

UNIVERSIDAD DE CANTABRIA

PROGRAMA DE DOCTORADO EN
BIOLOGÍA MOLECULAR Y BIOMEDICINA



Tesis doctoral

Análisis genómico dirigido en microbiomas

PhD Thesis

Targeted genomic analyses in microbiomes

Autor:

Juan Manuel Medina Méndez

Directores:

Javier Crespo García

Fernando de la Cruz Calahorra

This work has been carried out at “Instituto de Investigación Valdecilla” and “Instituto de Biomedicina y Biotecnología de Cantabria”, funded by the following projects:

- Dianas terapéuticas y biomarcadores para la medicina de precisión en MAFLD (PreMed-MAFLD). Project PMP21/00112. Funded by Instituto de Salud Carlos III (ISCIII). 2022-2025.
- De la medicina reactiva a la predictiva en NAFLD: Identificación de una nueva firma de progresión de NAFLD inicial. Project PI22/01853. Funded by Instituto de Salud Carlos III (ISCIII). 2023-2025.
- MAPMAR: Marine plasmids driving the spread of antibiotic resistances. Project PCI2021-121978. Funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGeneration EU/PRTR. 2021-2024.
- Desarrollo de herramientas para intervenciones en distintos microbiomas. Project 55-P221-640.04. 2023-2025.

Agradecimientos

TABLE OF CONTENTS

Abstract.....	13
Resumen	16
Graphical abstract	19
I. INTRODUCTION	21
1. Metagenomics in microbial ecology	23
2. Metataxonomics.....	25
3. Metagenomics.....	28
A) Levels of metagenomic analysis	28
B) Taxonomic profiling from reads.....	31
C) Metagenomic assembly and ORF prediction.....	31
D) Integrated pipelines for whole-metagenome analysis.....	34
E) Advantages and limitations	36
4. Targeted microbial enrichment strategies	39
5. Quantitative metagenomics.....	42
II. OBJECTIVES	50
1. General hypothesis.....	52
2. Objectives.....	52
CHAPTER I	55
1. Introduction	57
2. Materials and methods	60
A) Downloading of metagenomic sequencing data.....	60
B) Description of MASLD patient cohorts.....	61
C) Quality control of metagenomic reads.....	62
D) Taxonomic signatures associated with MASLD	62
E) Isolation of target gene families in the GM	62
F) Analysis of gene abundance by read-based quantification.....	63
G) Whole metagenomic analysis.....	64
H) Presence of candidate genes in human GM genomes and plasmids	64

I) <i>R packages and statistical analysis</i>	65
3. Results.....	65
A) <i>Agathobacter rectalis is consistently depleted in MASLD</i>	65
B) <i>Butyrate-producing genes are depleted in MASLD</i>	70
C) <i>SCA-producing genes are increased in MASLD</i>	73
D) <i>Genes involved in methane production are decreased in MASLD, whereas tor operons driving TMA accumulation are elevated</i>	77
E) <i>Candidate metabolic genes are accessory in the GM</i>	81
4. Discussion	85
5. Conclusions	91
6. Supplementary material	92
CHAPTER II	95
1. Introduction	97
2. Materials and methods	100
A) <i>Metagenome-assembled genomes</i>	100
B) <i>Metagenomic sequencing libraries</i>	101
C) <i>Quality control of metagenomic reads and de novo assembly</i>	102
D) <i>Relaxase and USCG prediction in MAGs and contigs</i>	102
E) <i>Validation of the ORF quantification method</i>	103
F) <i>Relaxase clustering and phylogeny</i>	104
G) <i>Antibiotic-resistance gene prediction</i>	105
H) <i>Detection of plasmid-specific sequences</i>	105
I) <i>Detection of prevalent marine plasmids</i>	105
J) <i>R packages and statistical analysis</i>	106
3. Results.....	106
A) <i>Relaxases are infrequent in marine bacterial genomes</i>	106
B) <i>Relaxases are scarce in aquatic ecosystems</i>	109
C) <i>Oceanic relaxases are phylogenetically diverse</i>	113
D) <i>Antibiotic resistance genes are depleted in the ocean</i>	114
E) <i>Genomic and metagenomic ORF abundances are correlated</i>	117
F) <i>Plasmid-specific sequences are limited in the sea</i>	119
4. Discussion	120

5. Conclusions.....	125
6. Supplementary material	126
V. GLOBAL DISCUSSION	136
VI. GLOBAL CONCLUSIONS	144
BIBLIOGRAPHY	147
List of abbreviations	177
List of figures	180
Supplementary figures	183
APPENDIX	185

Abstract

Metagenomes represent the complete set of genes and genomes present in microbial communities, which are directly sequenced from environmental samples to study their associated microbiomes. This culture-independent approach, known as metagenomics, has dramatically expanded our understanding of microbial diversity, particularly in ecosystems beyond the human microbiome, where most species remain uncultured and hence, functionally uncharacterized. Although taxonomic profiling based on marker genes such as those encoding 16S ribosomal RNA provides valuable insights into bacterial community composition, it lacks the resolution to capture intra-genomic variation and cannot reliably distinguish closely related strains whose functional differences may be critical for microbial activity, host interactions, or ecosystem functioning.

To address these limitations, gene-centric profiling shifts the focus from taxonomic groups to specific gene families with well-defined biological functions, enabling the analysis of molecular traits directly linked to microbial activity and phenotype. Nevertheless, metagenomic quantification remains methodologically challenging due to both the compositional nature of sequencing data and the absence of consensus on standardized metrics for estimating gene abundances. While advances in targeted microbial enrichment have improved the recovery and sequencing of specific genes and mobile genetic elements (MGEs), a unified framework for gene-level quantification across metagenomic datasets remains elusive.

This thesis addresses these challenges by applying a gene-centric strategy to large-scale metagenomic collections, focusing on gene families involved in microbial metabolism and plasmid-mediated horizontal gene transfer (HGT). The central aim is to uncover biologically meaningful patterns related to host-associated disease and environmental antibiotic resistance, patterns that would be overlooked by taxonomy-based approaches alone.

The thesis is structured around two distinct applications of this methodology. In Chapter I, 554 fecal metagenomes from three independent patient cohorts were analyzed to identify functional signatures associated with metabolic dysfunction-associated steatotic liver disease (MASLD). Over 50 target gene families involved in the microbial production of butyrate,

methane, trimethylamine (TMA), and short-chain alcohols (SCAs) were curated and quantified by aligning metagenomic reads against them. In Chapter II, nearly 1000 metagenomes from diverse aquatic and terrestrial ecosystems were used to assess plasmid prevalence by *de novo* assembling reads into contigs, predicting open reading frames, and identifying relaxase (RLX) and antibiotic-resistance genes (ARGs) using searches based on hidden Markov models. These elements were quantified as robust markers of plasmid mobility across biomes.

Results demonstrate that MASLD is marked by a consistent depletion of genes involved in butyrate and methane production, along with an enrichment of TMA- and SCA-producing enzymatic genes. These functional shifts reflect a reprogramming of gut microbial (GM) metabolism, largely driven by accessory and plasmid-encoded genes. In environmental samples, RLXs and ARGs are widespread but exhibit marked lower prevalence in oceanic, riverine, and soil ecosystems compared to the human GM and sewage. RLX gene abundance was up to two orders of magnitude higher in GM and sewage metagenomes, suggesting an elevated potential for HGT through conjugative plasmids in anthropogenically impacted biomes. Sewage samples also harbor ARG concentrations up to four times higher than those found in natural environments, underscoring their role as major reservoirs of resistance dissemination.

In conclusion, these findings underscore the power of targeted, gene-level metagenomic profiling to detect functionally relevant microbial signatures that remain undetected by taxonomy-based methods. By combining gene abundance profiling with metabolic pathway analysis and MGE tracking, this work provides a cohesive framework to elucidate microbial contributions to digestive pathologies and to monitor the spread of antibiotic resistance in natural and anthropic environments, by identifying clinically and ecologically relevant genetic biomarkers.

Resumen

Los metagenomas representan el conjunto de genes y genomas presentes en las comunidades microbianas, y pueden secuenciarse directamente a partir de muestras ambientales para estudiar sus microbiomas asociados. Esta tecnología independiente de cultivo, denominada metagenómica, ha ampliado en gran medida nuestro conocimiento sobre la diversidad microbiana, especialmente en ecosistemas distintos al microbioma humano, donde la mayoría de las especies no han sido cultivadas ni estudiadas funcionalmente. Aunque la caracterización taxonómica basada en genes marcadores como el gen codificante del RNA ribosomal 16S proporciona información clave sobre la composición de una comunidad bacteriana, carece de la resolución necesaria para detectar variaciones intra-genómicas, y por tanto no permite distinguir cepas estrechamente relacionadas, cuyas diferencias funcionales pueden ser determinantes para su actividad, su interacción con el hospedador o para el funcionamiento del ecosistema.

