and the statistical significance of the overall shift in quality between the two methods with Wilcoxon signed-rank tests.

3.5.15. Assessment of MAG chimericity

To assess whether MAGs obtained from co-assemblies of technical variants of different subjects were chimeric we merged technical variants into combined single-sample metagenomes (SSM), which contained only samples from a single individual, and mixed-subject metagenomes (MSM), containing samples from multiple individuals in different proportions. Due to computational requirements, we chose to compose most MSM combinations using only technical variants from subjects S1 and S2, specifically the six ones which yielded most MQ and HQ MAGs. Overall, we composed 23 SSMs of

4 samples (10x S_1, 10x S_2, 1x S_3, 1xS_4 and 1xS_5) and 52 MSMs with the following compositions: 3_{S1}:1_{S2} x 10, 2_{S1}:2_{S2} x 10, 1_{S1}:1_{S2} x 10, 1_{S1}:3_{S2} x 10, 1_{S1}:1_{S2} :3_{Sot} x 1 and 3_{S1}:3_{S2} :9_{Sot} x 1.

To assess whether integrating samples of the same subject at different time points would alter MAGs of the same SGB, we (co)assembled, for each one of subjects S2, S4, and S5, the following samples/combinations: 3 single T0 samples (PSZY2, PSZY4 and PSZY5), 6 single T1 and T2 samples, 6 T0+T1 and T0+T2 combinations, 3 T1+T2 combinations and 3 T0+T1+T2 combinations [Table S7]. Next, we reconstructed MAGs from (co)assembled contigs with the same pipeline as above. MQ and HQ MAGs were selected from each combination, and were assigned to their respective SGB taxonomy with phylophlan_metagenomic.py (see Taxonomic annotation of MAGs, methods). We then reconstructed the phylogeny of each SGB by applying PhyloPhlAn3.0 (Asnicar et al. 2020) to perform a multiple sequence alignment (MSA) of the nucleotides of the SGB-specific core genes and then build a phylogenetic tree with RAxML (parameters: -m GTRCAT -p 12, version: 8.1.15, (Kozlov et al. 2019). For each SGB, we extracted pairwise mutation rates for the comparison categories reported in the density distributions of Fig. 4. Density of mutation rates distributions was estimated and plotted with R package *ggridges* version 0.5.4 (Wilke 2024), using function *geom_density_ridge* (default settings, apart from *bandwidth* = 0.05). In all comparisons involving MAGs, we paired a reference MAG from a single sample SSM with a MAG from either a multi-sample SSM or an MSM. Since we had access to five single-sample SSMs for each of the two subjects, this meant that a single MSM MAG could potentially be compared to five different reference MAGs. To determine which MAG would serve as the reference for a given MSM MAG, we selected the one with the highest quality score (see Comparison of binning and co-binning, Methods). We kept pairwise mutation rates of MAGs assembled from an MSM and a single sample only if the MAG was chosen as a reference. Pairwise mutation rates involving non-reference MAGs were included only in the computation of the intra-subject median mutation rate (dashed line in Fig. 4B) and in the mutation rate distributions of comparisons between single samples (1st and) from different subjects (7th row of Fig. 4B).

3.6. Supplementary Tables

https://docs.google.com/spreadsheets/d/1SokTgGl6TAPVPnFmtOVuVJEKizQy3QQAYz-xt9pohZg/edit?usp=sharing

3.6.1. Supplementary Figures

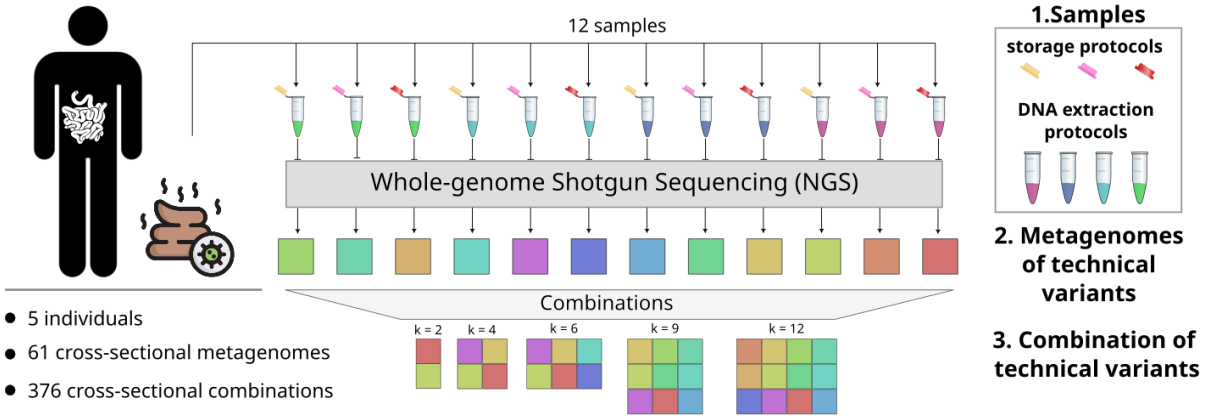

Fig. S1. Study design: stool samples from 5 individuals were collected and processed with slightly different pre-processing protocols (see Methods) and sequenced into biological redundant metagenomes, or *technical variants*. For each subject, technical variants were used to compose 376 intra-subject combinations of different sizes (Table S3), which were both co-assembled to reconstruct MAGs and profiled with a reference-based approach (MetaPhlAn4). Additionally, for a few subjects (S2, S4, and S5), four longitudinal samples were available and included in larger size combinations (not included in this figure, see Fig. 3). Cross-sectional technical variants were also used to compose 21 inter-subject combinations to assess how co-assembly of inter-subject metagenomes impacted the assembly of SGBs.

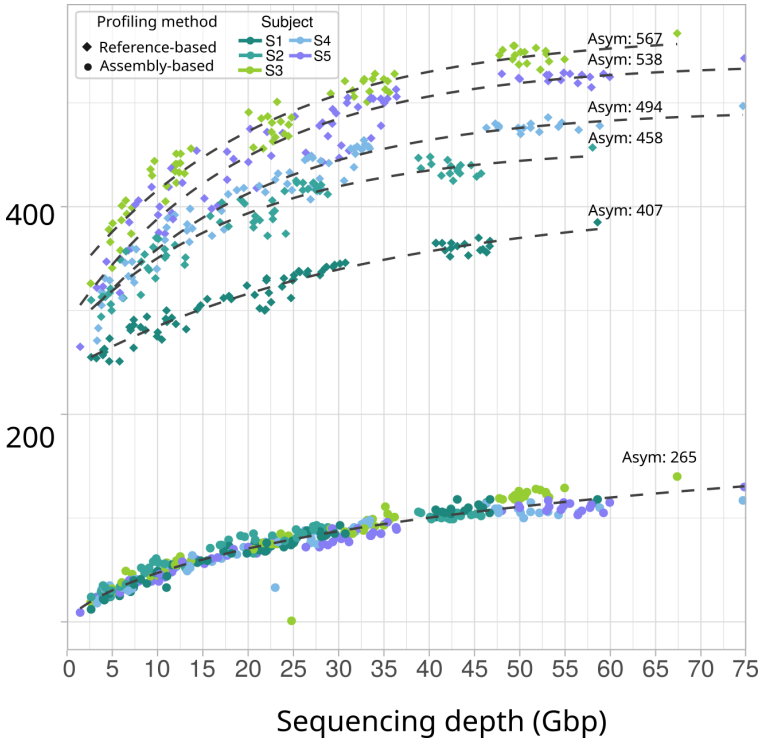

Fig. S2 Scatterplot showing the number of SGBs detected with reference-based (diamond shape) and assembly-based approach (circle shape). Black dashed lines show fitted species accumulation curves for estimating the theoretical number of species in the community (Statistics at Table S5H)

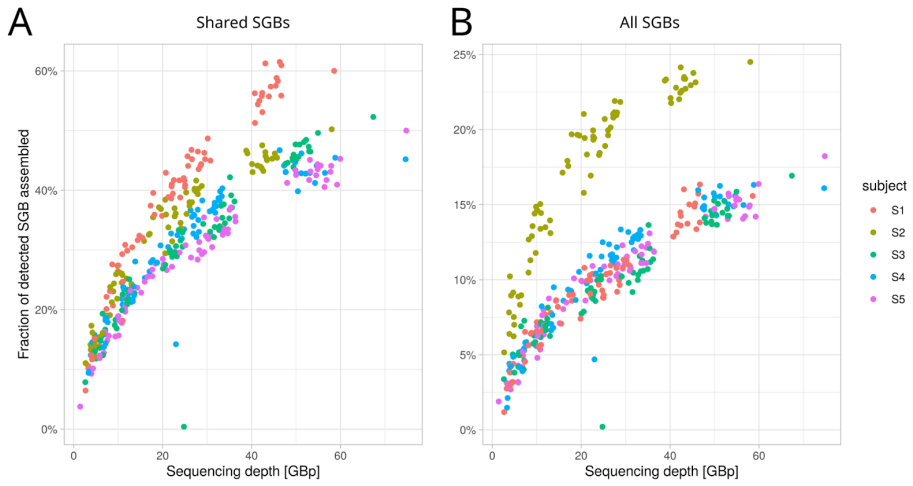

Fig. S3 Scatterplots showing how the fraction of SGBs that were both detected by reference-based profiling and assembly changes at increasing levels of sequencing depth. In A, the fraction is computed taking into consideration only the MetaPhlAn4 SGBs that were also assembled at least once, in B the fraction of SGBs assembled is computed by taking into consideration also all the SGBs that were never assembled.

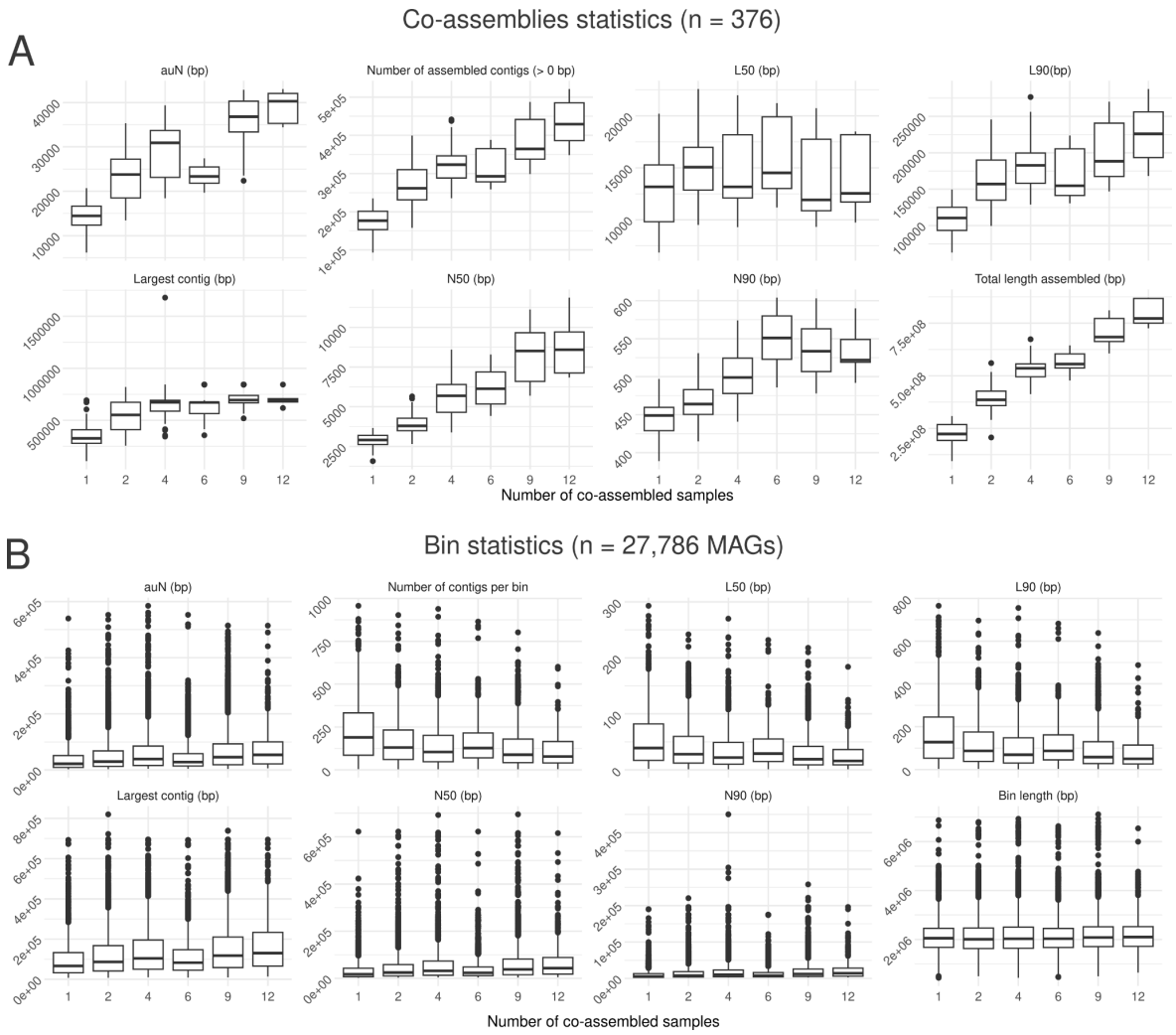

Fig. S4: Boxplots showing assembly statistics for the overall co-assemblies (A) and the MQ/HQ MAGs obtained from their binning (B)

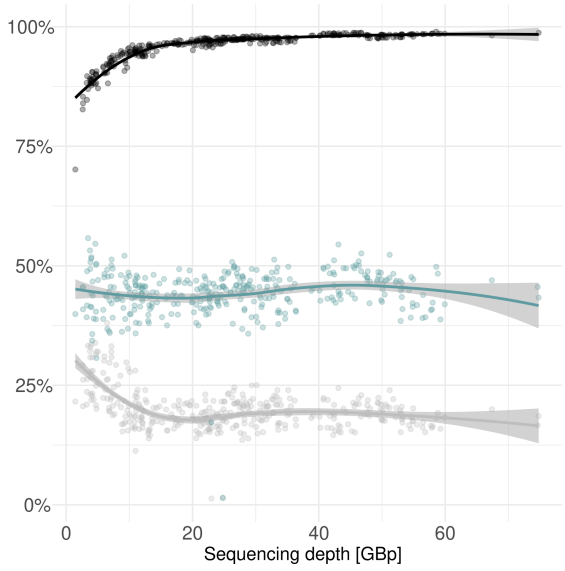

Fig. S5: Scatterplot showing (1) % of reads mapping to assembled contigs (black line), (2) assembled sequence length >1,500bp binned into medium/high-quality MAGs (light blue line), and (3) assembled contigs >1,500bp binned into medium/high-quality MAGs (grey line). Each point represents a different co-assembly combination, with trend lines showing the overall pattern.

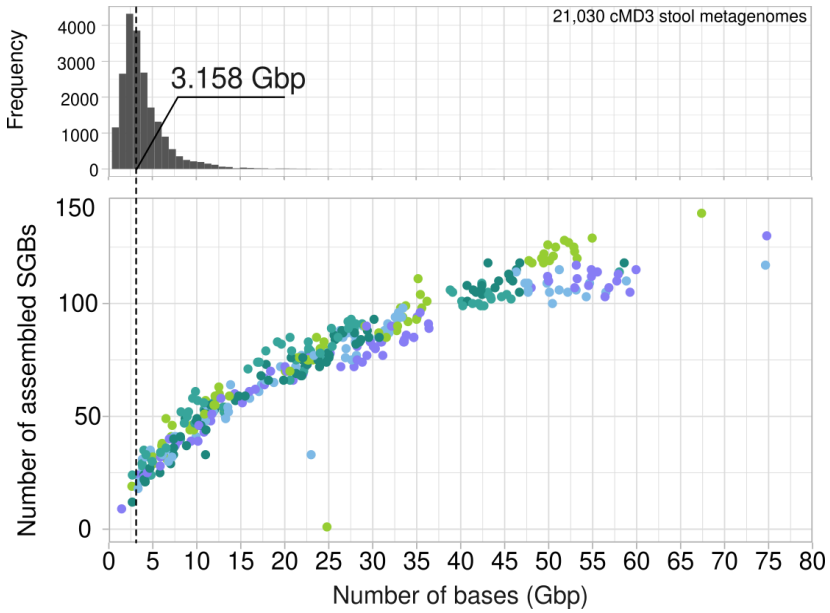

Fig. S6 Scatterplot showing the number of MAGs reconstructed from the same microbial communities of the five subjects in metagenomes of increasing read depth. Above the scatterplot, the read-depth distribution of all stool metagenomic samples available in curatedMetagenomicData3.0 (see Methods, DNA Sequencing and read pre-processing). The dashed line corresponds to the median read depth of metagenomes in cMD (median: 30.6×10^6 reads. IQR = 27.4×10^6) used in metagenomic studies and crosses the curve with the number of SGBs that have been assembled across 376 metagenomics combinations.