Para superar estas limitaciones, el análisis geno-céntrico cambia el enfoque en los grupos taxonómicos hacia familias génicas específicas con funciones biológicas bien definidas, lo que permite estudiar rasgos moleculares directamente vinculados con la actividad microbiana y el fenotipo. No obstante, la cuantificación en metagenómica sigue siendo un reto metodológico debido tanto a la naturaleza composicional de los datos de secuenciación como a la falta de consenso sobre métricas estandarizadas para estimar la abundancia génica. Si bien los avances en estrategias de enriquecimiento microbiano dirigido han mejorado la recuperación y secuenciación de genes específicos y elementos genéticos móviles (MGEs), todavía no se dispone de un marco unificado para su cuantificación a gran escala.

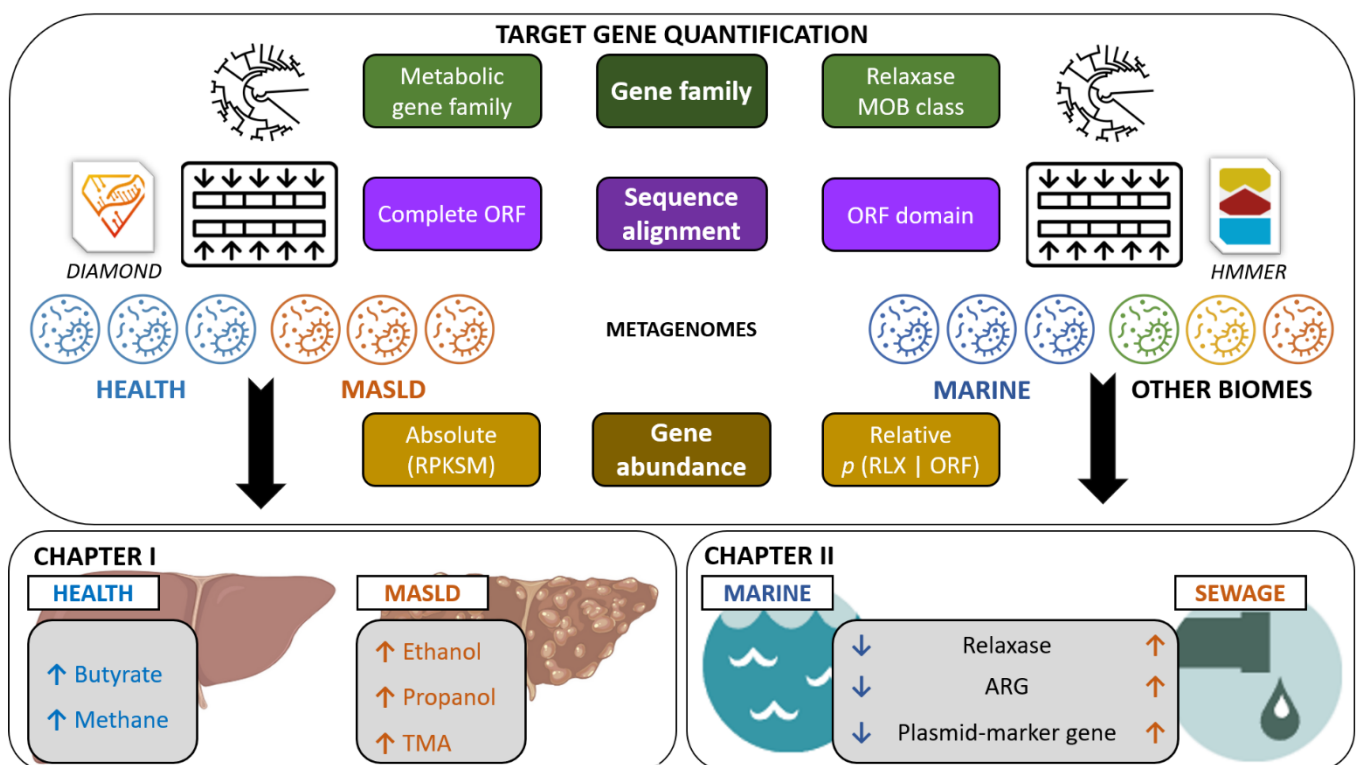
Esta tesis aborda dichos desafíos mediante una estrategia geno-céntrica aplicada a amplios conjuntos de datos metagenómicos, centrándose en familias génicas implicadas en el metabolismo microbiano y en la transferencia horizontal de genes (HGT) mediada por plásmidos. El objetivo central es identificar patrones biológicamente relevantes relacionados con enfermedades humanas y con la resistencia ambiental a antibióticos, señales que pasarían desapercibidas usando enfoques puramente metataxonómicos.

La tesis se estructura en dos aplicaciones distintas de esta metodología. En el Capítulo I se analizaron 554 metagenomas fecales de tres cohortes independientes de pacientes con el objetivo de identificar señales funcionales ligadas a la enfermedad hepática metabólica (MASLD). Más de 50 familias génicas involucradas en la producción microbiana de butirato, metano, trimetilamina (TMA) y alcoholes de cadena corta (SCAs) fueron seleccionadas y cuantificadas mediante el alineamiento de lecturas metagenómicas. En el Capítulo II se utilizaron cerca de 1000 metagenomas procedentes de diversos ecosistemas acuáticos y terrestres para evaluar la prevalencia de plásmidos mediante el ensamblaje *de novo* de lecturas en contigs, la predicción de marcos de lectura abiertos, y la identificación de genes codificantes de relaxasas (RLXs) y de resistencia a antibióticos (ARGs) mediante búsquedas basadas en modelos ocultos de Markov. Estos elementos fueron cuantificados como marcadores robustos de movilidad plasmídica a través de distintos biomas.

Los resultados muestran que MASLD se caracteriza por una disminución consistente de genes implicados en la producción de butirato y metano, y por un aumento de genes enzimáticos productores de TMA y SCAs. Estos cambios funcionales reflejan una reprogramación del metabolismo del microbioma intestinal (GM), determinada en gran medida por genes accesorios y codificados en plásmidos. En muestras ambientales, los genes codificantes de RLXs y los ARGs están ampliamente distribuidos, pero presentan una prevalencia mucho menor en ecosistemas oceánicos, fluviales y edáficos en comparación con el GM y las aguas residuales. La abundancia de genes RLX fue hasta dos órdenes de magnitud mayor en el GM y aguas residuales, lo que sugiere un mayor potencial de HGT mediada por plásmidos conjugativos en biomas antropizados. Asimismo, las aguas residuales albergan concentraciones de ARGs hasta cuatro veces superiores a las de los ambientes naturales, subrayando su papel como reservorios clave en la diseminación de resistencias.

En conjunto, estos hallazgos subrayan el poder del análisis metagenómico dirigido a nivel génico para detectar señales funcionales relevantes que escapen a los métodos basados en taxonomía. Integrando la abundancia génica con el análisis de rutas metabólicas y la detección de MGEs, esta tesis doctoral establece un marco unificado para comprender el papel microbiano en enfermedades digestivas, monitorizar la propagación de resistencias en entornos naturales y antropizados, e identificar biomarcadores genéticos con relevancia clínica y ecológica.

Graphical abstract



I. INTRODUCTION

1. Metagenomics in microbial ecology

In 1998, the term metagenome was first introduced to describe “the collective genomes and the biosynthetic machinery of soil microflora”¹. In this pioneering study, Handelsman and colleagues isolated and cloned DNA directly from a soil sample, allowing them to access the genetic content of uncultured microorganisms. This approach marked the first time the metagenome was treated as a distinct genomic unit, enabling the exploration of microbial diversity beyond cultured species. By directly extracting genetic material from environmental samples, it provided new insights into microbial ecology and biological functions, laying the foundation for modern metagenomic research. Over 25 years later, the term metagenome continues to refer to the collection of genomes and genes from a microbial community^{2,3}.

Many microorganisms cannot be cultured under standard laboratory conditions due to several factors that prevent their isolation and genomic analysis. In some cases, this is due to their reliance on complex metabolic interactions within native communities, where they function as interconnected networks of cells exchanging nutrients and biochemical functions⁴. Other reasons include the difficulty in identifying the specific growth conditions or substrates required for many microorganisms, as they depend on unique environmental or growth factors that are challenging to replicate in the lab⁵. Additionally, microbial populations can include dormant persister cells that survive in low-nutrient conditions, such as those found in the deep biosphere, where microbial activity occurs at much slower rates than at surface levels⁶. Moreover, DNA can remain in the environment even after cell death, meaning that sequencing data may not always reflect the active microbial population⁷.

These limitations have traditionally hindered our understanding of the microbial world, despite its crucial role in the maintenance of ecosystems and direct effects on human health⁸. Metagenomics offers a significant advantage over traditional culture-based methods by enabling the study of these uncultured microorganisms, which represent a large and often overlooked portion of microbial diversity^{9,10}. Recent work based on universal single-copy genes (USCG) shows that the majority of bacterial sequences detected in environmental samples belong to from phylogenetically novel, uncultured groups, with human-associated environments being a notable exception. In contrast, most biomes, such as terrestrial and aquatic habitats, are dominated by yet-uncultured genera¹⁰ (Figure I-1).

Furthermore, metagenomics provides rapid and direct genomic analysis, allowing the identification of low-abundance genomes within a community¹¹ and preserving the original structure of bacterial populations by integrating taxonomic profiles with metabolic functions. This approach can reveal patterns related to microbial-driven human diseases^{12,13} and has the potential to generate vast gene and genome databases, enhancing our understanding of bacterial functional ecology and metabolism¹⁴. Additionally, it sheds light on community interactions in the environment such as symbiosis, competition, and pathogenicity¹⁵.

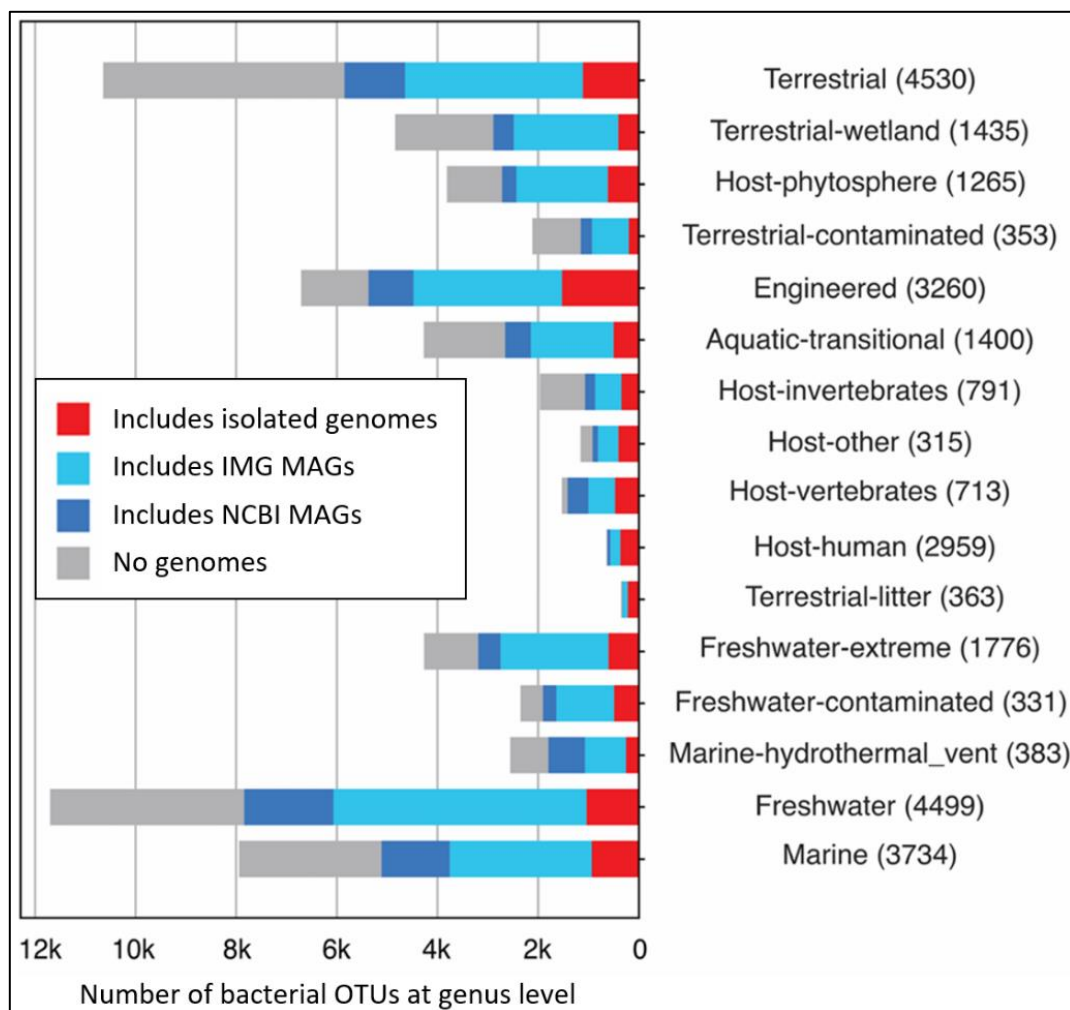


Figure I-1: Distribution of cultured bacterial genera across different biomes.

Horizontal barplot showing the number of bacterial OTUs detected in metagenomic samples from various environments (y-axis, with the number of samples in parentheses). OTUs are defined by USCG alanyl-tRNA synthetase sequences from genomes and metagenomes. Bars are stacked to reflect the genomic context of each OTU: red (includes cultured isolates), light blue (includes IMG/M MAGs, but no isolates), dark blue (includes only NCBI MAGs), and grey (includes only metagenome-derived genes, but no genomes). The x-axis indicates the total number of OTUs at genus-level. OTU: operational taxonomic unit. IMG: Integrated Microbial Genomes database. MAG: metagenome-assembled genome. Figure adapted from Wu *et al.*¹⁰.

The microbiome is defined as the collection of microorganisms present in an environment, along with their “theatre of activity”, which includes structural elements, metabolites, signal molecules, and the surrounding environmental conditions³. From a Molecular Biology perspective, its most relevant component is the metagenome of the underlying microbial community. There are two main approaches to studying a microbiome: metataxonomics and metagenomics.

Metataxonomics focuses on sequencing specific marker genes, such as the ribosomal RNA (rRNA) gene, which contain both conserved and variable regions that enable the classification of microbial communities and the profiling of their diversity. In contrast, metagenomics involves sequencing the metagenome of a sample, followed by assembly- or reference-based profiling and annotation to determine both the taxonomic composition and the functional potential of microbial communities, as well as to reconstruct whole-genome sequences.

These strategies are illustrated in Figures I-2 and I-3, respectively. Section I-2 reviews metataxonomic approaches, while Section I-3 focuses on metagenomic workflows. Section I-4 introduces targeted microbial enrichment strategies, which selectively enrich DNA prior to sequencing to enhance the detection and resolution of specific genes, organisms, or functions, representing an intermediate approach between metataxonomics and metagenomics. Finally, Section I-5 discusses quantitative metagenomics, highlighting the challenges that complicate accurate gene abundance measurements in metagenomes.

2. Metataxonomics

The main advantage of metataxonomics is its cost-effectiveness, requiring the sequencing of a single gene to identify a broad range of organisms across diverse environments. The most commonly used markers in metataxonomics are rRNA gene sequences, such as the 16S rRNA gene for bacteria, the 18S rRNA gene for eukaryotes (e.g. protists), and the internal transcribed spacer regions of the fungal ribosome for fungi. Other markers, such as the 23S rRNA gene, are used when the 16S rRNA gene’s resolution is insufficient for certain bacterial groups¹⁶. Additional markers, like ribulose-bisphosphate carboxylase large chain for photosynthetic organisms¹⁷ or cytochrome c oxidase I for metazoans¹⁸ help profiling organisms in specific ecosystems.

These markers are ideal for phylogenetic profiling due to their universal presence across populations, hypervariable regions for species differentiation, and conserved flanking regions amenable for amplification with universal primers. Another key benefit of rRNA analysis is the availability of extensive reference databases, such as SILVA¹⁹ or Greengenes²⁰ -the latter including around 20 million 16S rRNA sequences at the time of writing-, which provide taxonomic information for each reference. In contrast, the RefSeq collection²¹ contains less than half a million genome assemblies from Bacteria and Archaea, making specific rRNA gene databases more comprehensive for taxonomic classification than reference genome collections.

However, metataxonomics has notable limitations, paradoxically stemming from its main strength: targeting only specific genomic regions. First, this approach requires prior knowledge of the target gene or region. Because it typically focuses on a single marker gene, such as the 16S rRNA gene, it may misclassify two genetically distinct organisms that share identical sequences at the targeted locus. This limitation reflects the 16S rRNA gene's inability to capture intra-genomic heterogeneity, preventing it from distinguishing closely related strains with significant genetic differences in other genomic regions, such as those encoded on the extrachromosomal genome. Additionally, primer selection for 16S rRNA gene sequencing involves a trade-off between accurate taxonomic classification and abundance estimation, and minimizing host genome amplification, especially in low-biomass samples²². For example, primers targeting regions like V4 may lack sufficient sequence variation to distinguish between closely related species within a genus, such as *Lactobacillus acidophilus* and *Lactobacillus crispatus*, leading to both misclassification and overestimation of species abundances²³.

Another key limitation is the variability in 16S rRNA gene copy number across bacterial taxa. While some phyla carry a single copy, others such as some Firmicutes and Gammaproteobacteria possess multiple, sequence-divergent copies²⁴. Polymerase chain reaction (PCR) amplification further introduces biases by preferentially amplifying certain taxa due to mismatches between “universal” primers and target sequences, as well as differences in amplification efficiency²⁵. Together, these factors distort both diversity and abundance estimates, inflating representation for some clades and obscuring the true community composition. Moreover, the taxonomic resolution in metataxonomics may be limited when using short-read libraries. Long-read sequencing has demonstrated that full-length 16S rRNA

gene sequencing -approximately 1500 base pairs (bp)- offers higher resolution than targeting variable regions alone. Frequently used regions like V4 often lack sufficient variation to distinguish closely related species, reducing taxonomic accuracy²⁶.