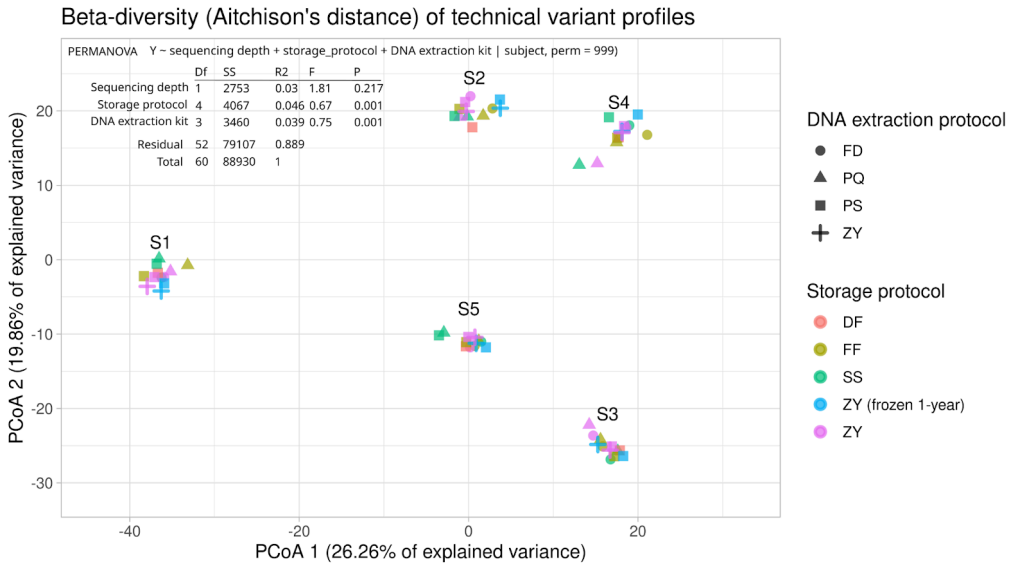

Fig. S7: Beta-diversity analysis of the technical variants of the five subjects. Dots are stratified by DNA extraction protocols (shape) and storage protocol employed (color). Subjects are reported above each cluster

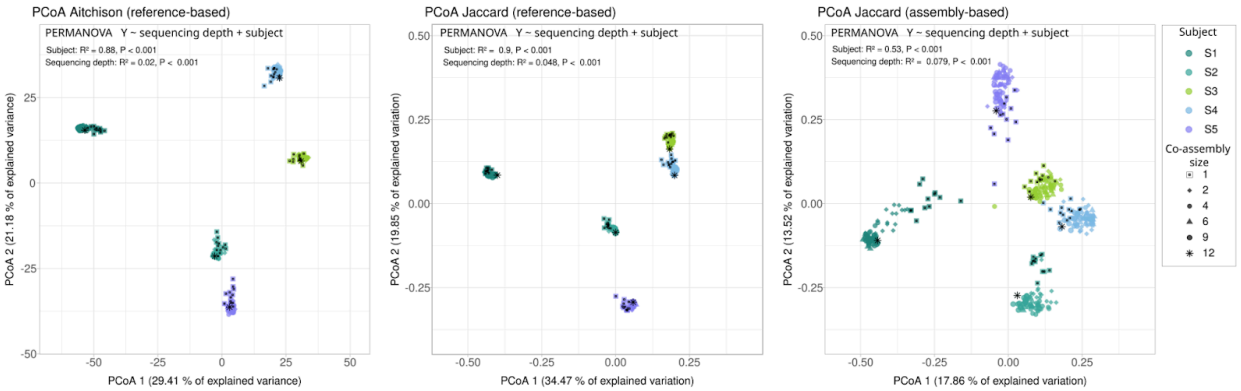

Fig. S8: Comparison of beta diversity across subjects and metagenomic combination sizes using three approaches: reference-based profiles with Aitchison distances (left), reference-based profiles with Jaccard distances (middle), and assembly-based profiles with Jaccard distances (right). Shapes indicate combination sizes, and colors represent different subjects.

https://drive.google.com/file/d/1QdB8SL6rypDHf5zrR62BSO2hOOcNFBmX/view?usp=drive_link

Fig. S9 an extended version of the heatmap in Fig. 2 containing all SGBs detected by MetaPhlAn4 and assembled into a MQ/HQ MAG

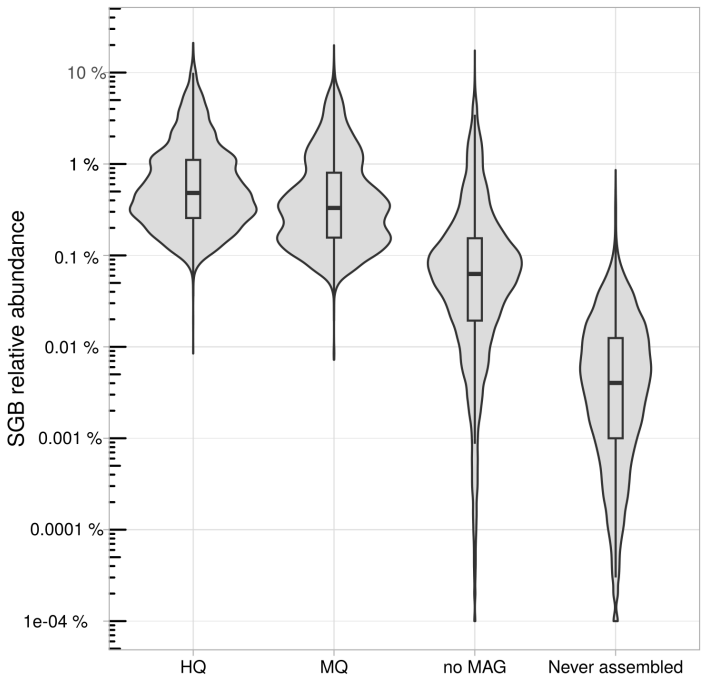

Fig. S10 Distributions of relative abundances for SGBs that were detected and assembled into HQ and MQ MAGs, not assembled but detected in the respective sample (no MAG) and SGBs that were detected by reference-based profiling but never assembled (Wilcoxon rank-sum test, P < 2.2e-16 for all comparisons).

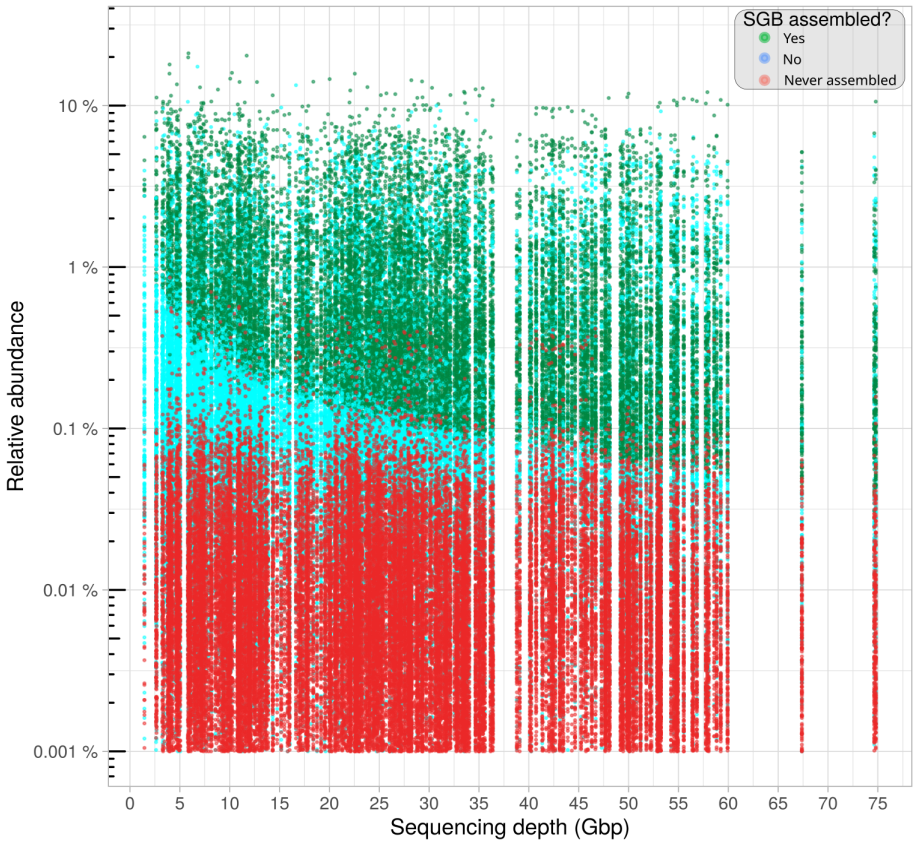

Fig. S11 Dotplot showing the relationship between the relative abundance of a species as detected by reference-based profiling and whether it has been successfully assembled into respective HQ/MQ MAGs. Each dot represents the relative abundance (y-axis) of a species detected in a specific technical-variant combination, along with the sequencing depth of the combination (x-axis). The color of the dot shows whether a MAG for that species in the metagenomic combination was successfully assembled (green), detected but not assembled (blue), or never assembled throughout the whole cohort (red).

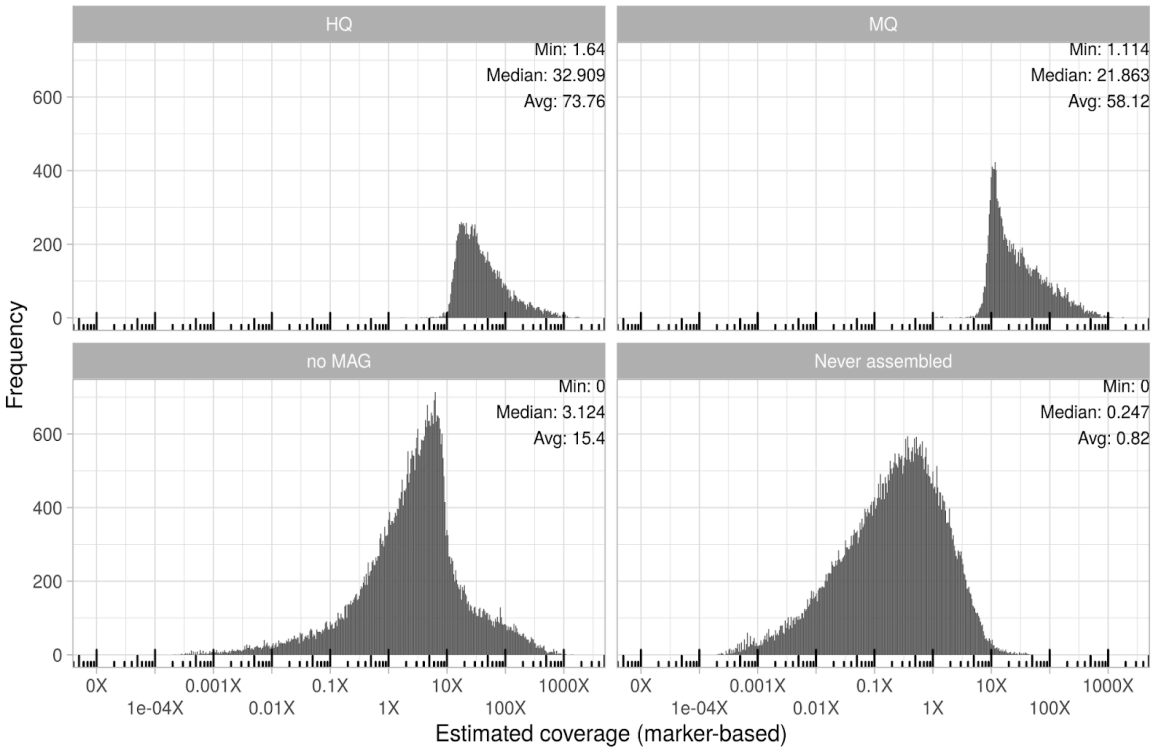

Fig. S12 distribution of the estimated marker-based coverages across SGBs that have been assembled as HQ, MQ, those that have been detected but not assembled in the respective metagenomic combination (no MAG), and those that have never been assembled but only detected throughout the whole cohort (Never assembled).

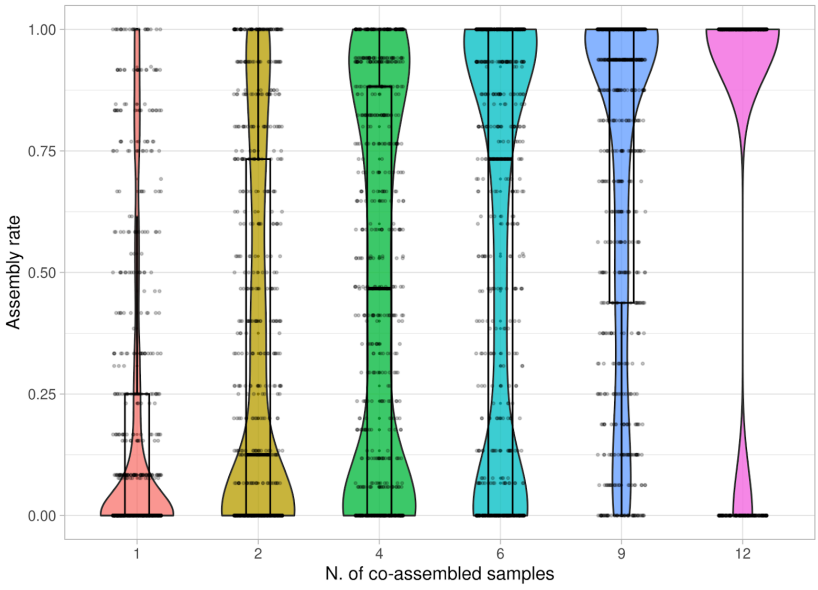

Fig. S13: Rate of assembly of SGB that have been reconstructed at least once in a subject's samples. Each dot represents the fraction of times (y-axis) an SGB has been assembled across metagenomic combinations of a certain category size (x-axis). An SGB was included in the plot only if it was assembled in at least one metagenomic combination of a subject. (Wilcoxon signed-rank test, P adjusted < 8.12e-58 for all comparisons, Table S5G)

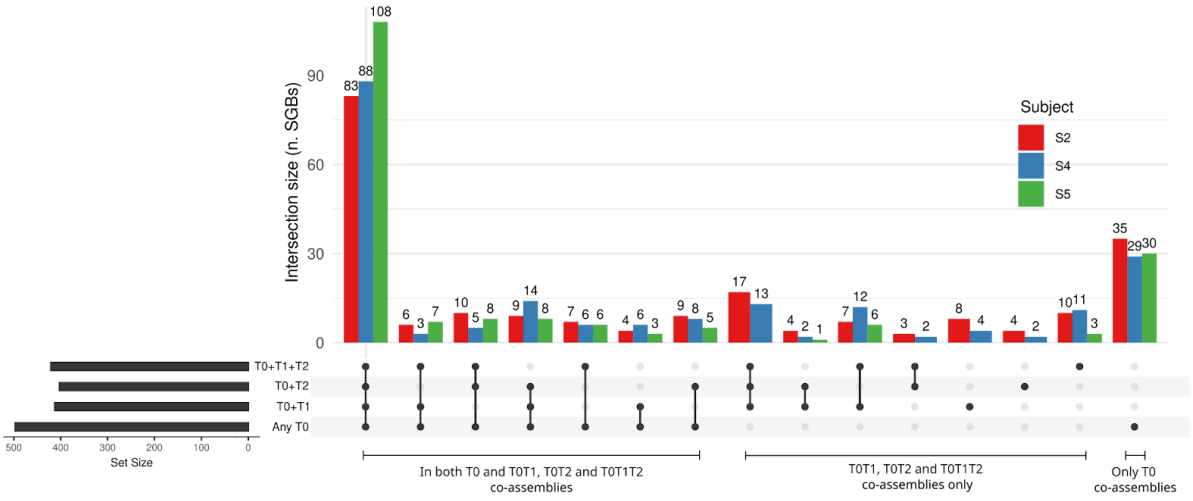

Fig. S14: Upset plot comparing SGBs reconstructed across three subjects' cross-sectional (T0) and longitudinal (T0+T1, T0+T2, T0+T1+T2) (co)assemblies. The vertical bars show the number of SGBs per subject that were shared among the intersecting (co)assembly categories, as indicated by the black dots and connecting lines in the matrix below. The horizontal bars on the left show the total number of distinct SGBs obtained from each (co)assembly category.

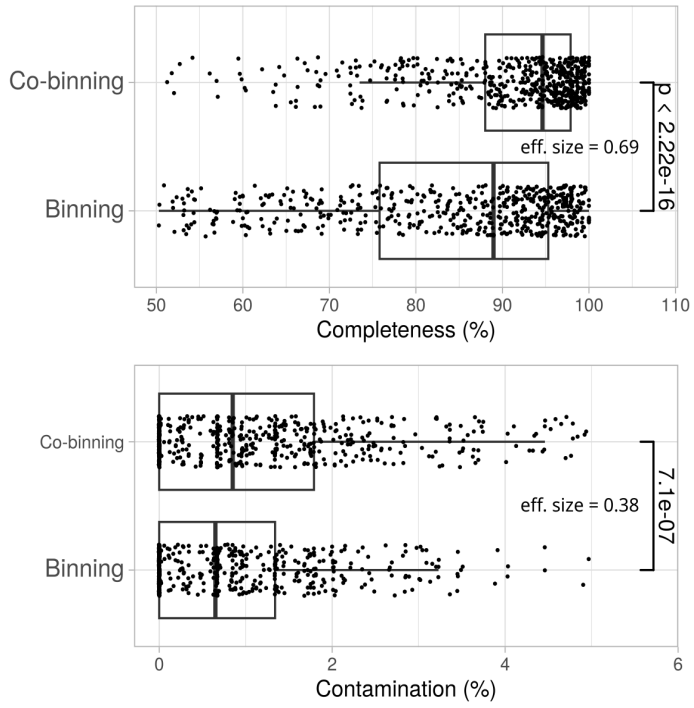

Fig. S15: Distribution of contamination and completeness of MAGs reconstructed using co-binning and binning on cross-sectional co-assemblies (T0, including also co-assemblies from subjects S1 and S3, excluded in the main figure). (Wilcoxon signed-rank test)

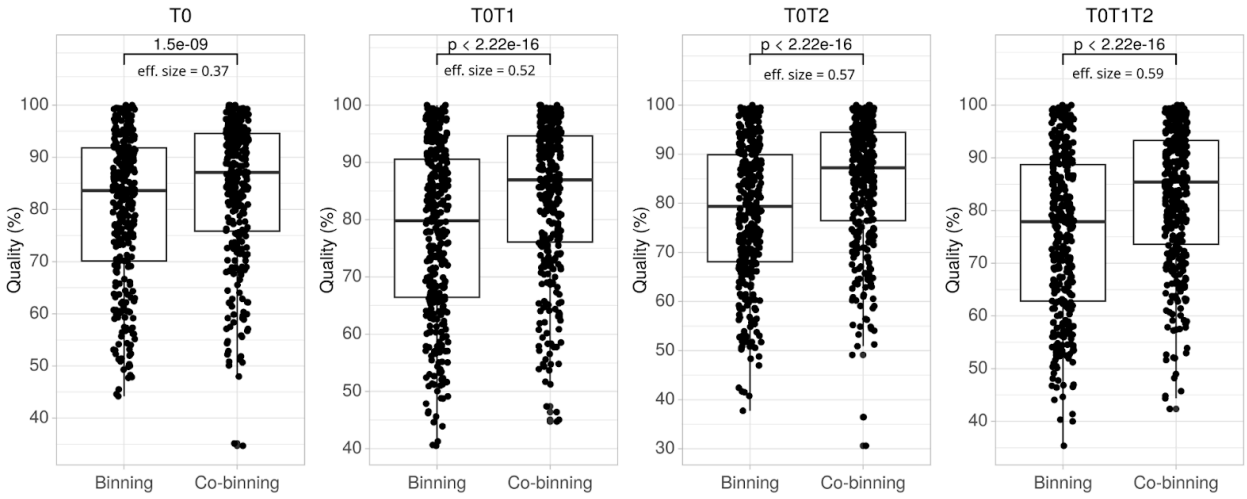

Fig. S16: Distribution of assembled MAG quality scores (Methods) reconstructed by binning as compared to co-binning on co-assemblies including longitudinal samples of the same subject (Wilcoxon signed-rank tests)