The most widely used metataxonomic profilers for 16S rRNA sequencing data are DADA2²⁷ and QIIME²⁸. However, these tools often face limitations in accurately resolving taxonomy at the genus and species levels and have sometimes been outperformed by metagenomic profilers²⁹. Although 16S rRNA data can be used for indirect functional inference through tools like PICRUSt2³⁰, the accuracy of these predictions is constrained³¹. Additionally, the lack of strain-level resolution and dependence on incomplete reference databases further complicate functional annotations, often obscuring subtle but meaningful shifts in community function due to background noise³².

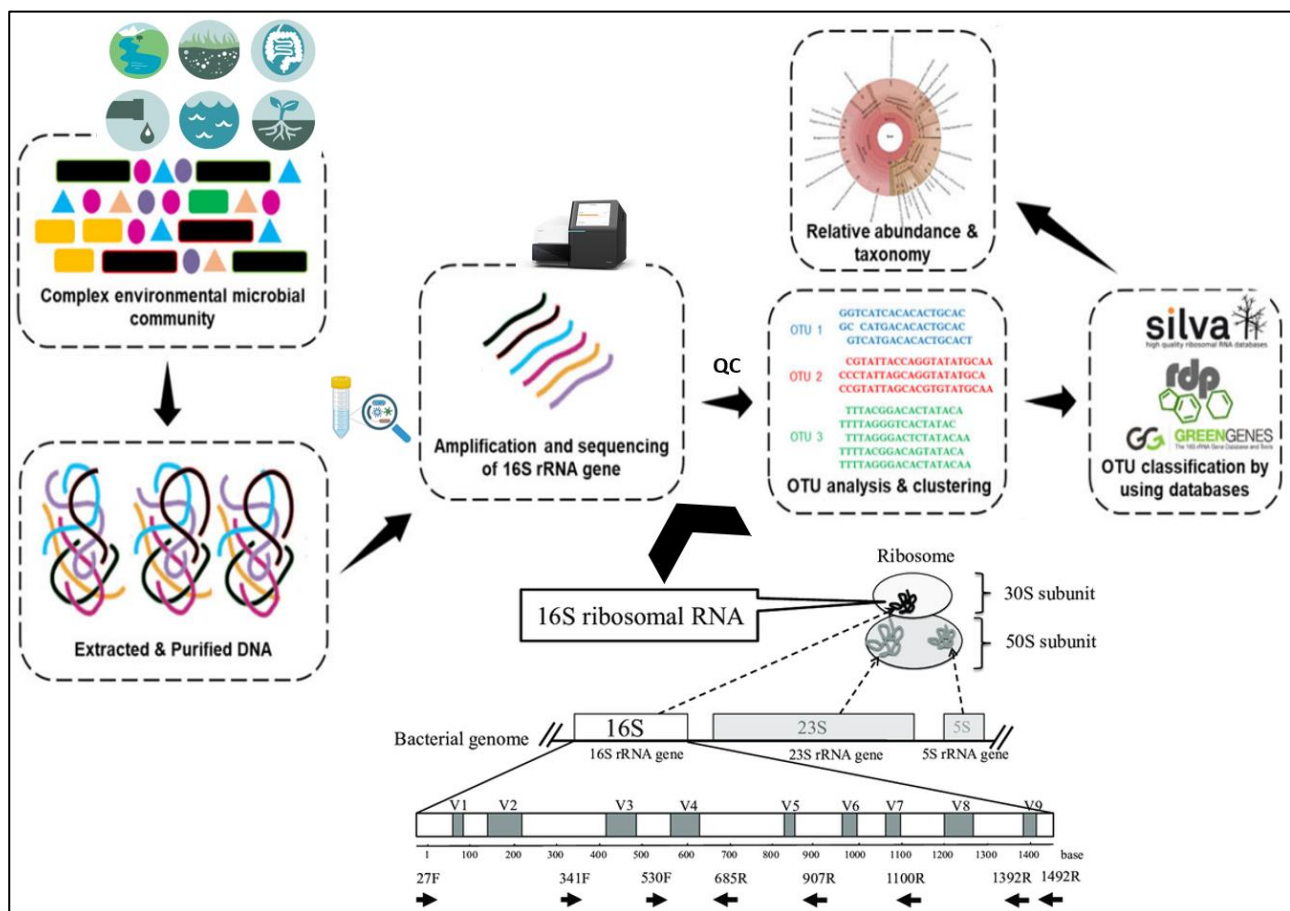


Figure I-2: Metataxonomics protocol for microbiome analysis using the 16S rRNA gene.

The process begins with the extraction of bacterial DNA from the biological sample, ensuring cell integrity is preserved. The 16S rRNA gene is then amplified via PCR with primers targeting conserved regions flanking hypervariable regions (e.g., V3, V4, V6, V8), which provide taxonomic resolution. The amplified 16S rRNA gene fragments are sequenced, and the resulting data undergo rigorous quality

control to remove low-quality reads, adapters, contaminants, and chimeric sequences. High-quality reads are clustered into operational taxonomic units based on similarity thresholds, which are taxonomically classified using reference databases. Finally, the relative abundance of microbial taxa is estimated to characterize the composition of the microbial community. Figure adapted from Ortiz-Estrada *et al.*³³.

3. Metagenomics

A) Levels of metagenomic analysis

Metagenomic data can be analyzed at three levels: reads, contigs and metagenome-assembled genomes (MAGs), each providing complementary insights into the composition and function of a microbial community (Figure I-3).

The first level consists of raw **sequencing reads**, which are DNA fragments directly generated by high-throughput sequencing platforms. Second-generation sequencers, primarily represented by Illumina technology, typically produce millions of short, highly-accurate reads of 100-300 bp. In contrast, third-generation platforms such as Oxford Nanopore and PacBio generate much longer reads -ranging from a few to tens of kilobases- but at the cost of lower throughput and higher base-calling error rates (between 5 and 15%)³⁴.

Short reads can be used for taxonomic profiling by mapping against reference databases. Tools like Kraken³⁵ and MetaPhlAn³⁶ are commonly applied to identify the presence and abundance of microbial taxa present in a metagenome. However, short-read data often provides limited taxonomic resolution, especially when differentiating closely related species or strains³⁷. In contrast, long reads can resolve complex genomic regions that are difficult to reconstruct with short reads alone, including repetitive elements³⁸ or complete microbial gene clusters³⁹. Long-read metagenomics is particularly valuable for characterizing novel or underrepresented taxa, closing gaps in genome assembly and resolving structural variants³⁸. Despite these advantages, long reads still face challenges in distinguishing highly similar genomes⁴⁰. Moreover, long-read datasets often require hybrid approaches, combining long and short reads, to balance sequence accuracy with assembly contiguity⁴¹.

Short reads are also valuable for functional profiling, where the presence of specific genes or metabolic pathways is inferred directly through read mapping to functional databases, such as Pfam⁴², KEGG⁴³ or MetaCyc⁴⁴, or indirectly using predictive tools like HUMAnN⁴⁵.

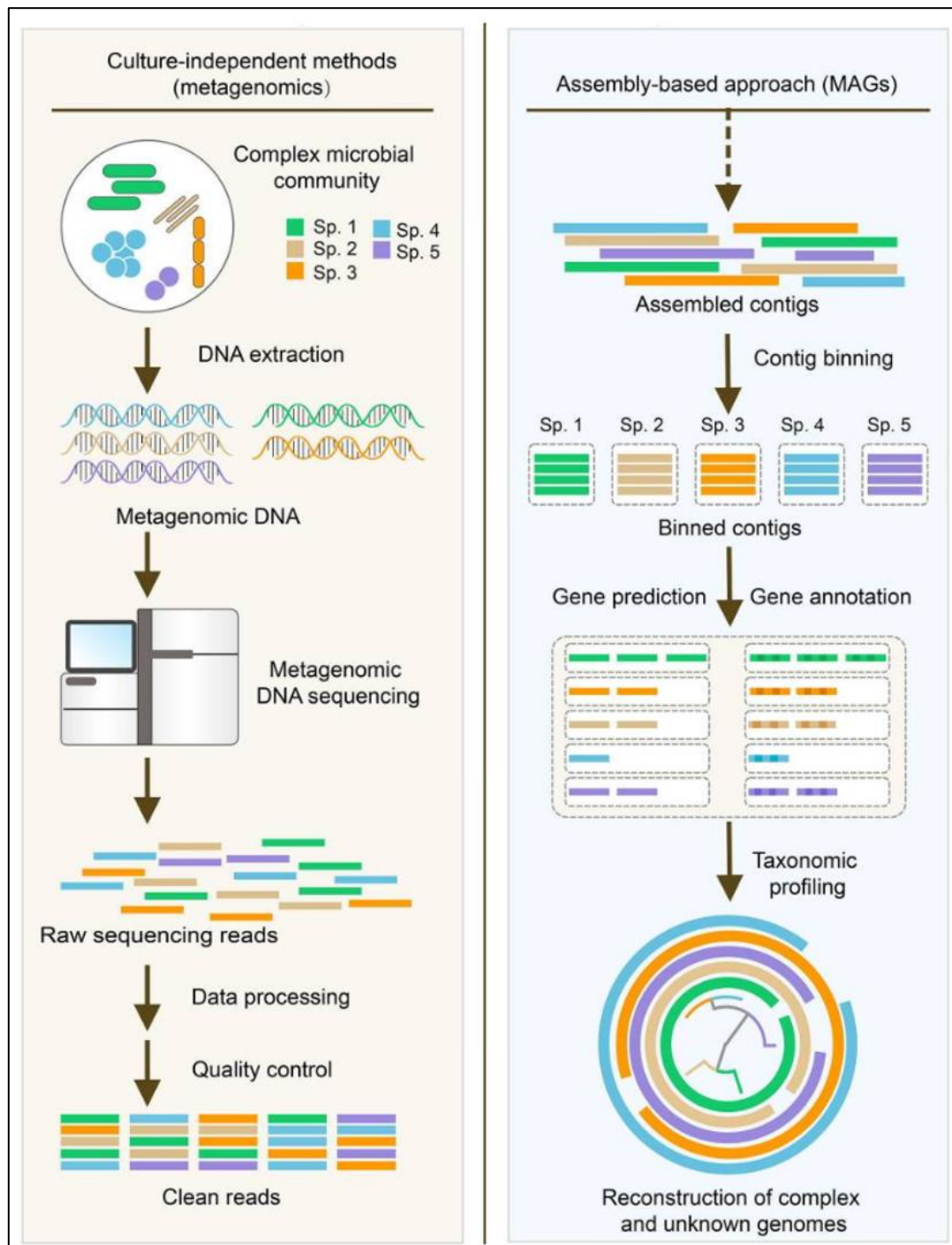


Figure I-3: Metagenomics protocol for microbiome analysis.

The process begins with the collection of biological samples and extraction of total DNA, ensuring minimal degradation and efficient lysis of microbial cells. The extracted DNA is then sequenced, producing short reads that undergo quality control to remove adapters, low-quality reads, and non-target DNA (e.g. host or eukaryotic DNA) to minimize sequence biases. High-quality reads are *de novo* assembled into contigs, which are subsequently grouped into MAGs through binning methods based on sequence composition and coverage patterns. Finally, MAGs are subjected to taxonomic and functional annotation to characterize the microbial community composition and its potential metabolic capabilities. Figure adapted from Yang *et al.*⁴⁶.

Reads can be assembled into longer, contiguous sequences (**contigs**) using computational tools that align overlapping reads to reconstruct larger fragments of the genome. This process is usually performed *de novo* (i.e., without a reference), although modern implementations are emerging to perform reference-guided assemblies⁴⁷. Tools like MEGAHIT⁴⁸ and metaSPAdes⁴⁹ are designed specifically for metagenomic assembly and can handle the complexity of mixed microbial communities. Contigs are fundamental to deconvolute microbial communities because they provide longer and more informative genomic sequences that can potentially represent complete genes or regions of microbial genomes. At this stage, open reading frame (ORF) prediction can be performed to identify genes within the assembled contigs. MetaProdigal⁵⁰ is the most widely used tool for ORF prediction in metagenomic data, providing the corresponding amino acid sequences. These predicted ORFs can be further clustered to identify novel protein families or functions using tools like CD-hit⁵¹ or MMseqs⁵². While contigs provide more resolution than individual reads, they are still fragmented and may not represent complete genomes.

The third level of analysis involves reconstructing **MAGs**, which are composite genomes of individual microorganisms recovered from metagenomes. The key step in this process is a binning protocol, where assembled contigs are grouped based on features such as sequence composition (e.g. GC content or tetranucleotide frequency) and similar abundance patterns (i.e., coverage across samples)⁵³. Tools such as CONCOCT⁵⁴, MetaBAT⁵⁵ and MetaBinner⁵⁶ are commonly used for this task. MAGs quality can vary depending on factors like sequencing depth, assembly fragmentation, and binning accuracy. Quality metrics such as completeness and contamination are typically assessed using tools like CheckM⁵⁷. High-quality MAGs (i.e., >90% completeness, <5% contamination)⁵⁸ are usually annotated for gene content and metabolic potential, using tools such as PROKKA⁵⁹ or eggNOG-mapper⁶⁰, enabling deeper insight into the functional roles of individual taxa within the community.

MAGs also provide higher taxonomic resolution, even at the species or strain level⁶¹, and can be placed within the microbial tree of life and integrated with reference genome phylogenies⁶². MAGs are particularly valuable for capturing rare or uncultured microorganisms that may be underrepresented in reference databases⁶³. Notably, the term MAG is sometimes used interchangeably with single-amplified genomes, although they are derived differently:

MAGs result from the assembly and binning of metagenomic data from microbial communities, while single-amplified genomes are obtained from the sequencing of isolated single cells⁶⁴.

Together, reads, contigs, and MAGs offer a multi-dimensional overview of the metagenome. Importantly, reads and contigs allow for quantitative analysis of taxonomic and functional abundance, while MAGs primarily allow qualitative characterization of microbiomes, offering a genome-centric view of microbial diversity and function.

B) Taxonomic profiling from reads

Taxonomic profiling from metagenomic reads aims to identify the microbial taxa present in a microbiome and quantify their relative abundances. Kraken³⁵ is a taxonomic classifier that assigns labels to sequencing reads by mapping their *k-mers* to a reference database of known genomes, classifying each read based on the majority of its *k-mers*. When a *k-mer* is found in two or more taxa, Kraken assigns it to their lowest-common ancestor, generating taxon-specific read bins and enabling abundance estimates based on read counts. In contrast, MetaPhlAn³⁶ is a taxonomic profiler that does not classify individual reads, but instead aligns them against a curated database of approximately 5 million clade-specific marker genes, which are unique to specific microbial lineages and enable accurate estimation of taxonomic composition in metagenomes⁶⁵. This marker-based strategy reduces biases from shared genomic regions and improves resolution at specific taxonomic levels, although it depends on the predefined set of genes⁶⁶. Other taxonomic classifiers include Kaiju⁶⁷, which aligns translated reads against a reference protein database, or Centrifuge⁶⁸, which efficiently classifies reads against compressed genome databases.

C) Metagenomic assembly and ORF prediction

Metagenomic sequencing involves extracting DNA from all cells in a microbial community, randomly fragmenting it, and sequencing the resulting short reads. Since long-range genomic information is lost during metagenomic library preparation, reconstructing genes and genomes from the sample requires inferring this information from the short reads. This is achieved through *de novo* assembly, which overlaps reads through a sliding window to construct a *de Bruijn* graph. In this graph, nodes represent *k-mers* (substrings of length *k*), and edges denote overlaps between them. The assembler then resolves this graph by traversing the nodes in the correct order, starting and ending at the same node, and connecting them through

edges if the $(k-1)$ -length suffix of one node is also a prefix of the next. This process results in non-branching paths that represent contigs. However, sequencing errors can lead to the formation of erroneous k -mer singletons -isolated reads that fail to assemble into contigs- or generate bubbles and hairs in the graph, leading to premature stops and paths with erroneous k -mers⁶⁹ (Figure I-4).