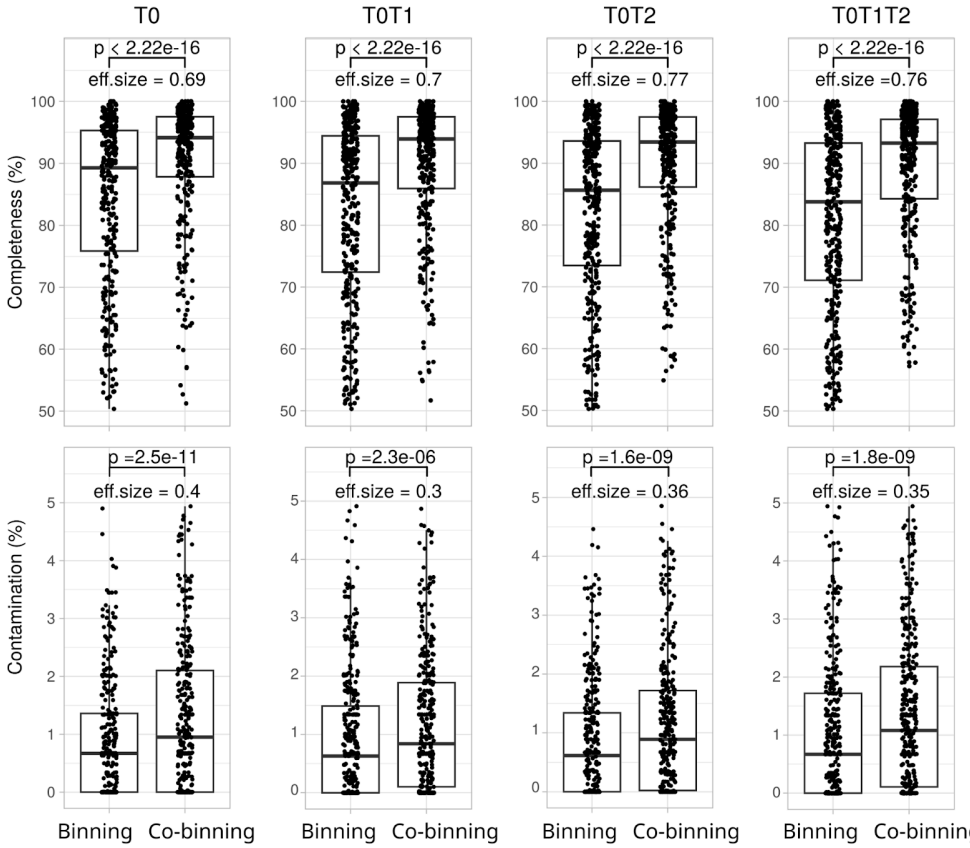

Fig. S17: Distributions of completeness and contamination across MAGs of SGB reconstructed both by binning and co-binning in cross-sectional and longitudinal co-assemblies (Wilcoxon signed-rank tests)

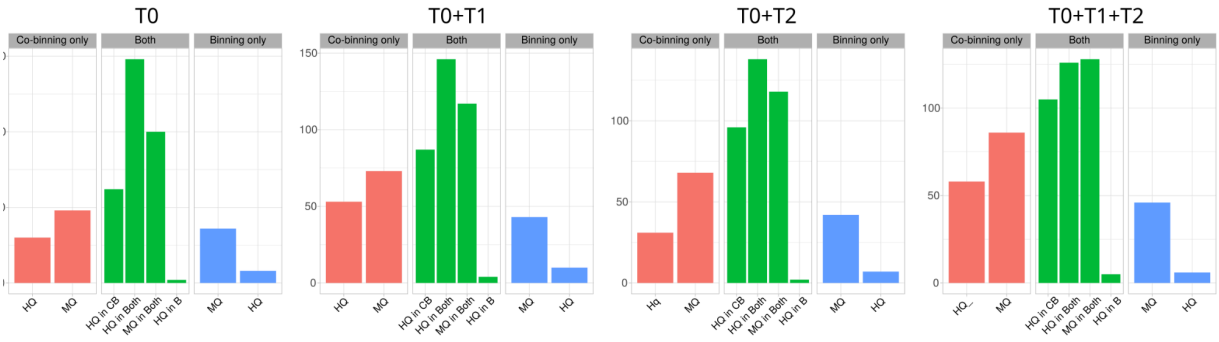

Fig. S18: Number of SGBs with HQ/MQ MAGs reconstructed exclusively by co-binning (red bars), binning (blue bars) or by both methods (green). Bars in green report whether shared SGBs passed from MQ to HQ with one method (HQ in CB or HQ in B), or were found as MQ/HQ with both.

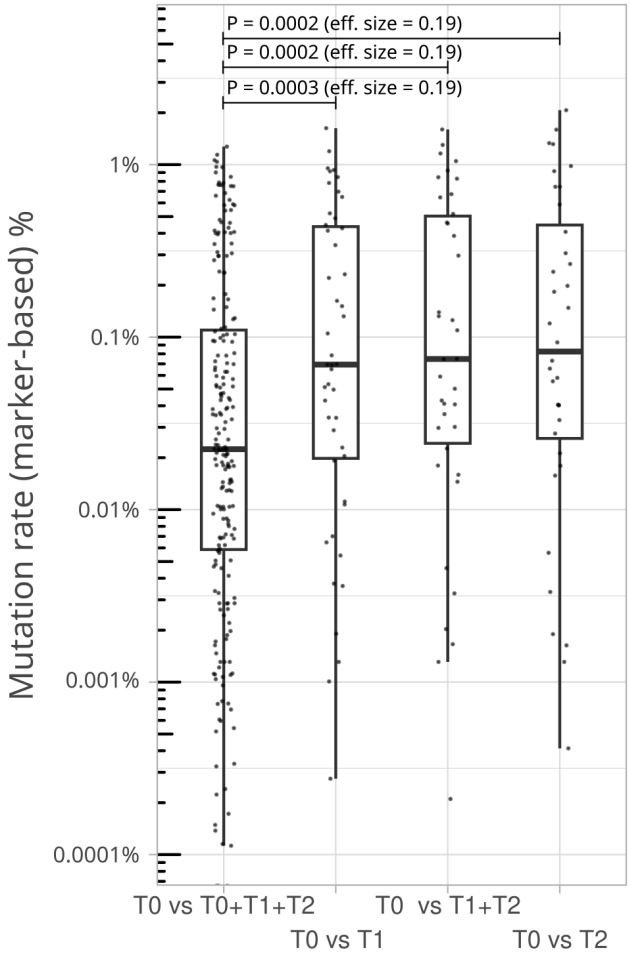

Fig. S19: Distributions of pairwise mutation rates between reference MAGs assembled from STM at T0 and MAGs assembled from STM at T1 (T0 vs T1), T2 (T0 vs T2), and MTMs that include T0 (T0 vs T0+T1+T2, which include pairwise comparisons with T0+T1, T0+T2 and T0+T1+T2) or not (T0 vs T1+T2).

Chapter 4 - Other Contributions

During my PhD, I contributed to projects both within and outside the laboratory. My work primarily involved taking care of the computational analyses of wet lab experiments, including reference-based metagenomic profiling of samples, assembly of bacterial isolates, and subsequent functional and genomic analyses. For each of the projects here presented, I report the main figure or analysis to which I contributed the most.

4.1 Screening, isolation, and cultivation of unknown gut microbial lineages

The first line of contribution consisted in the profiling of volunteers to screen for the presence of potential novel microbial lineages that were represented by MAGs but no available reference genomes. The aim was to isolate and cultivate these organisms and characterize these organisms through *in vitro* and *in silico* methods. This ongoing project resulted in three genome announcements. We provided preliminary descriptions of *Neopoerus faecalis*, representing a novel genus within the *Oscillospiraceae* family, and a putative novel *Catenibacterium* species. The third announcement encompassed a collection of 46 isolates, many of which were previously uncharacterized and await further analysis.

On the practical side, my contribution consisted of generating the metagenomic profiles, assembling the isolates' genomes into single amplified genomes (SAGs), performing preliminary *in silico* genomic and functional characterization, and comparative phylogenomic analyses to compare the unknown isolates to the closest available references.

4.1.1 Draft Genome Sequence of Neopoerus faecalis gen. nov., sp. nov., an Oscillospiraceae Strain Isolated from Human Feces

Marta Selma-Royo[#], Liviana Ricci[#], **Davide Golzato**, Charlotte Servais, Federica Armanini, Francesco Asnicar, Federica Pinto, Nicola Segata

[#]Contributed equally

Abstract: Here, we report the isolation and genome assembly of a strictly anaerobic bacterium from a previously uncharacterized species in the *Oscillospiraceae* family, isolated from a fecal sample from a healthy adult human. The name *Neopoerus faecalis* gen. nov., sp. nov. is proposed.

Published on *Microbiology Resource Announcements,* 21 June 2023

The analysis I made that resulted in the figure report below (Figure 1 of the paper) shows the phylogenetic analysis, based on 400 universal prokaryotic markers, performed between an isolate from a previously undescribed bacterial lineage and taxonomically annotated reference genomes from NCBI. Before this work, evidence for the existence of this microbial lineage was based on MAGs assembled across multiple cohorts, despite being quite prevalent (40.6%) across more than 24,500 metagenomes from five continents.

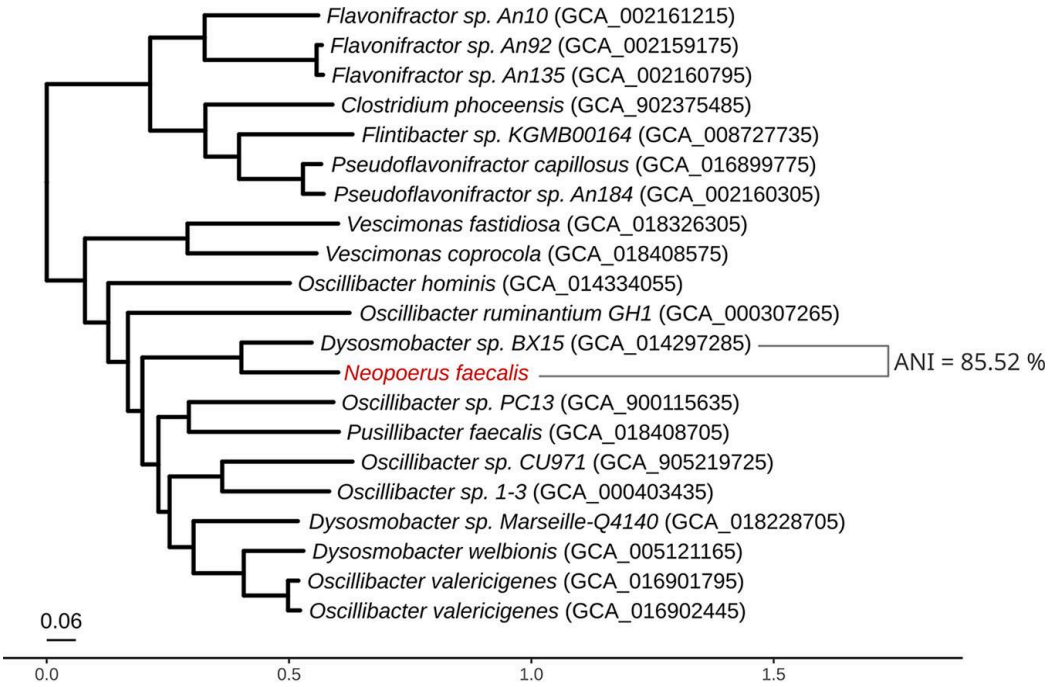

Fig. 5 Phylogenetic tree of the *Neopoerus faecalis* gen. nov, sp. nov. isolate and related taxa with available reference genomes based on whole-genome sequencing (WGS). Sequence accession numbers are given in parentheses. The closest taxonomically defined species is *Dysosmobacter welbionis* (GenBank accession number GCA_005121165), with an ANI of <80%. The closest available reference genome, *Dysosmobacter* sp. strain BX15 (GCA_014297285), is mislabeled at the genus level, since it has an ANI of <85% to the *Dysosmobacter* reference strain (*Dysosmobacter welbionis*).

The closest species to our isolate was labeled as a member of the *Dysosmobacter* genus and had an average nucleotide identity (ANI) of 85.52% with it, suggesting our isolate to be also from the same genus. However, the genome of *Dysosmobacter* sp. BX15 was considerably distant in terms of ANI (<85%) from the only taxonomically well-defined member (*D. welbionis*) of the genus, suggesting it was incorrectly labeled. This prompted us to warrant for the definition both of a novel genus and species for our isolate, *Neopoerus faecalis*.

I've also performed preliminary functional annotation of the *N. faecalis* genome, which revealed the presence of genes related to complex carbohydrate degradation, including cellulose and chitin. Below I report the table with the main genome composition features of the assembly.

This work, along with the following two genome announcements, exemplifies how integrating MAGs of unknown species into reference-based approaches facilitates their targeted isolation and characterization. By leveraging specific genomic information contained in the MAGs (e.g: SGB-specific markers) it becomes possible to design informed strategies for screening and culturing previously uncultured microbes. Although *N. faecalis* was not specifically among the species MAGs reconstructed the main work of chapter 3, the same approach used here could be applied to screen for potential donors for the 39 novel microbial lineages detected there and facilitate their informed isolation.

Parameter	Value
Total length (bp)	2,542,847
No. of scaffolds	36
GC content (%)	56.22
Mean coverage (×)	116
Size of longest scaffold (bp)	496,649
N50 (bp)	274,696
L50	4
No. of coding sequences	2,408
No. of rRNAs	7
No. of tRNAs	57
Estimated completeness (%)	98.66
Estimated contamination (%)	0

Table 1: Statistics of genomic features of the *Neopoerus faecalis* genome assembly

4.1.2 Draft genome sequence of a representative strain of the Catenibacterium genus isolated from human feces

Liviana Ricci[#], Marta Selma-Royo[#], **Davide Golzato**, Amir Nabinejad, Charlotte Servais, Federica Armanini, Francesco Asnicar, Federica Pinto, Sabrina Tamburini, Nicola Segata

[#]Contributed equally

Abstract: A strain from a previously undescribed species belonging to the *Catenibacterium* genus was isolated from the stool of a healthy volunteer. The strain is strictly anaerobic, and the genome encodes a CRISPR-Cas system and genes related to trimethylamine production.

Published on *Microbiology Resource Announcements,* 26 july 2023

The *Catenibacterium* genus belongs to the *Erysipelotrichaceae* family and the species *C. mitsuokai* is the only one that has been taxonomically well-defined. Currently, more than 1,000 MAGs belonging to this genus have been assembled, suggesting intra-genus phylogenetic heterogeneity is potentially still undescribed.

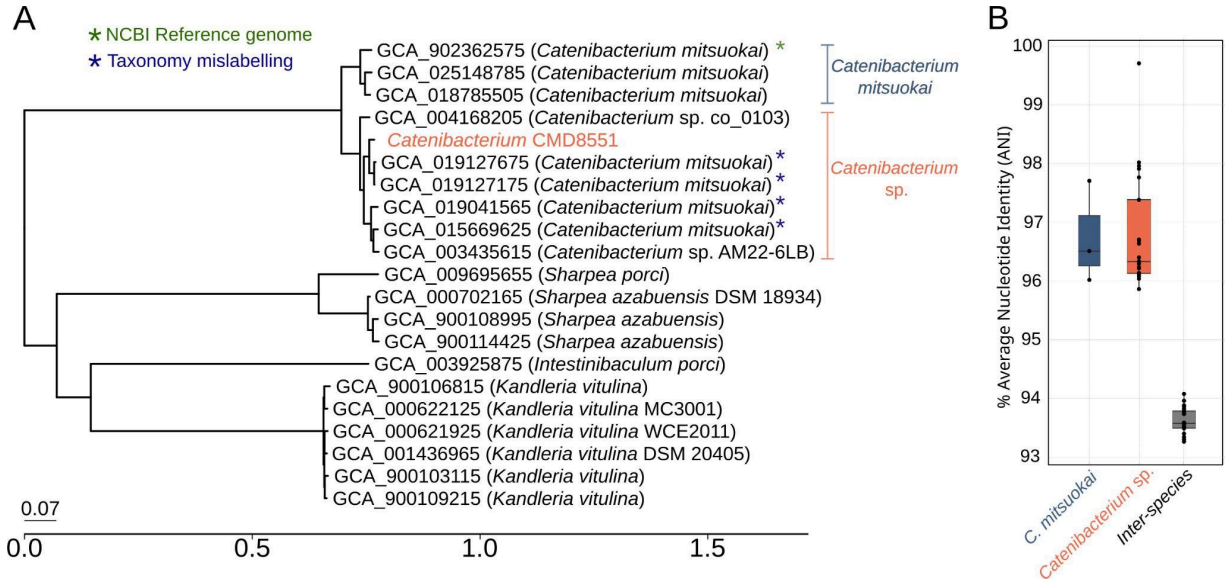

Fig. 6 (A) Phylogenetic tree of the new *Catenibacterium* isolate and related taxa with available reference genomes obtained from sequencing of isolates. The marked reference genomes were taxonomical missasigned (as *C. mitsuokai*) based on average nucleotide identity (ANI, <94%) to its reference strain (GCA_902362575). These reference genomes, along with the CMD8551 strain, cluster as a different taxonomic group distant from *C. mitsuokai* compatible with a new species inside the *Catenibacterium* genus. (B) Comparison of the ANI of the available genomes from *Catenibacterium*

genus inside each species genomes defined by the phylogenetic tree (*C. mitsuokai* and the proposed *C. tridentinum*) and between them (inter-species).

In this work, we isolated and cultivated a member of this undercharacterized genus. The figure above [**Fig. 1 A**] shows a phylogenetic analysis including our isolate assembly (CMD8551) in comparison to NCBI reference genomes, we observe that this isolate is close to other *Catenibacterium* genomes. However, it is possible to see that these genomes, almost all labeled as *C. mitsuokai,* separate into two distinct subtrees. The distribution of pairwise ANI, computed between the genomes of these two phylogenetic branches (inter-species boxplot, **Fig. 1 B**), is always below the 95% species-defining ANI threshold (Jain et al. 2018) warrants for the definition of a novel species to distinguish these of our isolate branch to the other one (*C.mitsuokai)*. Additionally, similar to our previous isolation work (4.1.1), I found that many genomes that cluster with our *C. tridentinum* isolate, are labeled as *C. mitsuokai*, but display ANI < 95% compared to the *C. mitsuokai* reference genome that first defined this species. This suggests these genomes have been mislabeled and instead belong to the *C. tridentinum* species defined here.

Interestingly, functional annotation of this *C. tridentinum* isolate genome has shown that it encodes for many glycoside hydrolases that suggest extended carbohydrate catabolic potential, and for genes putatively involved in antibiotic resistance to glycopeptides and carbapenems.

Parameter	**Value**
Total length (bp)	2,320,430
No. of scaffolds	113
GC content (%)	33.7
Mean coverage (×)	519×
Size of longest scaffold (bp)	105,579
N50 (bp)	44,791
L50	18
No. of coding sequences	2,239
No. of rRNAs	10 rRNA
No. of tRNAs	76 tRNA
Estimated completeness (%)	100%
Estimated contamination (%)	0%

Table 2: Statistics of genomic features of the new *Catenibacterium tridentinum* genome assembly

4.1.3 Draft genome sequences of multiple bacterial strains isolated from human feces

Liviana Ricci[#], Marta Selma-Royo[#], **Davide Golzato**[#], Amir Nabinejad, Charlotte Servais, Federica Armanini, Francesco Asnicar, Federica Pinto, Sabrina Tamburini, Nicola Segata

[#]Contributed equally

Abstract: Bacterial isolation is necessary for functional and mechanistic analyses, and the increased human microbiome diversity revealed by metagenomic sequencing is expanding the relevant cultivation targets. Here, we report 46 draft genome sequences of bacterial isolates obtained from fecal samples of healthy adults in Trento and Milan (Italy), including strains from seven taxonomically uncharacterized species

Published on *Microbiology Resource Announcements,* 29 May 2024

My contribution to this work consisted of assembling sequencing data from all the isolates, obtained from nine volunteers, so that we could reliably culture, taxonomically annotate, and curate them into the CibioCM collection of single-amplified genomes. Seven of these SAGs represent the first members of novel lineages from the *Oscillospiraceae*, *Coriobacteriaceae, Clostridia* and *Collinsella* taxa.