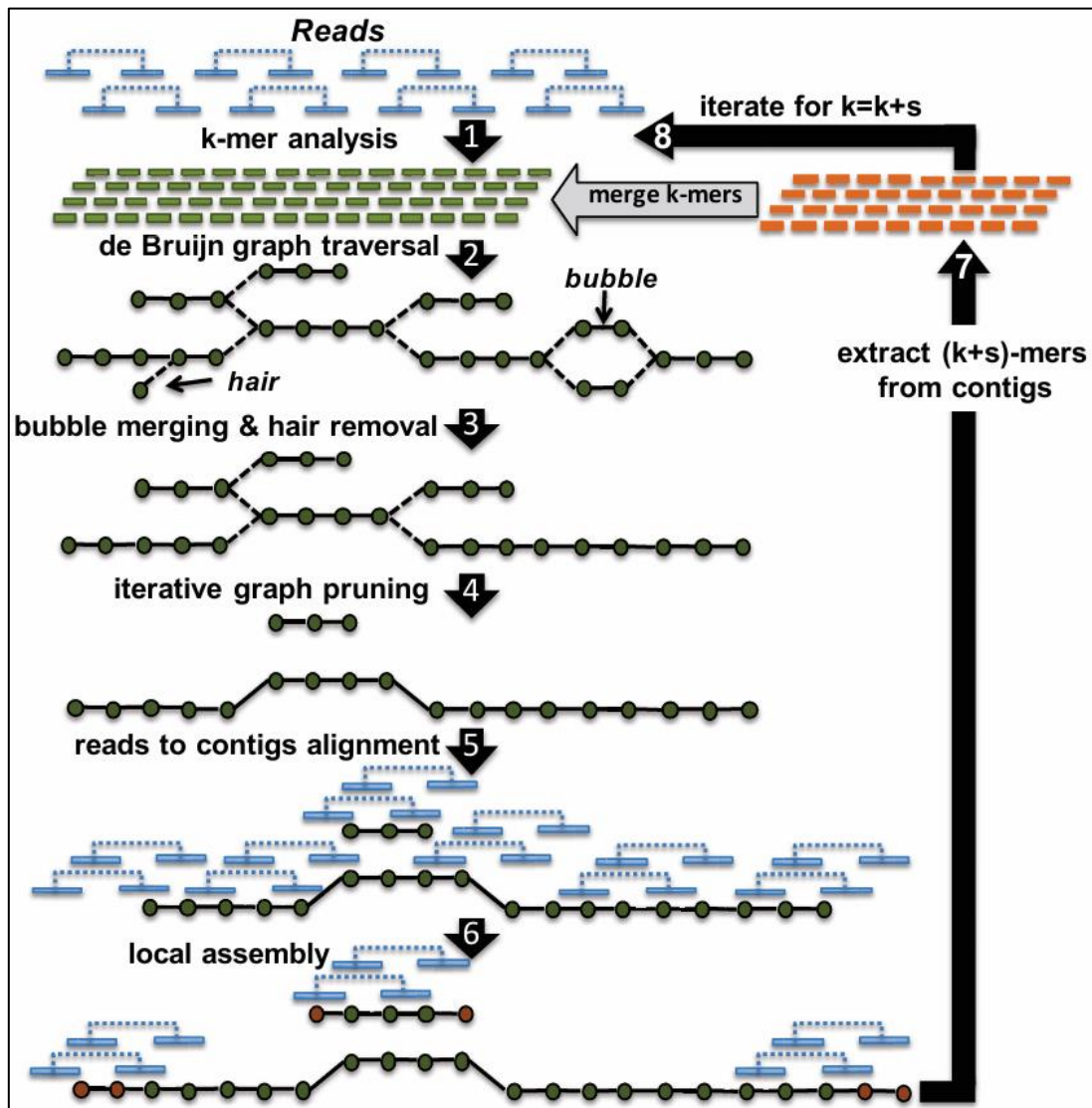


Figure I-4: Overview of *de novo* assembly algorithm.

The process begins with k -mer analysis, where sequencing reads are decomposed into overlapping substrings of length k . These k -mers form the nodes of the graph, and edges represent shared $(k-1)$ -length overlaps. A *de Bruijn* graph is then constructed and traversed, merging bubbles and removing hairs caused by sequencing errors or polymorphisms. Iterative graph pruning simplifies the structure by eliminating low-support paths. Reads are aligned back to the assembled contigs to refine connections, and local reassembly is performed to resolve ambiguities. Extended $(k+s)$ -mers are then extracted from contigs to increase resolution, where s is an incremental value. The graph is rebuilt with progressively larger k values to improve assembly quality and contiguity. Figure adapted from Georgeanas *et al.*⁷⁰.

The size of the *k-mers* is a crucial parameter in graph-based assembly. Smaller *k-mers* increase graph connectivity, which can aid in resolving low-coverage regions and filtering erroneous edges. However, they can complicate the assembly of repetitive genomic regions, as multiple edges lead to the same *k-mer*, creating many bubbles, and requiring more edges to be traversed to complete the graph. Additionally, if repeats are longer than the chosen *k-mer* size, they can tangle the graph and break the contigs into shorter fragments. On the other hand, larger *k-mers* reduce graph complexity and improve repeat resolution, but they may cause fragmentation due to missing overlaps and increased susceptibility to sequencing errors⁷¹. To address these challenges, modern assemblers, like metaSPAdes⁴⁹ and MEGAHIT⁴⁸, use iterative or multi *k-mer* approaches that balance the advantages and limitations of different *k-mer* sizes, enhancing assembly quality and robustness (Figure I-4).

A key limitation of *de novo* metagenomic assembly is the formation of chimeric contigs, which occur when sequences from different genomes are incorrectly joined into a single contig, often due to shared, similar regions. This misassembly is more likely in complex communities, where short reads from multiple species can be difficult to distinguish. Additionally, a proportion of reads -especially those from low-abundance organisms- often remain unassembled⁷¹. The overall performance of assembly depends on both the number of sequences and the complexity of the microbiome (species richness and evenness). Highly diverse environments, such as soil microbiomes, are particularly challenging, frequently resulting in more numerous and shorter contigs, and higher rates of chimerism⁷².

Identifying ORFs in metagenomic contigs is a key step in understanding the functional potential of microbial communities. However, metagenomic contigs are often fragmented, smaller than average genes, and may lack full-length genes, which makes ORF prediction more challenging⁷³. These short, incomplete and unidentified coding sequences complicate gene identification, especially when determining translation initiation sites. In prokaryotes, translation initiation is typically regulated by a ribosome binding site with a Shine-Dalgarno consensus located in the 5' untranslated region. However, about one-third of prokaryotic genomes in the RefSeq dataset lack this sequence or have not been annotated for it⁷⁴, many employing alternative mechanisms for translation initiation, such as leaderless translation, where the 5' untranslated region is absent⁷⁵. These complexities require specialized methods to accurately identify ORFs and translation initiation sites in metagenomes. Additionally, errors

may arise from incorrect assumptions on the genetic code or the presence of alternative genetic codes.

The high diversity of metagenomic data, coupled with the presence of incomplete or short genes, requires the use of specialized computational tools for ORF prediction. Traditional gene prediction methods often struggle with metagenomic sequences due to incomplete gene structures and variability in gene content across different species. To address these challenges, advanced metagenomic ORF predictors such as MetaProdigal⁵⁰ or MetaGeneMark⁷⁴ incorporate techniques to infer ribosome binding site sequences and promoter regions based on GC content, and they can handle alternative genetic codes. Additionally, filtering mechanisms based on confidence scores improve prediction accuracy, providing a clearer understanding of microbial functions and their metabolic capabilities in diverse environments.

D) Integrated pipelines for whole-metagenome analysis

There are two widely used automated pipelines that enable the complete analysis of a metagenome, integrating all steps needed to quantify both the taxonomic composition and metabolic potential of a microbial community, such as read quality control, *de novo* assembly, gene prediction, taxonomic and functional annotation, and MAG recovery through binning. These two workflows are HUMAnN⁴⁵ and SqueezeMeta⁷⁶, which differ primarily in their operational level: HUMAnN analyzes unassembled reads, while SqueezeMeta works at the contig level.

Briefly, HUMAnN begins by profiling the taxonomy of the metagenome using MetaPhlAn³⁶, as described in Section I-3B. Then, it constructs a sample-specific gene database by merging precomputed, functionally annotated pangenomes of the identified species. Reads are subsequently aligned at the nucleotide level against this database using Bowtie⁷⁷, while unmapped reads are aligned in the translated space against the non-redundant protein database UniRef⁷⁸ using DIAMOND⁷⁹. HUMAnN estimates gene family abundances by weighting read mappings based on alignment quality, gene length and gene coverage, both per organism and at the community level. Finally, it reconstructs and quantifies metabolic pathways by mapping gene families to enzyme functions and metabolic reactions through their Enzyme Commission (EC) number in MetaCyc⁴⁴ (Figure I-5).

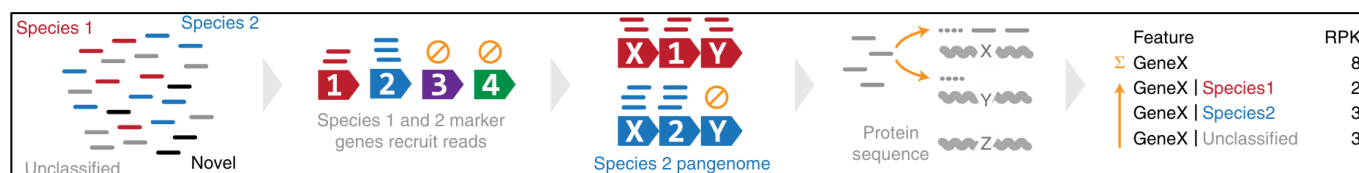


Figure I-5: HUMAnN workflow overview.

Metagenomic reads are taxonomically profiled with MetaPhlan to identify species present in the community. A database with the annotated proteins of these species is constructed. Reads are first aligned against species-specific functional pangenomes, and those that remain unclassified are translated and aligned against Uniref. Finally, gene family and pathway abundances are computed, stratified by species and at the community level. RPK: reads per kilobase. Figure adapted from Franzosa *et al.*⁴⁵.

SqueezeMeta, in contrast, supports the co-assembly of related metagenomes. It uses MetaProdigal⁵⁰ to predict ORFs on assembled contigs, and annotates them through an alignment against functional databases using DIAMOND⁷⁹. Specifically, DIAMOND predicts protein families by searching against the KEGG⁴³ and eggNOG⁸⁰ databases to assign ORFs to KEGG Orthology (KO) groups and clusters of orthologous genes, respectively. Gene abundances are quantified by mapping reads back to contigs, extracting the number of reads and aligned base pairs that map to each gene and contig. The pipeline then computes the average contig coverage, normalizing the relative gene abundances using the reads per kilobase per million reads (RPKM) method⁸¹. MAGs are then identified through binning, which benefits from co-assembly by exploiting shared abundance patterns and nucleotide composition across samples, since contigs from the same MAG share sequence features and tend to co-vary in abundance along the samples. A key advantage of SqueezeMeta is its ability to detect low-abundance genes by generating a shared contig reference set across samples, enabling the recruitment of reads to genes that would otherwise fail to assemble in individual metagenomes due to low coverage⁷⁶. Additionally, this pipeline supports downstream visualization and metabolic pathway reconstruction using KO groups⁸².