4.2 Distinct Chromosomal Mutation Associated with Cefiderocol Resistance in *Acinetobacter baumannii*: A Combined Bioinformatics and Mass Spectrometry Approach to unveil and validate the in-vivo acquired Chemoresistance

The emergence of multi-drug resistant species (MRS) is a significant threat to global health, with hospitals providing an environment that favors the rapid selection of resistant strains. In this study, we analyzed two clinical isolates of *Acinetobacter baumannii* from the same hospital, one resistant to a last-resort antibiotic and the other susceptible to it.

Lavinia Morosi[#], **Davide Golzato[#]**, Linda Bussini, Hygerda Guma, Federica Tordato, Federica Armanini, Zian Asif, Francesco Carella, Paola Morelli, Michele Bartoletti, Giorgio Da Rin, Erminia Casari, Giuseppe Martano, Maria Rescigno, Nicola Segata, Sara Carloni*, Valeria Cento*

[#]Contributed equally
*Contributed equally

Abstract: Antibiotic-resistant *Acinetobacter baumannii* strains pose a significant public health concern, particularly in clinical settings. Cefiderocol (FDC), a novel siderophore cephalosporin, shows promise as a rescue treatment due to its favorable resistance profile. However, the emergence of in-vivo acquired FDC-resistant *A. baumannii* strains necessitates rapid clinical profiling and identification of genomic mutations to improve clinical management and reduce treatment failures. This study analyzed two carbapenem-resistant *A. baumannii* isolates obtained from co-hospitalized patients with bacteremia. Initial multi-locus sequence typing (MLST) of whole-genome sequences suggested strain similarity. However, phenotypic minimum inhibitory concentration (MIC) analysis revealed divergent FDC susceptibility between the isolates. Genomic comparison through variant calling identified key differences, including a novel chromosomal mutation in a gene putatively encoding a homolog of the TonB-dependent receptor, involved in ferric-siderophore and heme uptake. This frameshift mutation resulted in a premature stop codon, suggesting acquired resistance due to loss of function. Functional characterization using liquid chromatography-tandem mass spectrometry (LC-MS/MS) demonstrated significantly impaired antibiotic uptake and intracellular accumulation in the FDC-resistant *A. baumannii* strain, consistent with the predicted loss of function. This study presents a novel clinical translational approach for recognizing and characterizing emerging genetic mutations against newly approved antimicrobials, revealing strain differences undetectable by MLST alone. The combination of variant calling analysis and

LC-MS/MS technology enables rapid detection and functional characterization of FDC resistance in *A. baumannii* isolates. This approach could be integrated into clinical practice for monitoring in-vivo resistant isolates and providing an unbiased epidemiological overview of resistance strains and mechanisms. Furthermore, this methodology may be applicable to other emerging antimicrobial agents and bacterial pathogens, potentially improving the overall management of antibiotic-resistant infections.

Published on *Frontiers in Microbiology* (doi: 10.3389/fmicb.2024.1480322)

	Isolate annotation						Variant calling of 5577 vs 5406			
Isolate 5577 Contig ID	**CDS start position**	**CDS end position**	**strand**	**Locus tag**	**Prokka product annotation**	**Blastp majoritary annotation**	**Variant position**	**REF**	**ALT**	**INFO**
EECGJPIA_2	6370	7032	-	EECGJPIA_00294	hypothetical proteins	Ribonuclease I	6861	G	A	SNP
EECGJPIA_2	202652	203074	-	EECGJPIA_00493		isoprenylcysteine carboxylmethyltransferase family protein	202679	A	AT	INS
EECGJPIA_22	100	2247	-	EECGJPIA_03186		TonB dependent receptor	777	C	T	SNP
EECGJPIA_22	100	2247	-	EECGJPIA_03186		TonB dependent receptor	126	CGAGGCTATAA	C	DEL

Table 3: This table reports the mutation detected by variant calling, contextualized to specific contigs and coding sequences (CDS) of the non-resistant isolate 5577, here used as reference. It includes which contigs of isolate 5577 assembly are found to be mutated in the resistant isolate 5406, the start and end positions of the mutated CDS within the contig and the strand orientation of the CDS. Additionally, we report the annotations of the putative encoded proteins encoded by these genes, obtained by Prokka and blastp to the NCBI-nr/nt database. Details on the mutation, such as the reference allele (REF) and the alternative (ALT), and the respective types (INFO, "ins" for insertion, "snp" for single nucleotide polymorphism, "del" for deletion) are reported.

This analysis highlights how routinely sequencing and performing genomic analysis of antibiotic-resistant isolates in hospitals can aid in the early detection of potential MRS outbreaks and provide experimental evidence on the role of specific proteins.

4.3 Short-term cocoa supplementation influences microbiota composition and serum markers of lipid metabolism in elite male soccer players

Diet impacts every aspect of human health and is one of the main shaping factors of the gut microbiome composition, which in turn also affects health. Sports nutrition is one of the many fields interested in developing specific dietary strategies to improve the athletic performances of professional athletes. In this study, we have assessed the impact of daily cocoa supplementation in 32 soccer players, by monitoring for changes in the blood lipid profiles, used as a marker for inflammation, and whether it also affected their gut microbial composition. My contribution to this study involved profiling the gut microbial profiles of the players, before and after the dietary intervention, to show how much the microbial community changed and specifically show which species changed their abundances.

Laura Mancin, Ian Rollo, **Davide Golzato**, Nicola Segata, Cristian Petri, Luca Pengue, Luca Vergani, Nicolò Cassone, Alessandro Corsini, Joao Felipe Mota, Stefania Sut, Stefano Dall'Acqua, Antonio Paoli

Abstract

Objectives**:* Dietary strategies to improve arachidonic acid:eicosapentaenoic acid (AA:EPA) ratios are of interest due to potential reductions in inflammation and oxidative stress following exercise. The aim of this study was to investigate the impact of a novel dietary intervention, that is, the ingestion of 30 g of dark chocolate, on blood lipid profiles and gut microbiota composition in elite male soccer players. Methods: Professional male soccer players were randomly assigned to the experimental group (DC) provided with 30 g of dark chocolate or to the control group (WC), provided with 30 g of white chocolate, for 30 days. Before and after intervention, blood, fecal sample, and anthropometry data were collected. For each outcome, two-way repeated-measure analysis of variance was used to identify differences between baseline and endpoint (Week 4), considering treatment (dark chocolate, white chocolate) as intersubjects' factors. Metagenomic analysis was performed following the general guidelines, which relies on the bioBakery computational environment. ***Results: DC group showed increased plasma polyphenols (from 154.7 ± 18.6 μg gallic acid equivalents/ml to 185.11 ± 57.6 μg gallic acid equivalents/ml, Δ pre vs. post = +30.41 ± 21.50) and significant improvements in lipid profiles: total cholesterol (Δ −32.47 ± 17.18 mg/dl DC vs. Δ −2.84 ± 6.25 mg/dl WC, Time × Treatment interaction $p < .001$), triglycerides (Δ −6.32 ± 4.96 mg/dl DC vs. Δ −0.42 ± 6.47 mg/dl WC, Time × Treatment interaction $p < .001$), low-density lipoprotein (Δ −18.42 ± 17.13 mg/dl vs. Δ −2.05 ± 5.19 mg/dl WC, Time × Treatment interaction $p < .001$), AA/EPA ratio (Δ −5.26 ± 2.35; −54.1% DC vs. Δ −0.47 ± 0.73,

−6.41% WC, Time × Treatment interaction p < .001) compared with WC group. In addition, 4 weeks of intervention showed a significant increase in high-density lipoprotein concentration in DC group (Δ + 3.26 ± 4.49 mg/dl DC vs. Δ −0.79 ± 5.12 mg/dl WC). Microbial communities in the DC group maintained a slightly higher microbial stability over time (exhibiting lower within-subject community dissimilarity). **Conclusion**: Ingesting 30 g of dark chocolate over 4 weeks positively improved AA:EPA ratio and maintained gut microbial stability. Dark chocolate ingestion represents an effective nutritional strategy to improve blood lipid profiles in professional soccer players. **What Are the Findings?** Ingesting 30 g of dark chocolate for 4 weeks positively influences blood lipid AA: EPA ratio while maintaining gut microbial stability. **What This Study Adds?** Dietary intake of specific foods such as dark chocolate represents an alternative strategy to support the health and recovery of elite soccer players. **What Impact Might This Have on Clinical Practice in the Future?** From a clinical and translational perspective, dark chocolate ingestion positively modulates favorable blood lipid profiles and polyunsaturated fatty acid metabolism while maintaining gut microbial stability. Dark chocolate ingestion may be considered as an effective nutritional strategy in elite sport environments during periods of high-intensity training and congested competitions. Further research is required to determine functional outcomes associated with the observed improvements in blood lipid profiles.

Published on *International Journal of Sport Nutrition and Exercise Metabolism*, 7 August 2024 (doi: 10.1123/ijsnem.2024-0012)

In the analysis below I specifically show that no particular pattern in the beta diversity changes appear to be specifically associated with the treatment, although the microbial compositions of the group of athletes consuming dark chocolate appear to be slightly more stable than that of athletes consuming white chocolate.

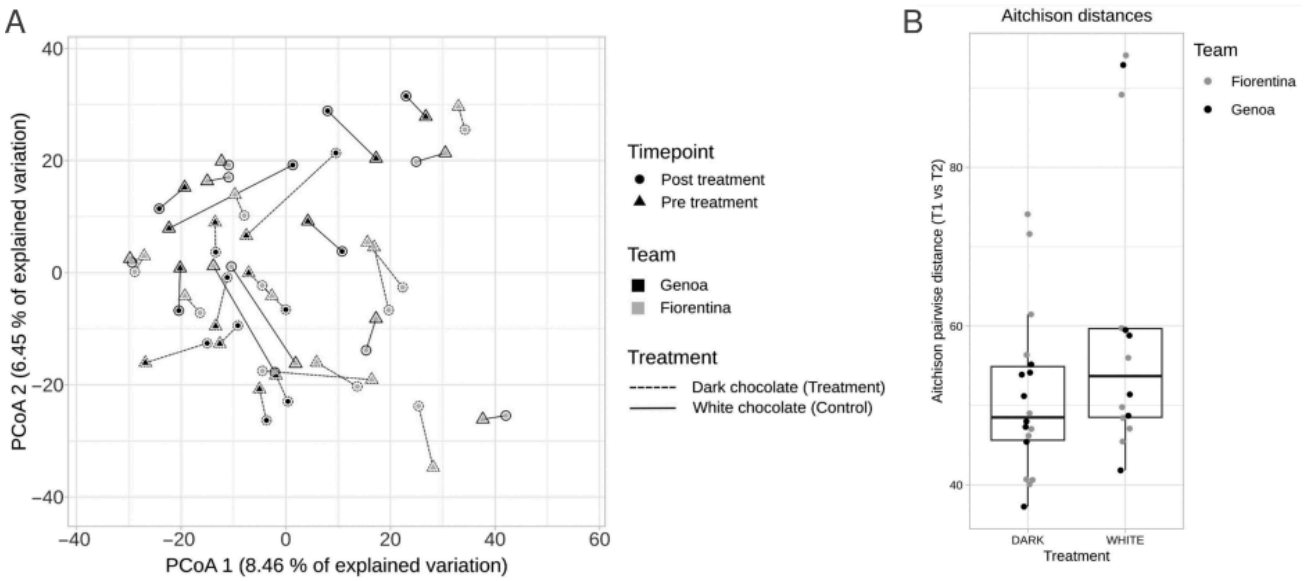

Fig. 7 (A) Beta-diversity of the two dietary intervention groups, showing no striking microbial composition shifts directly associated with the short-term cocoa supplementation. (B) However, microbial communities of players that consumed dark chocolate appear to be more slightly stable throughout time, as shown in the pairwise comparison of beta-diversity pairwise distances before and after treatment.

Although not reported here, a differential abundance analysis that I've performed shows that SGBs belonging to the *Blautia* and *Coprococcus* genera appear to be differentially abundant after the treatment in the group that consumed dark chocolate, but not in the control group consuming white chocolate.

4.4 Meta-analysis of 22,710 human metagenomes defines an index of oral to gut microbial introgression and associations with age, sex, BMI, and diseases

Statistical associations between microbial features and host-related metadata can suggest potential involvement of the microbiome in human health. For example, the presence or absence of certain microbial species can be predictive of health status or certain diseases. However, single-study associations may suffer from study-specific confounding biological and technical factors, potentially making them spurious and not generalizable to other studies. Meta-analysis provides a statistical framework to summarize associations across multiple studies, verifying consistency across independent cohorts. The curatedMetagenomicData3 (cMD3) database collects taxonomic and functional profiles from >20,000 metagenomes, structured around manually curated metadata in a standardized grammar. This study leveraged this resource to find multi-cohort supported associations between microbial taxa, or related metabolic functions, and host health.

Paolo Manghi, Lucas Schiffer, **Davide Golzato**, Jennifer Wokaty, Francesco Beghini, Chloe Mirzayi, Kaelyn Long, Kai Gravel-Pucillo, Gianmarco Piccinno, Samuel David Gamboa-Tuz, Arianna Bonetti, Giacomo D'Amato, Rimsha Azhar, Kelly Eckenrode, Fatima Zohra, Valentina Giunchiglia, Marisa Keller, Anna Pedrotti, Ilya Likhotkin, Shaimaa Elsafoury, Ludwig Geistlinger, Aitor Blanco-Miguez, Andrew Maltez Thomas, Moreno Zolfo, Marcel Ramos, Mireia Valles-Colomer, Sabrina Tamburini, Francesco Asnicar, Heidi Jones, Curtis Huttenhower, Vincent Carey, Sean Davis, Edoardo Pasolli, Sehyun Oh, Nicola Segata,#*, Levi Waldron,#*

Abstract

Publicly available metagenomic profiles of the human microbiome are now of sufficient scope for meta-analysis on a range of metabolic and health outcomes, but lack of accessibility and standardization are barriers to their utilization. Here, we use 22,710 uniformly processed metagenomes from 94 cohorts and 42 countries with manually-curated metadata from the newly developed curatedMetagenomicData (cMD) 3 to address several ongoing challenges in human microbiome epidemiology. These include assessing the relationship of the human microbiome with basic and clinical host outcomes, which can be performed in cMD 3 with a greater sample size and diversity of populations than previously possible. Using random-effects meta-analysis we identify species and metabolic features generally associated with sex (102 taxa, 15 MetaCyc pathways, and 226 KEGG-Orthologs (KO), FDR = 0.01), age (91 taxa and 1,991 functions, FDR = 0.01) or BMI (195 taxa and 1,018 functions,

FDR = 0.01), and generally with health or disease across 15 diseases (34 species and 215 pathways, FDR = 0.01). Typically oral-associated species are among those systematically elevated in disease, and we use these to define a simple “oral to gut introgression” score that can be easily calculated in any fecal sample as the summed relative abundance of 305 species identified from 857 oral metagenomes. This score is more frequently elevated than decreased in disease (29 of 30 studies, P = 5.8 x 10-8, binomial test), representing a quantitative index of deviation from health-associated microbiome configurations. Together these analyses identify modest but widely shared variation in human microbiomes that will serve as a reproducible and readily updatable reference.

In revision

My contribution to this work consisted in writing and coding guided tutorials to show how to reproduce the meta-analyses conducted in this work, additionally guiding potential cMD3 users showing practical examples. The tutorials are available at

- https://waldronlab.io/curatedMetagenomicDataAnalyses/articles/Age_metaanalysis_vignette.html

- https://waldronlab.io/curatedMetagenomicDataAnalyses/articles/Sex_metaanalysis_vignette.html

Chapter 5 - Conclusions and Future Perspectives

In this thesis I explored computational approaches to comprehensively profile a gut microbial community and characterize the hidden biodiversity that is systematically neglected in gut microbiome studies due to technological limitations.

A requisite for a full ecological and deterministic understanding of the host-gut microbiome system is the knowledge of all its discrete components and related functions, namely the genomes and genes of each bacterial lineage within the community. The abundance of a microbial species is not necessarily proportional to its ecological importance in the community, nor to its influence on host health. Rare and low-abundant members can thus also play essential roles. However, while detecting their presence in a sample when a reference genome is available is relatively straightforward, our ability to reconstruct their genomes is very limited due to low genomic coverage. Therefore, I consider the development and optimization of methods to enhance metagenomic data usage for characterizing such species to be of utmost importance.

I showed through the application of multi-sample approaches how most species present within a single microbial community can be uncovered and reconstructed only through deep WGS sequencing. My approach was to randomly merge technical variants (i.e. same microbial community but different library preparation protocols) into metagenomes of different sample sizes, allowing me to assess how assembly-based and assembly-free methods profile the same microbial community across a range of different sequencing depths and experimental conditions. This cohort's sample design is optimal for assessing the impact of such variables on metagenomic profiling, and although it is common in studies evaluating protocol optimization for stool metagenomics (Wesolowska-Andersen et al. 2014; Guan et al. 2021; Panek et al. 2018; Szóstak et al. 2022), to the best of my knowledge very few studies used technical variants to assess how multi-sample strategies for metagenomic profiling, especially assembly-based approaches, perform across different conditions.

Increased sequencing depth, obtained from the merge of cross-sectional samples, has a great impact on the number of MQ and HQ MAGs that can be recovered from the same microbial community increasing sublinearly with it. Interestingly, this trend does not appear to plateau even at higher depths, indicating that metagenomic assembly could benefit from pooling even more technical variants. On the other hand, reference-based profiling is not drastically affected by the multi-sample approach, with compositional alpha-diversity estimates being only slightly underestimated in the single samples. Even though reference-based richness is always greater than that of the

assembly-based approach, the former increases faster with respect to sequencing depth.

The higher richness observed in reference-based profiling reflects the extensive representation of gut metagenomes in current databases (Blanco-Míguez et al. 2023). However, this approach can only detect microbial species that have been assembled enough times to infer markers for their reliable detection, with assembly remaining the primary method to assess their presence. This work estimates that 10X genomic coverage is the minimum average threshold to assemble a species MAG, a level not achieved for most low-abundance species at standard sequencing depths. Consequently, current MAG databases contain thousands of MAGs for abundant species but few or none for most low-abundant ones. Rare species have likely been assembled primarily due to their occasional enrichment in specific individuals across thousands of globally sequenced samples.

This suggests that the gap in current gut microbiome genome catalogs exists not only due to insufficient sampling of inter-individual microbial variability but also due to technical aspects of sequencing. Indeed, co-assembly of merely five individual gut communities allowed the recovery of genomes from 39 putatively new bacterial taxa, some potentially representing the first members of new prokaryotic families, classes, or even orders. Based on this observation, I suggest increasing sampling and sequencing efforts when dealing with gut metagenomes from underrepresented environments (e.g., non-westernized populations, individuals affected by rare diseases, or those exposed to particular environmental conditions).

Furthermore, an additional aspect to consider is that sample processing protocols can introduce biases in observed composition, ultimately affecting genomic coverage of specific taxa and the ease of their assembly. My work suggests that instead of assembling a single sample at increased sequencing depth, it might be worth to produce and co-assemble two samples at lower depth using different protocols (e.g: different DNA extraction kit) so as to balance out potential biases which might affect DNA concentration of specific taxa. While producing technical variants is costlier than increasing sequencing depth, incorporating more technical variation could be a potential strategy to maximize microbial diversity reconstructed from samples whose collection is difficult.

Another advantage of sampling the same community multiple times (longitudinally or cross-sectionally) over a single time with increased sequencing depth is the ability to leverage abundance covariation of contigs to improve the specificity of contig binning, i.e., co-binning. Sets of contigs whose coverage covaries across different samples are more likely to belong to the same bin. In this context, I show that co-binning is generally superior to single-sample binning, as it retrieves more MAGs, typically of higher quality.

This is especially evident when binning co-assemblies that include longitudinal metagenomes. Although not demonstrated in this work, I hypothesize that using both binning approaches, followed by MAG de-replication to retain those of the highest quality, can maximize the retrieval of MAGs from an individual's microbial community.

Characterizing microbial diversity at the species level is challenging, and achieving strain-level resolution can be even more problematic. The limitations arise not only from insufficient sequencing depth but also from intrinsic constraints of the DBG-based assembly algorithms. Strain variability within a single species can lead to increased assembly fragmentation, and most assemblers tend to collapse subtle genomic differences into MAGs whose contigs represent a consensus of the various strains. In the final section of the chapter, I demonstrate how this issue can be exacerbated in co-assembly when the selection of samples to pool is not performed judiciously. In fact, co-assembly of samples from unrelated individuals generated apparently high-quality MAGs that are, in fact, inter-individual strain chimeras. While intra-individual chimeras still represent a genuine biological species population, inter-subject strain chimeras do not, and utilizing them for downstream analyses can potentially yield spurious results.

Although this limitation may restrict the applicability of co-assembly, as metagenomic studies often lack samples that are as biologically related as those employed in my analysis, such as technical variants, replicates, or negative controls, co-assembly can be effectively applied to longitudinal gut metagenomes. Co-assembling follow-up samples from the same subject produced same species MAGs that were phylogenetically nearly identical to those assembled from individual samples, allowing for the repurposing of publicly available metagenomic data to assemble novel species while minimizing the risk of strain chimeric MAGs.

Overall, the main work of this thesis and other works to which I contributed highlight the pivotal role of metagenomic assembly and the extensive sampling, sequencing and computational efforts that are still needed to bridge the gap in our current representation of the gut microbial dark matter. The optimization of protocols to optimize metagenomic assembly of multiple biologically similar samples can help in the recovery of the genomes of rare and low abundant species that may play crucial roles in human health.

Limitations of this thesis and future perspectives

My thesis aimed to reconstruct the genomic diversity of low-abundance species in the gut microbiome by evaluating and providing guidelines for leveraging multiple metagenomes, in order to obtain a more comprehensive picture of the gut microbiome. While my goal was to provide a comprehensive evaluation of these in the context of gut microbiome, and illustrate their utility in broader microbiome research, it is important to acknowledge that several limitations remain to be addressed and many aspects warrant further investigation.

One of the strongest limitations of co-assembly and co-binning highlighted in this thesis is that its maximal efficacy is achieved when applied to sets of samples that are biologically redundant (e.g: technical or biological replicates), rarely available in most metagenomic studies. Development of computationally efficient assembly and binning tools that are better able to resolve inter-sample strain variability into distinct MAGs could allow us to extend the approaches presented in this thesis also to unrelated samples.

While this study established general minimum coverage thresholds for species MAG assembly of satisfactory quality (Bowers et al. 2017), a further refinement of these thresholds based on genome characteristics could be achieved through the integration of synthetic metagenomes. Synthetic datasets, by allowing precise control over community composition and strain heterogeneity, could help systematically evaluate how specific genomic characteristics (such as G+C content, tetranucleotide frequency, and sequence redundancy) influence these thresholds. An analysis with the spike-in of synthetic reads, generated from references with known genome properties, in the single technical variants could be an interesting way to assess the efficacy of (co)assembly and (co)binning on the recovery of microbial species based on these parameters. However, it's important to note that current sequencing simulation tools struggle to fully model the complexity of real metagenomic data, particularly regarding technical variables like sample storage conditions and DNA extraction methods, which are important aspects of our cohort samples (Milhaven and Pfeifer 2023). Although synthetic datasets offer controlled testing conditions, they would not have substantially enhanced our practical guidelines beyond the insights gained from real metagenomic data.

Another specific aspect where synthetic metagenomes could provide valuable insights is in understanding how combining samples from different subjects affects MAG chimericity. Using synthetic data with known reference genomes as ground truth and precise control over each strain's coverage contribution would allow us to test whether the strain-specific polymorphisms incorporated into shared contigs during assembly are

proportional to the relative abundance of reads from each strain. Such analysis could help quantify how strain mixing in co-assembly relates to input strain proportions.

A significant limitation of this study stems from the computational constraints of assembling ultra-deep metagenomes, which forced us to prioritize software that balanced output quality with computational feasibility and time efficiency. With regards to metagenomic binning, different tools implement diverse algorithms and features that can make them better suited for specific scenarios (e.g., strain heterogeneity, metagenome complexity, availability of multiple samples). According to recent benchmarks, including the Critical Assessment of Metagenome Interpretation (Meyer et al. 2022), tools like CONCOCT, UltraBinner, and MetaBinner show better performances in recovering MAGs from the same assemblies, especially in strain-complex datasets. Moreover, ensemble binners that combine multiple stand-alone binners (e.g., MetaWRAP, DAStool, MetaBinner) typically achieve superior results in both MAG recovery and quality (Meyer et al. 2022; Qiu et al. 2024; Z. Wang et al. 2023). These advantages come with significantly higher computational costs, as most alternatives require at least an order of magnitude more time than MetaBAT2.

In the future, we plan to further explore multi-sample strategies by integrating more sophisticated approaches, such as haplotype-aware assemblers (Vicedomini et al. 2021; X. Kang, Luo, and Schönhuth 2022; Quince et al. 2021) and advanced binning refinement tools that leverage co-assembly DBG graph support for multi-bin assignment of overlapping contigs (Tolstoganov et al. 2022; Mallawaarachchi, Wickramarachchi, and Lin 2021). Additionally, while our current pipeline relies solely on CheckM's completeness and contamination metrics for MAG quality assessment, we aim to implement additional quality measures (Simão et al. 2015; Orakov et al. 2021; Mikheenko, Saveliev, and Gurevich 2016) to better detect potential contamination and chimerism that might currently go unnoticed.

Characterizing the putative novel microbial lineages identified in this study is a critical next step. The newly identified MAGs should be integrated into reference-based methods to determine whether these lineages are present in external metagenomic datasets, particularly for species assembled only in a single combination or from a single subject. Detection of these lineages in multiple independent samples would validate their legitimacy and mitigate concerns regarding co-assembly artifacts. Furthermore, comparative genomic analyses of the novel species' genomes should be conducted to validate whether their phylogenetic novelty also corresponds to functional distinctiveness.

Improving existing algorithms for processing short-read metagenomes remains crucial for repurposing the vast amount of available metagenomic data. However, despite the increasing affordability of deeper sequencing, the intrinsic limits of current short-read

sequencing technologies are unlikely to allow the reconstruction of closed bacterial genomes for all members of a microbial community.

Third-generation sequencing technologies, or long-read (LR) sequencing, have emerged as a solution to the shortcomings of short-read sequencing (SRS). LR-sequencing platforms developed by Oxford Nanopore and PacBio now produce reads in the order of kilobases or megabases with error rates comparable to those of Illumina SRS (Sereika et al. 2022; Hon et al. 2020). These long reads can completely span genomic regions that are challenging to resolve with short-read assembly, such as highly repetitive intra- and inter-genomic regions. This capability improves the contiguity of metagenomic assembly, resolves strain variability, and avoids amplification biases (Liu et al. 2022; C. Y. Kim, Ma, and Lee 2022). LR-sequencing methods have enabled the recovery of complete and fully circularized microbial genomes directly from metagenomes (Moss, Maghini, and Bhatt 2020).

However, achieving sequencing throughputs comparable to SRS remains costly for LR technologies, and high throughputs are necessary to comprehensively cover metagenomic biodiversity. Additionally, LR at the same depth recovers less MAGs compared to SRS. Consequently, short-read metagenome assembly will remain relevant in the next few years.

Hybrid approaches integrating short-read and long-read sequencing have been developed to leverage the advantages of both technologies (low error rates, high throughput). These methods have successfully reconstructed complete genomes from metagenomes and recovered MAGs missed by single-technology approaches (Antipov et al. 2016; Bertrand et al. 2019; Yorki et al. 2023; Gounot et al. 2022). To my knowledge, multi-sample approaches have not yet been evaluated for LR or hybrid metagenomes and could represent an interesting direction for the continuation of this thesis work.

Sequencing low-abundance members of the gut microbiome remains challenging, even for LR sequencing, as genome capture probability depends on relative abundance and sequencing depth. To address this, Oxford Nanopore Technologies has implemented a real-time DNA enrichment technology that discards molecules matching sequences in a reference database, potentially enhancing detection of rare species (Martin et al. 2022).

In conclusion, this thesis work and related contributions underscore the critical role of metagenomic assembly and the substantial efforts still required to characterize the gut microbial dark matter. Rare and low-abundance species represent an important reservoir of genetic diversity in the gut microbiome and potential markers for dysbiosis. Comprehensive knowledge of microbial diversity will enhance our understanding of microbiome-host interactions and enable targeted microbiome-based interventions.

References

Aitchison, John. 1982. *The Statistical Analysis of Compositional Data*.

Almeida, Alexandre, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley, and Robert D. Finn. 2019. “A New Genomic Blueprint of the Human Gut Microbiota.” *Nature* 568 (7753): 499–504.

Almeida, Alexandre, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S. Pollard, et al. 2021. “A Unified Catalog of 204,938 Reference Genomes from the Human Gut Microbiome.” *Nature Biotechnology* 39 (1): 105–14.

Alneberg, Johannes, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z. Ijaz, Leo Lahti, Nicholas J. Loman, Anders F. Andersson, and Christopher Quince. 2014. “Binning Metagenomic Contigs by Coverage and Composition.” *Nature Methods* 11 (11): 1144–46.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10.

Antipov, Dmitry, Anton Korobeynikov, Jeffrey S. McLean, and Pavel A. Pevzner. 2016. “hybridSPAdes: An Algorithm for Hybrid Assembly of Short and Long Reads.” *Bioinformatics* 32 (7): 1009–15.

Anyansi, Christine, Timothy J. Straub, Abigail L. Manson, Ashlee M. Earl, and Thomas Abeel. 2020. “Computational Methods for Strain-Level Microbial Detection in Colony and Metagenome Sequencing Data.” *Frontiers in Microbiology* 11 (August):1925.

Asnicar, Francesco, Sarah E. Berry, Ana M. Valdes, Long H. Nguyen, Gianmarco Piccinno, David A. Drew, Emily Leeming, et al. 2021. “Microbiome Connections with Host Metabolism and Habitual Diet from 1,098 Deeply Phenotyped Individuals.” *Nature Medicine* 27 (2): 321–32.

Asnicar, Francesco, Andrew Maltez Thomas, Francesco Beghini, Claudia Mengoni, Serena Manara, Paolo Manghi, Qiyun Zhu, et al. 2020. “Precise Phylogenetic Analysis of Microbial Isolates and Genomes from Metagenomes Using PhyloPhlAn 3.0.” *Nature Communications* 11 (1): 2500.

Ayling, Martin, Matthew D. Clark, and Richard M. Leggett. 2020. “New Approaches for Metagenome Assembly with Short Reads.” *Briefings in Bioinformatics* 21 (2): 584–94.

Beghini, Francesco, Lauren J. McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, et al. 2021. “Integrating Taxonomic, Functional, and Strain-Level Profiling of Diverse Microbial Communities with bioBakery 3.” *eLife* 10 (May). https://doi.org/10.7554/eLife.65088.

Berg, Gabriele, Daria Rybakova, Doreen Fischer, Tomislav Cernava, Marie-Christine Champomier Vergès, Trevor Charles, Xiaoyulong Chen, et al. 2020. “Microbiome Definition Re-Visited: Old Concepts and New Challenges.” *Microbiome* 8 (1): 1–22.

Bertrand, Denis, Jim Shaw, Manesh Kalathiyappan, Amanda Hui Qi Ng, M. Senthil Kumar, Chenhao Li, Mirta Dvornicic, et al. 2019. “Hybrid Metagenomic Assembly Enables High-Resolution Analysis of Resistance Determinants and Mobile Elements in Human Microbiomes.” *Nature Biotechnology* 37 (8): 937–44.

Bhute, Shrikant S., Saroj S. Ghaskadbi, and Yogesh S. Shouche. 2017. “Rare

Biosphere in Human Gut: A Less Explored Component of Human Gut Microbiota and Its Association with Human Health." In *Mining of Microbial Wealth and MetaGenomics*, edited by Vipin Chandra Kalia, Yogesh Shouche, Hemant J. Purohit, and Praveen Rahi, 133–42. Singapore: Springer Singapore.

Blanco-Míguez, Aitor, Francesco Beghini, Fabio Cumbo, Lauren J. McIver, Kelsey N. Thompson, Moreno Zolfo, Paolo Manghi, et al. 2023. "Extending and Improving Metagenomic Taxonomic Profiling with Uncharacterized Species Using MetaPhlAn 4." *Nature Biotechnology* 41 (11): 1633–44.

Boleij, Annemarie, Elizabeth M. Hechenbleikner, Andrew C. Goodwin, Ruchi Badani, Ellen M. Stein, Mark G. Lazarev, Brandon Ellis, et al. 2015. "The Bacteroides Fragilis Toxin Gene Is Prevalent in the Colon Mucosa of Colorectal Cancer Patients." *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* 60 (2): 208–15.

Boogaart, K. Gerald van den, and Raimon Tolosana-Delgado. 2013. *Analyzing Compositional Data with R*. Springer Science & Business Media.

Bouter, Kristien E., Daniël H. van Raalte, Albert K. Groen, and Max Nieuwdorp. 2017. "Role of the Gut Microbiome in the Pathogenesis of Obesity and Obesity-Related Metabolic Dysfunction." *Gastroenterology* 152 (7): 1671–78.

Bowers, Robert M., Alicia Clum, Hope Tice, Joanne Lim, Kanwar Singh, Doina Ciobanu, Chew Yee Ngan, Jan-Fang Cheng, Susannah G. Tringe, and Tanja Woyke. 2015. "Impact of Library Preparation Protocols and Template Quantity on the Metagenomic Reconstruction of a Mock Microbial Community." *BMC Genomics* 16 (October):856.

Bowers, Robert M., Nikos C. Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T. B. K. Reddy, Frederik Schulz, et al. 2017. "Minimum Information about a Single Amplified Genome (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea." *Nature Biotechnology* 35 (8): 725–31.

Bray, J. Roger, and J. T. Curtis. 1957. "An Ordination of the Upland Forest Communities of Southern Wisconsin." *Ecological Monographs* 27 (4): 325–49.

Castaner, Olga, Albert Goday, Yong-Moon Park, Seung-Hwan Lee, Faidon Magkos, Sue-Anne Toh Ee Shiow, and Helmut Schröder. 2018. "The Gut Microbiome Profile in Obesity: A Systematic Review." *International Journal of Endocrinology* 2018 (March):4095789.

Chaumeil, Pierre-Alain, Aaron J. Mussig, Philip Hugenholtz, and Donovan H. Parks. 2019. "GTDB-Tk: A Toolkit to Classify Genomes with the Genome Taxonomy Database." *Bioinformatics* 36 (6): 1925–27.

———. 2022. "GTDB-Tk v2: Memory Friendly Classification with the Genome Taxonomy Database." *Bioinformatics* 38 (23): 5315–16.

Chen, Lianmin, Daoming Wang, Sanzhima Garmaeva, Alexander Kurilshikov, Arnau Vich Vila, Ranko Gacesa, Trishla Sinha, et al. 2021. "The Long-Term Genetic Stability and Individual Specificity of the Human Gut Microbiome." *Cell* 184 (9): 2302–15.e12.

Chen, Linxing, Karthik Anantharaman, Alon Shaiber, A. Murat Eren, and Jillian F. Banfield. 2020. "Accurate and Complete Genomes from Metagenomes." *Genome Research* 30 (3): 315–33.

Chklovski, Alex, Donovan H. Parks, Ben J. Woodcroft, and Gene W. Tyson. 2023. “CheckM2: A Rapid, Scalable and Accurate Tool for Assessing Microbial Genome Quality Using Machine Learning.” *Nature Methods* 20 (8): 1203–12.

Churcheward, Benjamin, Maxime Millet, Audrey Bihouée, Guillaume Fertin, and Samuel Chaffron. 2022. “MAGNETO: An Automated Workflow for Genome-Resolved Metagenomics.” *mSystems* 7 (4): e0043222.

Cole, James R., Qiong Wang, Jordan A. Fish, Benli Chai, Donna M. McGarrell, Yanni Sun, C. Titus Brown, Andrea Porras-Alfaro, Cheryl R. Kuske, and James M. Tiedje. 2014. “Ribosomal Database Project: Data and Tools for High Throughput rRNA Analysis.” *Nucleic Acids Research* 42 (Database issue): D633–42.

Cryan, John F., Kenneth J. O’Riordan, Kiran Sandhu, Veronica Peterson, and Timothy G. Dinan. 2020. “The Gut Microbiome in Neurological Disorders.” *Lancet Neurology* 19 (2): 179–94.

Curtis, Dirk, Rob, Sahar, Jonathan, and Asif. 2012. “Human Microbiome Project Consortium: Structure, Function and Diversity of the Healthy Human Microbiome.” *Nature*.

Darling, Aaron E., Guillaume Jospin, Eric Lowe, Frederick A. Matsen 4th, Holly M. Bik, and Jonathan A. Eisen. 2014. “PhyloSift: Phylogenetic Analysis of Genomes and Metagenomes.” *PeerJ* 2 (January):e243.

Das, Taraprasad, Rajagopalaboopathi Jayasudha, Samakalyana Chakravarthy, Gumpili Sai Prashanthi, Archana Bhargava, Mudit Tyagi, Padmaja Kumari Rani, Rajeev Reddy Pappuru, Savitri Sharma, and Sisinthy Shivaji. 2021. “Alterations in the Gut Bacterial Microbiome in People with Type 2 Diabetes Mellitus and Diabetic Retinopathy.” *Scientific Reports* 11 (1): 2738.

Davar, Diwakar, Amiran K. Dzutsev, John A. McCulloch, Richard R. Rodrigues, Joe-Marc Chauvin, Robert M. Morrison, Richelle N. Deblasio, et al. 2021. “Fecal Microbiota Transplant Overcomes Resistance to Anti-PD-1 Therapy in Melanoma Patients.” *Science* 371 (6529): 595–602.