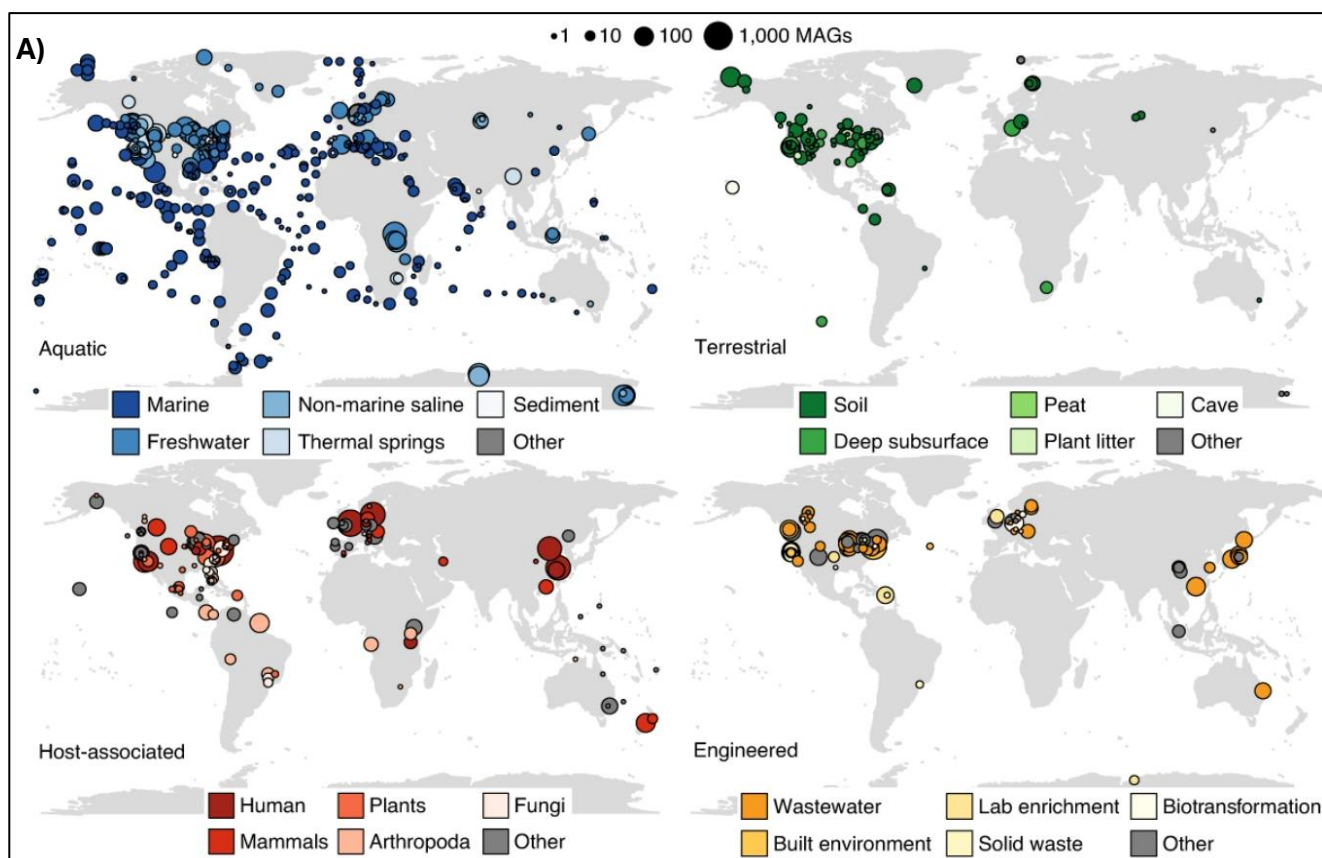
It is also worth mentioning the software Anvi'o⁸³, an open-source, community-driven platform that integrates a broad range of metagenomic analysis tools into a unified, modular environment with a strong emphasis on visual exploration. In addition to standard metagenomic analyses, its interactive interface supports manual exploration, allowing the import and refinement of contigs, genome bins and gene clusters. Unlike SqueezeMeta or HUMAnN, Anvi'o is less automated, but it is specifically designed for interactive, exploratory

workflows, offering visual inspection capabilities that enable in-depth analysis through rich graphical displays.

E) Advantages and limitations

Metagenomics overcomes the limitations of metataxonomics by sequencing the entire DNA content of a microbial community, rather than relying on a single marker gene such as the 16S rRNA gene. This enables the accurate detection of both well-characterized and novel organisms, as it captures genetic diversity across the whole genome. By examining the full genetic repertoire, metagenomics can distinguish between closely related strains and identify intra-genomic variations that are overlooked in 16S rRNA sequencing, including those found in mobile genetic elements (MGEs). This comprehensive approach allows for a more precise characterization of microbial communities, including the identification of rare or previously undetected taxa, as well as a deeper understanding of their functional potential. Importantly, this ability to identify novel species and strain-level diversity makes metagenomics an invaluable tool in environmental microbiology, clinical research, and biotechnological applications.

One of the major strengths of metagenomics is its ability to generate extensive catalogs of microbial gene and genomes, providing valuable insights into the functional and taxonomic diversity for microbial communities across diverse environments. These catalogs have been developed for the human gut microbiome (GM)^{84,85}, mouse⁸⁶, ruminant⁸⁷, and domestic animal GMs^{88,89}, oceans^{90,91}, cold seeps⁹² and soils⁹³ (Figure I-6). They serve as essential resources for understanding the genetic potential of microorganisms, often including genes and genomes from previously unculturable organisms. By capturing the vast genetic diversity of microbial ecosystems, these catalogs greatly enhance our understanding of global biogeochemical cycles, microbial functions, and the potential applications of microbial diversity across various fields, such as health, agriculture and biotechnology.



B)

Biome	Number of genes	Number of genomes	Reference
Human GM	170,602,708	204,938	84,85
Mouse GM	4,600,000	1,296	86
Ruminant GM	154,335,274	10,373	87
Pig GM	17,237,052	6,339	88
Chicken GM	16,565,684	12,339	89
Ocean	56,600,000	34,799	90,91
Cold seep	147,289,169	3,164	92
Soil	Unknown	40,039	93

Figure I-6: Genomic catalogues from metagenomic studies.

(A) Global distribution of MAGs recovered from metagenomic samples, grouped by biome based on environmental metadata. Each point represents a sampling location associated with one or more MAGs. Only MAGs with >50% completeness and <5% contamination were included. Figure adapted from Nayfach *et al.*⁹⁴. **(B)** Summary table of metagenomic gene and genome catalogs from some of the environments represented in **(A)**, including host-associated microbiomes and natural ecosystems.

Although most metagenomic studies are cross-sectional, the increasing number of longitudinal studies provides key advantages. These include the ability to distinguish inter- from intra-individual variability and to monitor microbiome changes over time. Longitudinal metagenomics enable the investigation of microbiome stability and dynamics, such as fluctuations in population sizes, within-host variation of GM metabolites, and evolutionary processes such as strain-level selection^{95,96}. They also reveal consistent shifts in microbial communities, such as increases in facultative anaerobes and decreases in obligate anaerobes in inflammatory bowel disease⁹⁷, and capture disease-associated disruptions in microbial gene expression, metabolite profiles, and host immune responses^{97,98}.

While metagenomics is a powerful tool for studying microbiomes, it faces several important limitations. Reference-based approaches rely heavily on the completeness and quality of microbial databases, which are often insufficient for poorly characterized environments. This limits confident assignment of reads and adds uncertainty due to sparse and inconsistent functional annotations that may misrepresent community capabilities³⁷. In addition, genome catalogues are biased towards model organisms and human pathogens, and this bias extends to metagenomic tools that rely on these catalogues. Many microbial genes also lack validated functional annotations, further limiting interpretability of alignments against gene families where only a few sequences are biochemically characterized, with the rest inferred by similarity⁹⁹.

Short-read sequencing further complicates analysis by losing the positional and genomic context of reads, requiring computationally intensive assembly to reconstruct genomes. Although taxonomic and functional profiles can be assigned directly from unassembled reads⁴⁵, this approach faces two major drawbacks: the high computational cost of aligning millions of reads against large reference databases⁴¹, and the limited accuracy of short reads for precise assignment¹⁰⁰. Assembly-based methods address the limitations of read length but demand substantial computational resources to process and store both raw and assembled data⁷¹.

Compared to metataxonomics, metagenomics involves higher costs, and remains sensitive to technical variability introduced during sample collection¹⁰¹, storage^{102,103}, DNA extraction¹⁰², and library preparation^{104,105}, all of which can affect both taxonomic and functional profiles. Sequencing biases, including GC content distortion¹⁰⁶, platform-specific

error rates¹⁰⁵, and PCR amplification artifacts, further compromise data quality. Biological and technical confounders, such as host DNA contamination, batch effects, and small sample sizes can obscure meaningful associations with phenotypes¹⁰⁷.

Other approaches to study a microbiome are also worth briefly acknowledging. These include metatranscriptomics, which analyzes gene expression in microbiomes through high-throughput sequencing of meta-cDNAs; metaproteomics, which profiles the proteome using liquid chromatography coupled with mass spectrometry for peptide identification; and metabolomics, which characterizes metabolite profiles using techniques such as nuclear magnetic resonance spectroscopy or mass spectrometry linked to liquid chromatography.