De Filippis, F., E. Pasolli, and D. Ercolini. 2020. “Newly Explored Faecalibacterium Diversity Is Connected to Age, Lifestyle, Geography, and Disease.” *Current Biology : CB* 30 (24). https://doi.org/10.1016/j.cub.2020.09.063.

Delgado, Luis Fernando, and Anders F. Andersson. 2022. “Evaluating Metagenomic Assembly Approaches for Biome-Specific Gene Catalogues.” *Microbiome* 10 (1): 72.

Delmont, Tom O., Christopher Quince, Alon Shaiber, Özcan C. Esen, Sonny Tm Lee, Michael S. Rappé, Sandra L. McLellan, Sebastian Lücker, and A. Murat Eren. 2018. “Nitrogen-Fixing Populations of Planctomycetes and Proteobacteria Are Abundant in Surface Ocean Metagenomes.” *Nature Microbiology* 3 (7): 804–13.

Dhakan, D. B., A. Maji, A. K. Sharma, R. Saxena, J. Pulikkan, T. Grace, A. Gomez, J. Scaria, K. R. Amato, and V. K. Sharma. 2019. “The Unique Composition of Indian Gut Microbiome, Gene Catalogue, and Associated Fecal Metabolome Deciphered Using Multi-Omics Approaches.” *GigaScience* 8 (3): giz004.

Diakite, Ami, Grégory Dubourg, Niokhor Dione, Pamela Afouda, Sara Bellali, Issa Isaac Ngom, Camille Valles, Mamadou Lamine Tall, Jean-Christophe Lagier, and Didier Raoult. 2020. “Optimization and Standardization of the Culturomics Technique for Human Microbiome Exploration.” *Scientific Reports* 10.

https://doi.org/10.1038/s41598-020-66738-8.
Dixon, P. 2003. “VEGAN, a Package of R Functions for Community Ecology.” *Journal of Vegetation Science: Official Organ of the International Association for Vegetation Science*. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1654-1103.2003.tb02228.x?casa_token=ltISWuvt2H8AAAAA:1L62-WZEhj5X9M3j-sYKqrppwC9O5CGCVVLhdY7TDYTqcyObg_Ue-Zql167nOlFFLf9FehSGk005Xg.
Eiseman, B., W. Silen, Bascom Gs, and Kauvar Aj. 1958. “Fecal Enema as an Adjunct in the Treatment of Pseudomembranous Enterocolitis.” *Surgery* 44 (5). https://pubmed.ncbi.nlm.nih.gov/13592638/.
Fattizzo, Bruno, Francesca Cavallaro, Francesco Folino, and Wilma Barcellini. 2021. “Recent Insights into the Role of the Microbiome in Malignant and Benign Hematologic Diseases.” *Critical Reviews in Oncology/hematology* 160 (April):103289.
Fierer, Noah. 2017. “Embracing the Unknown: Disentangling the Complexities of the Soil Microbiome.” *Nature Reviews. Microbiology* 15 (10): 579–90.
Finn, Robert D., Jody Clements, and Sean R. Eddy. 2011. “HMMER Web Server: Interactive Sequence Similarity Searching.” *Nucleic Acids Research* 39 (Web Server issue): W29–37.
Flemer, Burkhardt, Denise B. Lynch, Jillian M. R. Brown, Ian B. Jeffery, Feargal J. Ryan, Marcus J. Claesson, Micheal O’Riordain, Fergus Shanahan, and Paul W. O’Toole. 2017. “Tumour-Associated and Non-Tumour-Associated Microbiota in Colorectal Cancer.” *Gut* 66 (4): 633–43.
Forry, Samuel P., Stephanie L. Servetas, Jason G. Kralj, Keng Soh, Michalis Hadjithomas, Raul Cano, Martha Carlin, et al. 2024. “Variability and Bias in Microbiome Metagenomic Sequencing: An Interlaboratory Study Comparing Experimental Protocols.” *Scientific Reports* 14 (1): 9785.
Franzosa, Eric A., Lauren J. McIver, Gholamali Rahnavard, Luke R. Thompson, Melanie Schirmer, George Weingart, Karen Schwarzberg Lipson, et al. 2018. “Species-Level Functional Profiling of Metagenomes and Metatranscriptomes.” *Nature Methods* 15 (11): 962–68.
Fu, Jiongxing, Yan Zheng, Ying Gao, and Wanghong Xu. 2022. “Dietary Fiber Intake and Gut Microbiota in Human Health.” *Microorganisms* 10 (12). https://doi.org/10.3390/microorganisms10122507.
Fujita, Kazutoshi, Makoto Matsushita, Eri Banno, Marco A. De Velasco, Koji Hatano, Norio Nonomura, and Hirotsugu Uemura. 2022. “Gut Microbiome and Prostate Cancer.” *International Journal of Urology: Official Journal of the Japanese Urological Association* 29 (8): 793–98.
Gacesa, R., A. Kurilshikov, A. Vich Vila, T. Sinha, M. A. Y. Klaassen, L. A. Bolte, S. Andreu-Sánchez, et al. 2022. “Environmental Factors Shaping the Gut Microbiome in a Dutch Population.” *Nature* 604 (7907): 732–39.
Ghensi, Paolo, Paolo Manghi, Moreno Zolfo, Federica Armanini, Edoardo Pasolli, Mattia Bolzan, Alberto Bertelle, et al. 2020. “Strong Oral Plaque Microbiome Signatures for Dental Implant Diseases Identified by Strain-Resolution Metagenomics.” *NPJ Biofilms and Microbiomes* 6 (1): 47.
Ghoshal, Uday C., Ratnakar Shukla, Ujjala Ghoshal, Kok-Ann Gwee, Siew C. Ng, and

Eamonn M. M. Quigley. 2012. “The Gut Microbiota and Irritable Bowel Syndrome: Friend or Foe?” *International Journal of Inflammation* 2012. https://doi.org/10.1155/2012/151085.
Ghurye, Jay S., Victoria Cepeda-Espinoza, and Mihai Pop. 2016. “Metagenomic Assembly: Overview, Challenges and Applications.” *The Yale Journal of Biology and Medicine* 89 (3): 353–62.
Goodrich, Julia K., Jillian L. Waters, Angela C. Poole, Jessica L. Sutter, Omry Koren, Ran Blekhman, Michelle Beaumont, et al. 2014. “Human Genetics Shape the Gut Microbiome.” *Cell* 159 (4): 789–99.
Gounot, Jean-Sebastien, Minghao Chia, Denis Bertrand, Woei-Yuh Saw, Aarthi Ravikrishnan, Adrian Low, Yichen Ding, et al. 2022. “Genome-Centric Analysis of Short and Long Read Metagenomes Reveals Uncharacterized Microbiome Diversity in Southeast Asians.” *Nature Communications* 13 (1): 6044.
Greenwald, William W., Niels Klitgord, Victor Seguritan, Shibu Yooseph, J. Craig Venter, Chad Garner, Karen E. Nelson, and Weizhong Li. 2017. “Utilization of Defined Microbial Communities Enables Effective Evaluation of Meta-Genomic Assemblies.” *BMC Genomics* 18 (1): 296.
Grice, Elizabeth A., and Julia A. Segre. 2012. “The Human Microbiome: Our Second Genome.” *Annual Review of Genomics and Human Genetics* 13 (June):151–70.
Guan, Huihui, Yanni Pu, Chenglin Liu, Tao Lou, Shishang Tan, Mengmeng Kong, Zhonghan Sun, et al. 2021. “Comparison of Fecal Collection Methods on Variation in Gut Metagenomics and Untargeted Metabolomics.” *mSphere* 6 (5): e0063621.
Guinane, Caitriona M., and Paul D. Cotter. 2013. “Role of the Gut Microbiota in Health and Chronic Gastrointestinal Disease: Understanding a Hidden Metabolic Organ.” *Therapeutic Advances in Gastroenterology* 6 (4): 295–308.
Gu, Zuguang, Roland Eils, and Matthias Schlesner. 2016. “Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data.” *Bioinformatics* 32 (18): 2847–49.
Hacker, Jörg, and Gabriele Blum-Oehler. 2007. “In Appreciation of Theodor Escherich.” *Nature Reviews. Microbiology* 5 (12): 902–902.
Ha, Connie W. Y., and Suzanne Devkota. 2020. “The New Microbiology: Cultivating the Future of Microbiome-Directed Medicine.” *American Journal of Physiology. Gastrointestinal and Liver Physiology* 319 (6): G639–45.
Hajishengallis, George, Shuang Liang, Mark A. Payne, Ahmed Hashim, Ravi Jotwani, Mehmet A. Eskan, Megan L. McIntosh, et al. 2011. “Low-Abundance Biofilm Species Orchestrates Inflammatory Periodontal Disease through the Commensal Microbiota and Complement.” *Cell Host & Microbe* 10 (5): 497–506.
Halfvarson, Jonas, Colin J. Brislawn, Regina Lamendella, Yoshiki Vázquez-Baeza, William A. Walters, Lisa M. Bramer, Mauro D’Amato, et al. 2017. “Dynamics of the Human Gut Microbiome in Inflammatory Bowel Disease.” *Nature Microbiology* 2 (February):17004.
Hall, Andrew Brantley, Moran Yassour, Jenny Sauk, Ashley Garner, Xiaofang Jiang, Timothy Arthur, Georgia K. Lagoudas, et al. 2017. “A Novel Ruminococcus Gnavus Clade Enriched in Inflammatory Bowel Disease Patients.” *Genome Medicine* 9 (1): 103.
Hallmaier-Wacker, Luisa K., Simone Lueert, Christian Roos, and Sascha Knauf. 2018.

"The Impact of Storage Buffer, DNA Extraction Method, and Polymerase on Microbial Analysis." *Scientific Reports* 8 (1): 1–9.
Handelsman, Jo. 2004. "Metagenomics: Application of Genomics to Uncultured Microorganisms." *Microbiology and Molecular Biology Reviews: MMBR* 68 (4): 669–85.
Handelsman, J., M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman. 1998. "Molecular Biological Access to the Chemistry of Unknown Soil Microbes: A New Frontier for Natural Products." *Chemistry & Biology* 5 (10): R245–49.
Han, Geongoo, Hien Luong, and Shipra Vaishnava. 2022. "Low Abundance Members of the Gut Microbiome Exhibit High Immunogenicity." *Gut Microbes* 14 (1): 2104086.
Han, Geongoo, and Shipra Vaishnava. 2023. "Microbial Underdogs: Exploring the Significance of Low-Abundance Commensals in Host-Microbe Interactions." *Experimental & Molecular Medicine* 55 (12): 2498–2507.
Haryono, Mindia A. S., Ying Yu Law, Krithika Arumugam, Larry C-W Liew, Thi Quynh Ngoc Nguyen, Daniela I. Drautz-Moses, Stephan C. Schuster, Stefan Wuertz, and Rohan B. H. Williams. 2022. "Recovery of High Quality Metagenome-Assembled Genomes From Full-Scale Activated Sludge Microbial Communities in a Tropical Climate Using Longitudinal Metagenome Sampling." *Frontiers in Microbiology* 13 (June):869135.
Heather, James M., and Benjamin Chain. 2016. "The Sequence of Sequencers: The History of Sequencing DNA." *Genomics* 107 (1): 1–8.
Hitch, Thomas C. A., Afrizal Afrizal, Thomas Riedel, Antonios Kioukis, Dirk Haller, Ilias Lagkouvardos, Jörg Overmann, and Thomas Clavel. 2021. "Recent Advances in Culture-Based Gut Microbiome Research." *International Journal of Medical Microbiology: IJMM* 311 (3): 151485.
Hofer, Ursula. 2018. "The Majority Is Uncultured." *Nature Reviews. Microbiology* 16 (12): 716–17.
Hofmeyr, Steven, Rob Egan, Evangelos Georganas, Alex C. Copeland, Robert Riley, Alicia Clum, Emiley Eloe-Fadrosh, et al. 2020. "Terabase-Scale Metagenome Coassembly with MetaHipMer." *Scientific Reports* 10 (1): 10689.
Hong, Sunhee, John Bunge, Chesley Leslin, Sunok Jeon, and Slava S. Epstein. 2009. "Polymerase Chain Reaction Primers Miss Half of rRNA Microbial Diversity." *The ISME Journal* 3 (12): 1365–73.
Hon, Ting, Kristin Mars, Greg Young, Yu-Chih Tsai, Joseph W. Karalius, Jane M. Landolin, Nicholas Maurer, et al. 2020. "Highly Accurate Long-Read HiFi Sequencing Data for Five Complex Genomes." *bioRxiv*. bioRxiv. https://doi.org/10.1101/2020.05.04.077180.
Human Microbiome Project Consortium. 2012. "Structure, Function and Diversity of the Healthy Human Microbiome." *Nature* 486 (7402): 207–14.
Hunt, Martin, Taisei Kikuchi, Mandy Sanders, Chris Newbold, Matthew Berriman, and Thomas D. Otto. 2013. "REAPR: A Universal Tool for Genome Assembly Evaluation." *Genome Biology* 14 (5): R47.
Hyatt, Doug, Gwo-Liang Chen, Philip F. Locascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." *BMC Bioinformatics* 11 (March):119.
Jain, Chirag, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis,

and Srinivas Aluru. 2018. “High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries.” *Nature Communications* 9 (1): 5114.
Jégousse, Clara, Pauline Vannier, René Groben, Frank Oliver Glöckner, and Viggó Marteinsson. 2021. “A Total of 219 Metagenome-Assembled Genomes of Microorganisms from Icelandic Marine Waters.” *PeerJ* 9 (April):e11112.
Jia, Baolei, Xiao Han, Kyung Hyun Kim, and Che Ok Jeon. 2022. “Discovery and Mining of Enzymes from the Human Gut Microbiome.” *Trends in Biotechnology* 40 (2): 240–54.
Jost, Lou. 2007. “Partitioning Diversity into Independent Alpha and Beta Components.” *Ecology* 88 (10): 2427–39.
Kang, Dongwan D., Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. 2019. “MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies.” *PeerJ* 7 (July):e7359.
Kang, Xiongbin, Xiao Luo, and Alexander Schönhuth. 2022. “StrainXpress: Strain Aware Metagenome Assembly from Short Reads.” *Nucleic Acids Research* 50 (17): e101.
Karcher, Nicolai, Eleonora Nigro, Michal Punčochář, Aitor Blanco-Míguez, Matteo Ciciani, Paolo Manghi, Moreno Zolfo, et al. 2021. “Genomic Diversity and Ecology of Human-Associated Akkermansia Species in the Gut Microbiome Revealed by Extensive Metagenomic Assembly.” *Genome Biology* 22 (1): 209.
Karcher, Nicolai, Edoardo Pasolli, Francesco Asnicar, Kun D. Huang, Adrian Tett, Serena Manara, Federica Armanini, et al. 2020. “Analysis of 1321 Eubacterium Rectale Genomes from Metagenomes Uncovers Complex Phylogeographic Population Structure and Subspecies Functional Adaptations.” *Genome Biology* 21 (1): 138.
Kelsen, Judith R., and Gary D. Wu. 2012. “The Gut Microbiota, Environment and Diseases of Modern Society.” *Gut Microbes* 3 (4): 374–82.
Khedher, Mariem Ben, Sophie Alexandra Baron, Toilhata Riziki, Raymond Ruimy, Didier Raoult, Seydina M. Diene, and Jean-Marc Rolain. 2020. “Massive Analysis of 64,628 Bacterial Genomes to Decipher Water Reservoir and Origin of Mobile Colistin Resistance Genes: Is There Another Role for These Enzymes?” *Scientific Reports* 10 (1): 5970.
Kim, Chan Yeong, Muyoung Lee, Sunmo Yang, Kyungnam Kim, Dongeun Yong, Hye Ryun Kim, and Insuk Lee. 2021. “Human Reference Gut Microbiome Catalog Including Newly Assembled Genomes from under-Represented Asian Metagenomes.” *Genome Medicine* 13 (1): 134.
Kim, Chan Yeong, Junyeong Ma, and Insuk Lee. 2022. “HiFi Metagenomic Sequencing Enables Assembly of Accurate and Complete Genomes from Human Gut Microbiota.” *Nature Communications* 13 (1): 6367.
Kim, Nayeon, Junyeong Ma, Wonjong Kim, Jungyeon Kim, Peter Belenky, and Insuk Lee. 2024. “Genome-Resolved Metagenomics: A Game Changer for Microbiome Medicine.” *Experimental & Molecular Medicine*, July. https://doi.org/10.1038/s12276-024-01262-7.
Knox, Natalie C., Jessica D. Forbes, Gary Van Domselaar, and Charles N. Bernstein. 2019. “The Gut Microbiome as a Target for IBD Treatment: Are We There Yet?”

Current Treatment Options in Gastroenterology 17 (1): 115–26.
Kogawa, Masato, Masahito Hosokawa, Yohei Nishikawa, Kazuki Mori, and Haruko Takeyama. 2018. “Obtaining High-Quality Draft Genomes from Uncultured Microbes by Cleaning and Co-Assembly of Single-Cell Amplified Genomes.” *Scientific Reports* 8 (1): 2059.
Kostic, Aleksandar D., Dirk Gevers, Heli Siljander, Tommi Vatanen, Tuulia Hyötyläinen, Anu-Maaria Hämäläinen, Aleksandr Peet, et al. 2015. “The Dynamics of the Human Infant Gut Microbiome in Development and in Progression toward Type 1 Diabetes.” *Cell Host & Microbe* 17 (2): 260–73.
Kozlov, Alexey M., Diego Darriba, Tomáš Flouri, Benoit Morel, and Alexandros Stamatakis. 2019. “RAxML-NG: A Fast, Scalable and User-Friendly Tool for Maximum Likelihood Phylogenetic Inference.” *Bioinformatics* 35 (21): 4453–55.
Krishnan, Smitha, Nicholas Alden, and Kyongbum Lee. 2015. “Pathways and Functions of Gut Microbiota Metabolism Impacting Host Physiology.” *Current Opinion in Biotechnology* 36 (December):137–45.
Lagier, J-C, F. Armougom, M. Million, P. Hugon, I. Pagnier, C. Robert, F. Bittar, et al. 2012. “Microbial Culturomics: Paradigm Shift in the Human Gut Microbiome Study.” *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 18 (12): 1185–93.
Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59.
Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. 2009. “Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.” *Genome Biology* 10 (3): R25.
Le Chatelier, Emmanuelle, Trine Nielsen, Junjie Qin, Edi Prifti, Falk Hildebrand, Gwen Falony, Mathieu Almeida, et al. 2013. “Richness of Human Gut Microbiome Correlates with Metabolic Markers.” *Nature* 500 (7464): 541–46.
Leviatan, Sigal, Saar Shoer, Daphna Rothschild, Maria Gorodetski, and Eran Segal. 2022. “An Expanded Reference Map of the Human Gut Microbiome Reveals Hundreds of Previously Unknown Species.” *Nature Communications* 13 (1): 3863.
Levy, Maayan, Aleksandra A. Kolodziejczyk, Christoph A. Thaiss, and Eran Elinav. 2017. “Dysbiosis and the Immune System.” *Nature Reviews. Immunology* 17 (4): 219–32.
Ley, Ruth E., Daniel A. Peterson, and Jeffrey I. Gordon. 2006. “Ecological and Evolutionary Forces Shaping Microbial Diversity in the Human Intestine.” *Cell* 124 (4): 837–48.
Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. 2015. “MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph.” *Bioinformatics* 31 (10): 1674–76.
Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79.
Líndez, Pau Piera, Joachim Johansen, Svetlana Kutuzova, Arnor Ingi Sigurdsson, Jakob Nybo Nissen, and Simon Rasmussen. 2023. “Adversarial and Variational

Autoencoders Improve Metagenomic Binning." *Communications Biology* 6 (1): 1073.
Liu, Lei, Yu Yang, Yu Deng, and Tong Zhang. 2022. "Nanopore Long-Read-Only Metagenomics Enables Complete and High-Quality Genome Reconstruction from Mock and Complex Metagenomes." *Microbiome* 10 (1): 209.
Li, Zhenyu, Yanxiang Chen, Desheng Mu, Jianying Yuan, Yujian Shi, Hao Zhang, Jun Gan, et al. 2012. "Comparison of the Two Major Classes of Assembly Algorithms: Overlap-Layout-Consensus and de-Bruijn-Graph." *Briefings in Functional Genomics* 11 (1): 25–37.
Li, Zhuoxin, Jie Zhou, Hao Liang, Li Ye, Liuyan Lan, Fang Lu, Qing Wang, et al. 2022. "Differences in Alpha Diversity of Gut Microbiota in Neurological Diseases." *Frontiers in Neuroscience* 16 (June):879318.
Lomsadze, Alexandre, Christophe Bonny, Francesco Strozzi, and Mark Borodovsky. 2021. "GeneMark-HM: Improving Gene Prediction in DNA Sequences of Human Microbiome." *NAR Genomics and Bioinformatics* 3 (2): lqab047.
Lozupone, Catherine A. 2018. "Unraveling Interactions between the Microbiome and the Host Immune System To Decipher Mechanisms of Disease." *mSystems* 3 (2). https://doi.org/10.1128/mSystems.00183-17.
Lozupone, Catherine A., Jesse Stombaugh, Antonio Gonzalez, Gail Ackermann, Doug Wendel, Yoshiki Vázquez-Baeza, Janet K. Jansson, Jeffrey I. Gordon, and Rob Knight. 2013. "Meta-Analyses of Studies of the Human Microbiota." *Genome Research* 23 (10): 1704.
Lozupone, Catherine, Manuel E. Lladser, Dan Knights, Jesse Stombaugh, and Rob Knight. 2011. "UniFrac: An Effective Distance Metric for Microbial Community Comparison." *The ISME Journal* 5 (2): 169–72.
Luckey, T. D. 1972. "Introduction to Intestinal Microecology." *The American Journal of Clinical Nutrition* 25 (12): 1292–94.
Lu, Jennifer, Natalia Rincon, Derrick E. Wood, Florian P. Breitwieser, Christopher Pockrandt, Ben Langmead, Steven L. Salzberg, and Martin Steinegger. 2022. "Metagenome Analysis Using the Kraken Software Suite." *Nature Protocols* 17 (12): 2815–39.
Lund, David, Roelof Dirk Coertze, Marcos Parras-Moltó, Fanny Berglund, Carl-Fredrik Flach, Anna Johnning, D. G. Joakim Larsson, and Erik Kristiansson. 2023. "Extensive Screening Reveals Previously Undiscovered Aminoglycoside Resistance Genes in Human Pathogens." *Communications Biology* 6 (1): 812.
Luo, Chengwei, Despina Tsementzi, Nikos C. Kyrpides, and Konstantinos T. Konstantinidis. 2012. "Individual Genome Assembly from Complex Community Short-Read Metagenomic Datasets." *The ISME Journal* 6 (4): 898–901.
Malard, Florent, Joel Dore, Béatrice Gaugler, and Mohamad Mohty. 2021. "Introduction to Host Microbiome Symbiosis in Health and Disease." *Mucosal Immunology* 14 (3): 547–54.
Mallawaarachchi, Vijini G., Anuradha S. Wickramarachchi, and Yu Lin. 2021. "Improving Metagenomic Binning Results with Overlapped Bins Using Assembly Graphs." *Algorithms for Molecular Biology: AMB* 16 (1): 3.
Manara, Serena, Edoardo Pasolli, Daniela Dolce, Novella Ravenni, Silvia Campana, Federica Armanini, Francesco Asnicar, et al. 2018. "Whole-Genome Epidemiology,

Characterisation, and Phylogenetic Reconstruction of Staphylococcus Aureus Strains in a Paediatric Hospital." *Genome Medicine* 10 (1): 82.
Manichanh, C., L. Rigottier-Gois, E. Bonnaud, K. Gloux, E. Pelletier, L. Frangeul, R. Nalin, et al. 2006. "Reduced Diversity of Faecal Microbiota in Crohn's Disease Revealed by a Metagenomic Approach." *Gut* 55 (2): 205–11.
Manor, Ohad, and Elhanan Borenstein. 2017. "Systematic Characterization and Analysis of the Taxonomic Drivers of Functional Shifts in the Human Microbiome." *Cell Host & Microbe* 21 (2): 254–67.
Marchesi, Julian R., and Jacques Ravel. 2015. "The Vocabulary of Microbiome Research: A Proposal." *Microbiome* 3 (July):31.
Martin, Samuel, Martin Ayling, Livia Patrono, Mario Caccamo, Pablo Murcia, and Richard M. Leggett. 2023. "Capturing Variation in Metagenomic Assembly Graphs with MetaCortex." *Bioinformatics (Oxford, England)* 39 (1): btad020.
Martin, Samuel, Darren Heavens, Yuxuan Lan, Samuel Horsfield, Matthew D. Clark, and Richard M. Leggett. 2022. "Nanopore Adaptive Sampling: A Tool for Enrichment of Low Abundance Species in Metagenomic Samples." *Genome Biology* 23 (1): 11.
Ma, Yue, Zhengyan Guo, Binbin Xia, Yuwei Zhang, Xiaolin Liu, Ying Yu, Na Tang, et al. 2022. "Identification of Antimicrobial Peptides from the Human Gut Microbiome Using Deep Learning." *Nature Biotechnology* 40 (6): 921–31.
McDonald, Daniel, Gail Ackermann, Ludmila Khailova, Christine Baird, Daren Heyland, Rosemary Kozar, Margot Lemieux, et al. 2016. "Extreme Dysbiosis of the Microbiome in Critical Illness." *mSphere* 1 (4). https://doi.org/10.1128/mSphere.00199-16.
Medvedev, Paul, Konstantinos Georgiou, Gene Myers, and Michael Brudno. 2007. "Computability of Models for Sequence Assembly." In *Algorithms in Bioinformatics*, 289–301. Springer Berlin Heidelberg.
Meijenfeldt, F. A. Bastiaan von, Ksenia Arkhipova, Diego D. Cambuy, Felipe H. Coutinho, and Bas E. Dutilh. 2019. "Robust Taxonomic Classification of Uncharted Microbial Sequences and Bins with CAT and BAT." *Genome Biology* 20 (1): 217.
Mendes, Catarina Inês, Pedro Vila-Cerqueira, Yair Motro, Jacob Moran-Gilad, João André Carriço, and Mário Ramirez. 2022. "LMAS: Evaluating Metagenomic Short de Novo Assembly Methods through Defined Communities." *GigaScience* 12 (December). https://doi.org/10.1093/gigascience/giac122.
Menni, Cristina, Jialing Zhu, Caroline I. Le Roy, Olatz Mompeo, Kristin Young, Casey M. Rebholz, Elizabeth Selvin, et al. 2020. "Serum Metabolites Reflecting Gut Microbiome Alpha Diversity Predict Type 2 Diabetes." *Gut Microbes* 11 (6): 1632–42.
Menzel, Peter, Kim Lee Ng, and Anders Krogh. 2016. "Fast and Sensitive Taxonomic Classification for Metagenomics with Kaiju." *Nature Communications* 7 (April):11257.
Meštrović, Tomislav. 2016. "Leveling the Human Microbiota Playing Field: A Rederivation of Gut Bacteria to Human Cells Ratio in Physiological Conditions and in Inflammatory Bowel Disease." *Inflammatory Bowel Diseases* 22 (8): E27.
Meyer, Fernando, Adrian Fritz, Zhi-Luo Deng, David Koslicki, Till Robin Lesker, Alexey Gurevich, Gary Robertson, et al. 2022. "Critical Assessment of Metagenome

Interpretation: The Second Round of Challenges." *Nature Methods* 19 (4): 429–40.
Meyer, Fernando, Till-Robin Lesker, David Koslicki, Adrian Fritz, Alexey Gurevich, Aaron E. Darling, Alexander Sczyrba, Andreas Bremges, and Alice C. McHardy. 2021. "Tutorial: Assessing Metagenomics Software with the CAMI Benchmarking Toolkit." *Nature Protocols* 16 (4): 1785–1801.
Mikheenko, Alla, Vladislav Saveliev, and Alexey Gurevich. 2016. "MetaQUAST: Evaluation of Metagenome Assemblies." *Bioinformatics* 32 (7): 1088–90.
Milanese, Alessio, Daniel R. Mende, Lucas Paoli, Guillem Salazar, Hans-Joachim Ruscheweyh, Miguelangel Cuenca, Pascal Hingamp, et al. 2019. "Microbial Abundance, Activity and Population Genomic Profiling with mOTUs2." *Nature Communications* 10 (1): 1014.
Milhaven, Mark, and Susanne P. Pfeifer. 2023. "Performance Evaluation of Six Popular Short-Read Simulators." *Heredity* 130 (2): 55–63.
Minot, Samuel S., and Amy D. Willis. 2019. "Clustering Co-Abundant Genes Identifies Components of the Gut Microbiome That Are Reproducibly Associated with Colorectal Cancer and Inflammatory Bowel Disease." *Microbiome* 7 (1): 110.
Morowitz, Michael J., Vincent J. Denef, Elizabeth K. Costello, Brian C. Thomas, Valeriy Poroyko, David A. Relman, and Jillian F. Banfield. 2011. "Strain-Resolved Community Genomic Analysis of Gut Microbial Colonization in a Premature Infant." *Proceedings of the National Academy of Sciences of the United States of America* 108 (3): 1128–33.
Moss, Eli L., Dylan G. Maghini, and Ami S. Bhatt. 2020. "Complete, Closed Bacterial Genomes from Microbiomes Using Nanopore Sequencing." *Nature Biotechnology* 38 (6): 701–7.
Mukherjee, Supratim, Rekha Seshadri, Neha J. Varghese, Emiley A. Eloe-Fadrosh, Jan P. Meier-Kolthoff, Markus Göker, R. Cameron Coates, et al. 2017. "1,003 Reference Genomes of Bacterial and Archaeal Isolates Expand Coverage of the Tree of Life." *Nature Biotechnology* 35 (7): 676–83.
Nagpal, Jatin, and John F. Cryan. 2021. "Host Genetics, the Microbiome & Behaviour-a 'Holobiont' Perspective." *Cell Research* 31 (8): 832–33.
Nayfach, Stephen, Zhou Jason Shi, Rekha Seshadri, Katherine S. Pollard, and Nikos C. Kyrpides. 2019. "New Insights from Uncultivated Genomes of the Global Human Gut Microbiome." *Nature* 568 (7753): 505–10.
Nelson, Karen E., George M. Weinstock, Sarah K. Highlander, Kim C. Worley, Heather Huot Creasy, Jennifer Russo Wortman, Douglas B. Rusch, et al. 2010. "A Catalog of Reference Genomes from the Human Microbiome." *Science* 328 (5981): 994–99.
Nielsen, H. Bjørn, Mathieu Almeida, Agnieszka Sierakowska Juncker, Simon Rasmussen, Junhua Li, Shinichi Sunagawa, Damian R. Plichta, et al. 2014. "Identification and Assembly of Genomes and Genetic Elements in Complex Metagenomic Samples without Using Reference Genomes." *Nature Biotechnology* 32 (8): 822–28.
Nissen, Jakob Nybo, Joachim Johansen, Rosa Lundbye Allesøe, Casper Kaae Sønderby, Jose Juan Almagro Armenteros, Christopher Heje Grønbech, Lars Juhl Jensen, et al. 2021. "Improved Metagenome Binning and Assembly Using Deep Variational Autoencoders." *Nature Biotechnology* 39 (5): 555–60.
Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner. 2017.

"metaSPAdes: A New Versatile Metagenomic Assembler." *Genome Research* 27 (5): 824–34.
Ochkalova, Sofia, Ivan Tolstoganov, Alla Lapidus, and Anton Korobeynikov. 2023. "Protocol for Refining Metagenomic Binning with BinSPreader." *STAR Protocols* 4 (3): 102417.
Oh, Tae Gyu, Susy M. Kim, Cyrielle Caussy, Ting Fu, Jian Guo, Shirin Bassirian, Seema Singh, et al. 2020. "A Universal Gut-Microbiome-Derived Signature Predicts Cirrhosis." *Cell Metabolism* 32 (5): 878–88.e6.
O'Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. "Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation." *Nucleic Acids Research* 44 (D1): D733–45.
Olson, Nathan D., Todd J. Treangen, Christopher M. Hill, Victoria Cepeda-Espinoza, Jay Ghurye, Sergey Koren, and Mihai Pop. 2017. "Metagenomic Assembly through the Lens of Validation: Recent Advances in Assessing and Improving the Quality of Genomes Assembled from Metagenomes." *Briefings in Bioinformatics* 20 (4): 1140–50.
Ondov, Brian D., Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. 2016. "Mash: Fast Genome and Metagenome Distance Estimation Using MinHash." *Genome Biology* 17 (1): 132.
Orakov, Askarbek, Anthony Fullam, Luis Pedro Coelho, Supriya Khedkar, Damian Szklarczyk, Daniel R. Mende, Thomas S. B. Schmidt, and Peer Bork. 2021. "GUNC: Detection of Chimerism and Contamination in Prokaryotic Genomes." *Genome Biology* 22 (1): 178.
Oulas, Anastasis, Christina Pavloudi, Paraskevi Polymenakou, Georgios A. Pavlopoulos, Nikolas Papanikolaou, Georgios Kotoulas, Christos Arvanitidis, and Loannis Iliopoulos. 2015. "Metagenomics: Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies." *Bioinformatics and Biology Insights*, May. https://doi.org/10.4137/BBI.S12462.
Palmu, Joonatan, Aaro Salosensaari, Aki S. Havulinna, Susan Cheng, Michael Inouye, Mohit Jain, Rodolfo A. Salido, et al. 2020. "Association Between the Gut Microbiota and Blood Pressure in a Population Cohort of 6953 Individuals." *Journal of the American Heart Association* 9 (15): e016641.
Panek, Marina, Hana Čipčić Paljetak, Anja Barešić, Mihaela Perić, Mario Matijašić, Ivana Lojkić, Darija Vranešić Bender, Željko Krznarić, and Donatella Verbanac. 2018. "Methodology Challenges in Studying Human Gut Microbiota - Effects of Collection, Storage, DNA Extraction and next Generation Sequencing Technologies." *Scientific Reports* 8 (1): 5143.
Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer. 2004. "APE: Analyses of Phylogenetics and Evolution in R Language." *Bioinformatics* 20 (2): 289–90.
Parks, Donovan H., Maria Chuvochina, Pierre-Alain Chaumeil, Christian Rinke, Aaron J. Mussig, and Philip Hugenholtz. 2020. "A Complete Domain-to-Species Taxonomy for Bacteria and Archaea." *Nature Biotechnology* 38 (9): 1079–86.
Parks, Donovan H., Maria Chuvochina, David W. Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2018. "A Standardized Bacterial Taxonomy Based on Genome Phylogeny Substantially Revises the Tree

of Life.” *Nature Biotechnology* 36 (10): 996–1004.
Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. “CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes.” *Genome Research* 25 (7): 1043–55.
Parks, Donovan H., Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz, and Gene W. Tyson. 2017. “Recovery of Nearly 8,000 Metagenome-Assembled Genomes Substantially Expands the Tree of Life.” *Nature Microbiology* 2 (11): 1533–42.
Pascal, Victoria, Marta Pozuelo, Natalia Borruel, Francesc Casellas, David Campos, Alba Santiago, Xavier Martinez, et al. 2017. “A Microbial Signature for Crohn’s Disease.” *Gut* 66 (5): 813–22.
Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, et al. 2019. “Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle.” *Cell* 176 (3): 649–62.e20.
Pasolli, Edoardo, Lucas Schiffer, Paolo Manghi, Audrey Renson, Valerie Obenchain, Duy Tin Truong, Francesco Beghini, et al. 2017. “Accessible, Curated Metagenomic Data through ExperimentHub.” *Nature Methods* 14 (11): 1023–24.
Pedron, Renato, Alfonso Esposito, Irene Bianconi, Edoardo Pasolli, Adrian Tett, Francesco Asnicar, Mario Cristofolini, Nicola Segata, and Olivier Jousson. 2019. “Genomic and Metagenomic Insights into the Microbial Community of a Thermal Spring.” *Microbiome* 7 (1): 8.
Peng, Yu, Henry C. M. Leung, S. M. Yiu, and Francis Y. L. Chin. 2012. “IDBA-UD: A de Novo Assembler for Single-Cell and Metagenomic Sequencing Data with Highly Uneven Depth.” *Bioinformatics* 28 (11): 1420–28.
Peppercorn, M. A., and P. Goldman. 1972. “The Role of Intestinal Bacteria in the Metabolism of Salicylazosulfapyridine.” *The Journal of Pharmacology and Experimental Therapeutics* 181 (3). https://pubmed.ncbi.nlm.nih.gov/4402374/.
Plaza Oñate, Florian, Emmanuelle Le Chatelier, Mathieu Almeida, Alessandra C. L. Cervino, Franck Gauthier, Frédéric Magoulès, S. Dusko Ehrlich, and Matthieu Pichaud. 2019. “MSPminer: Abundance-Based Reconstitution of Microbial Pan-Genomes from Shotgun Metagenomic Data.” *Bioinformatics (Oxford, England)* 35 (9): 1544–52.
Poulsen, Casper Sahl, Rolf Sommer Kaas, Frank M. Aarestrup, and Sünje Johanna Pamp. 2021. “Standard Sample Storage Conditions Have an Impact on Inferred Microbiome Composition and Antimicrobial Resistance Patterns.” *Microbiology Spectrum* 9 (2): e0138721.
Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2009. “FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix.” *Molecular Biology and Evolution* 26 (7): 1641–50.
———. 2010. “FastTree 2--Approximately Maximum-Likelihood Trees for Large Alignments.” *PloS One* 5 (3): e9490.
Probst, Alexander J., Thomas Weinmaier, Todd Z. DeSantis, Jorge W. Santo Domingo, and Nicholas Ashbolt. 2015. “New Perspectives on Microbial Community Distortion after Whole-Genome Amplification.” *PloS One* 10 (5): e0124158.

Pust, Marie-Madlen, and Burkhard Tümmler. 2022. “Bacterial Low-Abundant Taxa Are Key Determinants of a Healthy Airway Metagenome in the Early Years of Human Life.” *Computational and Structural Biotechnology Journal* 20:175–86.
Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, et al. 2010. “A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing.” *Nature* 464 (7285): 59–65.
Qiu, Zhiguang, Li Yuan, Chun-Ang Lian, Bin Lin, Jie Chen, Rong Mu, Xuejiao Qiao, et al. 2024. “BASALT Refines Binning from Metagenomic Data and Increases Resolution of Genome-Resolved Metagenomic Analysis.” *Nature Communications* 15 (1): 2179.
Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2013. “The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools.” *Nucleic Acids Research* 41 (Database issue): D590–96.
Quince, Christopher, Tom O. Delmont, Sébastien Raguideau, Johannes Alneberg, Aaron E. Darling, Gavin Collins, and A. Murat Eren. 2017. “DESMAN: A New Tool for de Novo Extraction of Strains from Metagenomes.” *Genome Biology* 18 (1): 181.
Quince, Christopher, Sergey Nurk, Sebastien Raguideau, Robert James, Orkun S. Soyer, J. Kimberly Summers, Antoine Limasset, A. Murat Eren, Rayan Chikhi, and Aaron E. Darling. 2021. “STRONG: Metagenomics Strain Resolution on Assembly Graphs.” *Genome Biology* 22 (1): 214.
Quince, Christopher, Alan W. Walker, Jared T. Simpson, Nicholas J. Loman, and Nicola Segata. 2017. “Shotgun Metagenomics, from Sampling to Analysis.” *Nature Biotechnology* 35 (9): 833–44.
Ramakodi, Meganathan P. 2021. “Effect of Amplicon Sequencing Depth in Environmental Microbiome Research.” *Current Microbiology* 78 (3): 1026–33.
Razumov, A. S. 1932. “The Direct Method of Calculation of Bacteria in Water: Comparison with the Koch Method.” *Mikrobiologija* 1:131–46.
Rinke, Christian, Patrick Schwientek, Alexander Sczyrba, Natalia N. Ivanova, Iain J. Anderson, Jan-Fang Cheng, Aaron Darling, et al. 2013. “Insights into the Phylogeny and Coding Potential of Microbial Dark Matter.” *Nature* 499 (7459): 431–37.
Rowland, Ian, Glenn Gibson, Almut Heinken, Karen Scott, Jonathan Swann, Ines Thiele, and Kieran Tuohy. 2018. “Gut Microbiota Functions: Metabolism of Nutrients and Other Food Components.” *European Journal of Nutrition* 57 (1): 1–24.
Royalty, Taylor M., and Andrew D. Steen. 2019. “Theoretical and Simulation-Based Investigation of the Relationship between Sequencing Effort, Microbial Community Richness, and Diversity in Binning Metagenome-Assembled Genomes.” *mSystems* 4 (5). https://doi.org/10.1128/mSystems.00384-19.
Ruscheweyh, H. J., A. Milanese, L. Paoli, A. Sintsova, D. R. Mende, G. Zeller, and S. Sunagawa. 2021. “mOTUs: Profiling Taxonomic Composition, Transcriptional Activity and Strain Populations of Microbial Communities.” *Current Protocols* 1 (8). https://doi.org/10.1002/cpz1.218.
Salazar, Vinícius W., Babak Shaban, Maria Del Mar Quiroga, Robert Turnbull, Edoardo Tescari, Vanessa Rossetto Marcelino, Heroen Verbruggen, and Kim-Anh Lê Cao.

2022. “Metaphor-A Workflow for Streamlined Assembly and Binning of Metagenomes.” *GigaScience* 12 (December). https://doi.org/10.1093/gigascience/giad055.
Saw, Jimmy H. W. 2021. “Characterizing the Uncultivated Microbial Minority: Towards Understanding the Roles of the Rare Biosphere in Microbial Communities.” *mSystems* 6 (4): e0077321.
Schmidt, Thomas S. B., Anthony Fullam, Pamela Ferretti, Askarbek Orakov, Oleksandr M. Maistrenko, Hans-Joachim Ruscheweyh, Ivica Letunic, et al. 2024. “SPIRE: A Searchable, Planetary-Scale mIcrobiome REsource.” *Nucleic Acids Research* 52 (D1): D777–83.
Scholz, Matthias, Doyle V. Ward, Edoardo Pasolli, Thomas Tolio, Moreno Zolfo, Francesco Asnicar, Duy Tin Truong, Adrian Tett, Ardythe L. Morrow, and Nicola Segata. 2016. “Strain-Level Microbial Epidemiology and Population Genomics from Shotgun Metagenomics.” *Nature Methods* 13 (5): 435–38.
Schwengers, Oliver, Lukas Jelonek, Marius Alfred Dieckmann, Sebastian Beyvers, Jochen Blom, and Alexander Goesmann. 2021. “Bakta: Rapid and Standardized Annotation of Bacterial Genomes via Alignment-Free Sequence Identification.” *Microbial Genomics* 7 (11). https://doi.org/10.1099/mgen.0.000685.
Sczyrba, Alexander, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, et al. 2017. “Critical Assessment of Metagenome Interpretation-a Benchmark of Metagenomics Software.” *Nature Methods* 14 (11): 1063–71.
Seemann, Torsten. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” *Bioinformatics* 30 (14): 2068–69.
Segata, Nicola, Levi Waldron, Annalisa Ballarini, Vagheesh Narasimhan, Olivier Jousson, and Curtis Huttenhower. 2012. “Metagenomic Microbial Community Profiling Using Unique Clade-Specific Marker Genes.” *Nature Methods* 9 (8): 811–14.
Sereika, Mantas, Rasmus Hansen Kirkegaard, Søren Michael Karst, Thomas Yssing Michaelsen, Emil Aarre Sørensen, Rasmus Dam Wollenberg, and Mads Albertsen. 2022. “Oxford Nanopore R10.4 Long-Read Sequencing Enables the Generation of near-Finished Bacterial Genomes from Pure Cultures and Metagenomes without Short-Read or Reference Polishing.” *Nature Methods* 19 (7): 823–26.
Sieber, Christian M. K., Alexander J. Probst, Allison Sharrar, Brian C. Thomas, Matthias Hess, Susannah G. Tringe, and Jillian F. Banfield. 2018. “Recovery of Genomes from Metagenomes via a Dereplication, Aggregation and Scoring Strategy.” *Nature Microbiology* 3 (7): 836–43.
Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. “BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs.” *Bioinformatics* 31 (19): 3210–12.
Staley, J. T., and A. Konopka. 1985. “Measurement of in Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats.” *Annual Review of Microbiology* 39:321–46.
Stewart, Robert D., Marc D. Auffret, Amanda Warr, Alan W. Walker, Rainer Roehe, and Mick Watson. 2019. “Compendium of 4,941 Rumen Metagenome-Assembled

Genomes for Rumen Microbiome Biology and Enzyme Discovery." *Nature Biotechnology* 37 (8): 953–61.
Stewart, Robert D., Marc D. Auffret, Amanda Warr, Andrew H. Wiser, Maximilian O. Press, Kyle W. Langford, Ivan Liachko, et al. 2018. "Assembly of 913 Microbial Genomes from Metagenomic Sequencing of the Cow Rumen." *Nature Communications* 9 (1): 870.
Sun, Jiaao, Feng Chen, and Guangzhen Wu. 2023. "Potential Effects of Gut Microbiota on Host Cancers: Focus on Immunity, DNA Damage, Cellular Pathways, and Anticancer Therapy." *The ISME Journal* 17 (10): 1535–51.
Szóstak, Natalia, Agata Szymanek, Jan Havránek, Katarzyna Tomela, Magdalena Rakoczy, Anna Samelak-Czajka, Marcin Schmidt, et al. 2022. "The Standardisation of the Approach to Metagenomic Human Gut Analysis: From Sample Collection to Microbiome Profiling." *Scientific Reports* 12 (1): 8470.
Tatusova, Tatiana, Michael DiCuccio, Azat Badretdin, Vyacheslav Chetvernin, Eric P. Nawrocki, Leonid Zaslavsky, Alexandre Lomsadze, Kim D. Pruitt, Mark Borodovsky, and James Ostell. 2016. "NCBI Prokaryotic Genome Annotation Pipeline." *Nucleic Acids Research* 44 (14): 6614–24.
Tett, Adrian, Kun D. Huang, Francesco Asnicar, Hannah Fehlner-Peach, Edoardo Pasolli, Nicolai Karcher, Federica Armanini, et al. 2019. "The Prevotella Copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations." *Cell Host & Microbe* 26 (5): 666–79.e7.
Thaiss, Christoph A., Niv Zmora, Maayan Levy, and Eran Elinav. 2016. "The Microbiome and Innate Immunity." *Nature* 535 (7610): 65–74.
Thomas, Andrew Maltez, Francesco Asnicar, Guido Kroemer, and Nicola Segata. 2021. "Genes Encoding Microbial Acyl Coenzyme A Binding Protein/Diazepam-Binding Inhibitor Orthologs Are Rare in the Human Gut Microbiome and Show No Links to Obesity." *Applied and Environmental Microbiology* 87 (12): e0047121.
Thomas, Andrew Maltez, Paolo Manghi, Francesco Asnicar, Edoardo Pasolli, Federica Armanini, Moreno Zolfo, Francesco Beghini, et al. 2019. "Metagenomic Analysis of Colorectal Cancer Datasets Identifies Cross-Cohort Microbial Diagnostic Signatures and a Link with Choline Degradation." *Nature Medicine* 25 (4): 667–78.
Thomas, Andrew Maltez, and Nicola Segata. 2019. "Multiple Levels of the Unknown in Microbiome Research." *BMC Biology* 17 (1): 48.
Tolstoganov, Ivan, Yuri Kamenev, Roman Kruglikov, Sofia Ochkalova, and Anton Korobeynikov. 2022. "BinSPreader: Refine Binning Results for Fuller MAG Reconstruction." *iScience* 25 (8): 104770.
Tremaroli, Valentina, and Fredrik Bäckhed. 2012. "Functional Interactions between the Gut Microbiota and Host Metabolism." *Nature* 489 (7415): 242–49.
Tremblay, Julien, Lars Schreiber, and Charles W. Greer. 2022. "High-Resolution Shotgun Metagenomics: The More Data, the Better?" *Briefings in Bioinformatics* 23 (6): bbac443.
Truong, Duy Tin, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. 2015. "MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling." *Nature Methods* 12 (10): 902–3.
Truong, Duy Tin, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, and Nicola Segata.

2017. “Microbial Strain-Level Population Structure and Genetic Diversity from Metagenomes.” *Genome Research* 27 (4): 626–38.
Turnbaugh, Peter J., Micah Hamady, Tanya Yatsunenko, Brandi L. Cantarel, Alexis Duncan, Ruth E. Ley, Mitchell L. Sogin, et al. 2009. “A Core Gut Microbiome in Obese and Lean Twins.” *Nature* 457 (7228): 480–84.
Tyson, Gene W., Jarrod Chapman, Philip Hugenholtz, Eric E. Allen, Rachna J. Ram, Paul M. Richardson, Victor V. Solovyev, Edward M. Rubin, Daniel S. Rokhsar, and Jillian F. Banfield. 2004. “Community Structure and Metabolism through Reconstruction of Microbial Genomes from the Environment.” *Nature* 428 (6978): 37–43.
Uritskiy, Gherman V., Jocelyne DiRuggiero, and James Taylor. 2018. “MetaWRAP-a Flexible Pipeline for Genome-Resolved Metagenomic Data Analysis.” *Microbiome* 6 (1): 158.
Valles-Colomer, Mireia, Aitor Blanco-Míguez, Paolo Manghi, Francesco Asnicar, Leonard Dubois, Davide Golzato, Federica Armanini, et al. 2023. “The Person-to-Person Transmission Landscape of the Gut and Oral Microbiomes.” *Nature* 614 (7946): 125–35.
Van Rossum, Thea, Pamela Ferretti, Oleksandr M. Maistrenko, and Peer Bork. 2020. “Diversity within Species: Interpreting Strains in Microbiomes.” *Nature Reviews. Microbiology* 18 (9): 491–506.
Vicedomini, Riccardo, Christopher Quince, Aaron E. Darling, and Rayan Chikhi. 2021. “Strainberry: Automated Strain Separation in Low-Complexity Metagenomes Using Long Reads.” *Nature Communications* 12 (1): 4485.
Vosloo, Solize, Linxuan Huo, Christopher L. Anderson, Zihan Dai, Maria Sevillano, and Ameet Pinto. 2021. “Evaluating de Novo Assembly and Binning Strategies for Time Series Drinking Water Metagenomes.” *Microbiology Spectrum* 9 (3): e0143421.
Wang, Yuwei, Jin Zhao, Yunlong Qin, Zixian Yu, Yumeng Zhang, Xiaoxuan Ning, and Shiren Sun. 2022. “The Specific Alteration of Gut Microbiota in Diabetic Kidney Diseases-A Systematic Review and Meta-Analysis.” *Frontiers in Immunology* 13 (June):908219.
Wang, Ziye, Pingqin Huang, Ronghui You, Fengzhu Sun, and Shanfeng Zhu. 2023. “MetaBinner: A High-Performance and Stand-Alone Ensemble Binning Method to Recover Individual Genomes from Complex Microbial Communities.” *Genome Biology* 24 (1): 1.
Wesolowska-Andersen, Agata, Martin Iain Bahl, Vera Carvalho, Karsten Kristiansen, Thomas Sicheritz-Pontén, Ramneek Gupta, and Tine Rask Licht. 2014. “Choice of Bacterial DNA Extraction Method from Fecal Material Influences Community Structure as Evaluated by Metagenomic Analysis.” *Microbiome* 2 (June):19.
Wilke, Claus O. 2024. “Ridgeline Plots in ggplot2.” 2024. https://wilkelab.org/ggridges/index.html.
Woese, C. R., and G. E. Fox. 1977. “Phylogenetic Structure of the Prokaryotic Domain: The Primary Kingdoms.” *Proceedings of the National Academy of Sciences of the United States of America* 74 (11): 5088–90.
Wood, Derrick E., Jennifer Lu, and Ben Langmead. 2019. “Improved Metagenomic Analysis with Kraken 2.” *Genome Biology* 20 (1): 1–13.
Xu, Meng-Qi, Fei Pan, Li-Hua Peng, and Yun-Sheng Yang. 2024. “Advances in the

Isolation, Cultivation, and Identification of Gut Microbes." *Military Medical Research* 11 (1): 34.
Ye, Lin. 2017. *Exploring Microbial Community Structures and Functions of Activated Sludge by High-Throughput Sequencing*.
Yorki, Sosie, Terrance Shea, Christina A. Cuomo, Bruce J. Walker, Regina C. LaRocque, Abigail L. Manson, Ashlee M. Earl, and Colin J. Worby. 2023. "Comparison of Long- and Short-Read Metagenomic Assembly for Low-Abundance Species and Resistance Genes." *Briefings in Bioinformatics* 24 (2). https://doi.org/10.1093/bib/bbad050.
Yuan, Cheng, Jikai Lei, James Cole, and Yanni Sun. 2015. "Reconstructing 16S rRNA Genes in Metagenomic Data." *Bioinformatics (Oxford, England)* 31 (12): i35–43.
Zaheer, Rahat, Noelle Noyes, Rodrigo Ortega Polo, Shaun R. Cook, Eric Marinier, Gary Van Domselaar, Keith E. Belk, Paul S. Morley, and Tim A. McAllister. 2018. "Impact of Sequencing Depth on the Characterization of the Microbiome and Resistome." *Scientific Reports* 8 (1): 5890.
Zepeda-Rivera, Martha, Samuel S. Minot, Heather Bouzek, Hanrui Wu, Aitor Blanco-Míguez, Paolo Manghi, Dakota S. Jones, et al. 2024. "A Distinct Fusobacterium Nucleatum Clade Dominates the Colorectal Cancer Niche." *Nature* 628 (8007): 424–32.
Zhang, Zhenyan, Qi Zhang, Tingzhang Wang, Nuohan Xu, Tao Lu, Wenjie Hong, Josep Penuelas, et al. 2022. "Assessment of Global Health Risk of Antibiotic Resistance Genes." *Nature Communications* 13 (1): 1553.
Zha, Yuguo, Hui Chong, Pengshuo Yang, and Kang Ning. 2022. "Microbial Dark Matter: From Discovery to Applications." *Genomics, Proteomics & Bioinformatics* 20 (5): 867–81.
Zheng, Danping, Timur Liwinski, and Eran Elinav. 2020. "Interaction between Microbiota and Immunity in Health and Disease." *Cell Research* 30 (6): 492–506.
Zhi, Cuiting, Jingqing Huang, Jin Wang, Hua Cao, Yan Bai, Jiao Guo, and Zhengquan Su. 2019. "Connection between Gut Microbiome and the Development of Obesity." *European Journal of Clinical Microbiology & Infectious Diseases: Official Publication of the European Society of Clinical Microbiology* 38 (11): 1987–98.
Zhu, Huiyuan, Man Li, Dexi Bi, Huiqiong Yang, Yaohui Gao, Feifei Song, Jiayi Zheng, et al. 2024. "Fusobacterium Nucleatum Promotes Tumor Progression in KRAS p.G12D-Mutant Colorectal Cancer by Binding to DHX15." *Nature Communications* 15 (1): 1688.
Zysset-Burri, Denise C., Irene Keller, Lieselotte E. Berger, Peter J. Neyer, Christian Steuer, Sebastian Wolf, and Martin S. Zinkernagel. 2019. "Retinal Artery Occlusion Is Associated with Compositional and Functional Shifts in the Gut Microbiome and Altered Trimethylamine-N-Oxide Levels." *Scientific Reports* 9 (1): 15303.

List of abbreviations

MAG: Metagenome-assembled genome
NGS: Next-generation sequencing
WGS: Whole-genome Shotgun Sequencing
GI: Gastrointestinal Tract
IBD: Inflammatory Bowel Disease
NAFLD: Non-Alcoholic Fatty Liver Disease
ORF: Open Reading Frame
rRNA: ribosomal RiboNucleic Acid
HGT: Horizontal Gene Transfer
PCR: Polymerase Chain Reaction
LCA: (Lowest|Last) Common Ancestor
SGB: Species-level Genome Bin
OLC: Overlap-Layout-Consensus
DBG: DeBruijn Graph
SNP: Single-Nucleotide Polymorphism
PCA: Principal Component Analysis
TNF: TetraNucleotide Frequency
SCG: Single Copy Gene
MSA: Multiple Sequence Alignment
ANI: Average Nucleotide Identity
CAZy: Carbohydrate Active Enzymes
UMAP: Uniform Manifold Approximation and Projection
nMDS: Non-metric multidimensional scaling
tSNE: t-distributed stochastic neighbor embedding
SSM: Single-Subject Metagenome
MSM: Multi-Subject Metagenome
STM: Single-Timepoint Metagenome
MTM: Multi-Timepoint Metagenome
SRS: Short-Read Sequencing
LR: Long-read Sequencing
SAG: Single Amplified Genome
FC: Fold-Change