4. Targeted microbial enrichment strategies

Targeted microbial enrichment strategies enable the selective capture and sequencing of specific genetic elements such as antibiotic resistance genes (ARGs), MGEs or even complete viral genomes. Instead of sequencing all DNA present in a sample, these approaches perform targeted sequencing, allowing for a deep profiling of low-abundance or predefined targets, and enhancing detection and characterization without the need for cost-prohibitive, complete metagenomic sequencing.

Multiple displacement amplification is a method used to enrich microbial DNA from low-biomass samples prior to high-throughput sequencing. It uses Phi29 DNA polymerase, which has high processivity and proofreading activity, to amplify DNA with minimal fragmentation and generate long products (>10 kilobases). This increases total DNA yield and enables whole-genome amplification from isolated cells or scarce samples, making them accessible to metagenomic workflows¹⁰⁸. This technique has been widely applied in contexts such as single-cell genomics, virome studies¹⁰⁹, and low-biomass environments like groundwater¹⁰⁸. Despite its utility, it introduces biases such as under-amplification of high-GC regions, over-amplification of small circular genomes, chimera formation -mainly through inversions⁶⁴-, and uneven coverage¹⁰⁹. These biases can affect taxonomic profiles and assemblies, particularly at low DNA input.

Multiplex PCR-based sequencing extends conventional metataxonomic approaches by enabling the simultaneous amplification of numerous target genes in a single reaction. One early application targeted the 16S rRNA V1-V2 regions using forward primers concatenated with

sequencing adapters and unique barcodes, allowing multiplexing of up to 16 clinical specimens per run. Applied to sputum samples from cystic fibrosis patients, this method revealed low-abundance pathogens undetectable by culture and achieved species-level resolution through deep sequencing¹¹⁰. In another example, a panel of over 540 primer pairs was used to detect nearly 50 viral hemorrhagic fever agents directly from clinical material, providing rapid and sensitive diagnosis¹¹¹. More recently, multiplex PCR targeting genome fragments of common meningitis pathogens was paired with Oxford Nanopore sequencing to characterize associated ARGs¹¹².

Tiled-PCR amplification, a related method, uses a large number of overlapping primers to cover the entire genome of a target organism. This strategy ensures comprehensive coverage and supports accurate genome reconstruction, even in regions affected by mutations or low sequencing depth. Tiled-PCR is widely used for whole-genome sequencing of viruses from both environmental¹¹³ and clinical¹¹⁴ samples. Like all PCR-based methods, both multiplex and tiled-PCR approaches are limited by amplification biases and the precision of primer design, which can lead to uneven representation of targets due to mismatches or variable efficiency. Moreover, both approaches require prior sequence knowledge to guide primer development.

Genetic bait-capture platforms, by contrast, enrich target DNA by using custom-designed nucleic acid probes that hybridize to specific genetic regions, such as ARGs or MGEs, within a metagenome (Figure I-7A). These methods do not rely on PCR, reducing amplification bias and allowing greater tolerance to sequence variation. One of the first demonstrations of this technique used an array of 385,000 probes to reveal horizontal gene transfer (HGT) phenomena from a bacteriophage coinfection in a mixture of host, bacterial, and viral DNA¹¹⁵. The MEGaRICH platform, for example, enriched low-abundance resistome and virome components from fecal and wastewater-derived microbiomes, achieving more than a 100-fold increase in ARG-associated reads¹¹⁶. Similarly, the ResCap platform targets ARGs and MGE-encoding genes, significantly enhancing their detection and diversity compared to metagenomic sequencing¹¹⁷. Using this system, the authors achieved a 279-fold increase in the proportion of reads mapping to ARGs. ResCap has also been applied to environmental samples, revealing resistome dynamics in soil¹¹⁸. A yet-unpublished study has developed a different platform composed of 263,111 unique probes to enrich and analyze 14 key metabolic genes involved in nitrogen and methane cycling¹¹⁹.

Upstream **cell sorting-based** enrichment complements genetic bait-capture platforms by physically isolating target bacterial cells prior to sequencing (Figure I-7B). One early method integrated single-cell sorting with whole-genome amplification and fosmid library construction to recover light-harvesting genes from novel marine *Synechococcus* genomes¹²⁰. A subsequent approach combined fluorescent taxon-specific probes with flow cytometry to enrich defined clades for downstream genomic analysis. This strategy produced low-diversity metagenomes and enabled the recovery of high-quality single-cell assembled genomes, as shown in studies of an uncultured flavobacterial clade associated with phytoplankton¹²¹. More recently, cell sorting of antibiotic-resistant bacteria followed by metagenomic sequencing has been used to identify uncultured pathogenic strains and their associated ARGs and MGEs in soil¹²². Combined, these methods help overcome limitations of conventional metagenomic approaches, which can suffer from low sensitivity and specificity, hindering the detection of low-abundance microbial populations that fall below the detection threshold, as well as the identification of allelic variants linked to distinct phenotypic traits.

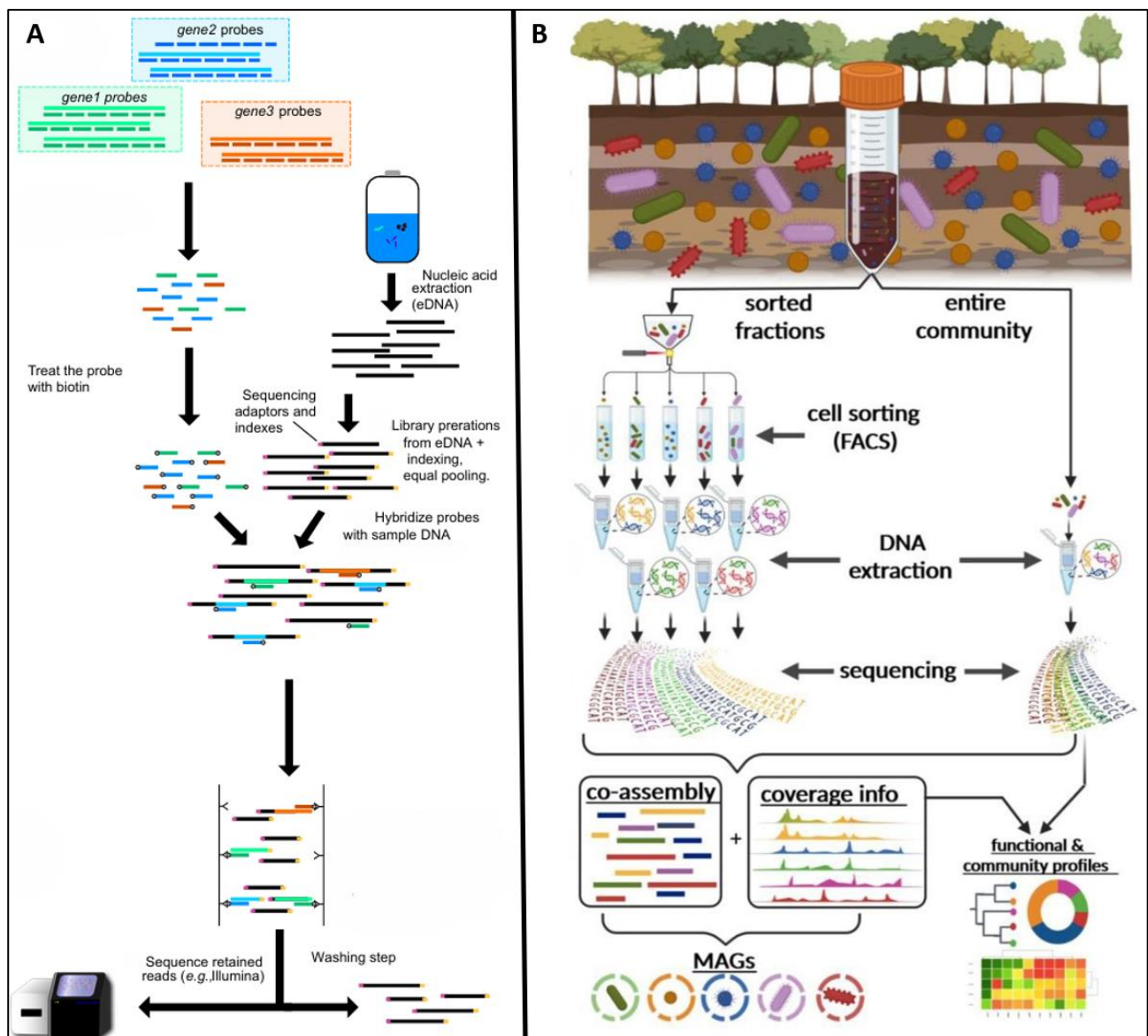


Figure I-7: Microbial genomic enrichment strategies.

(A) Overview of a genetic-bait capture platform illustrating the use of custom-designed probes to enrich three specific target genes. Figure adapted from Siljanen *et al.*¹¹⁹. **(B)** Overview of an upstream cell sorting-based enrichment approach applied to a metagenomic sample (left), compared to direct metagenomic analysis (right). Figure adapted from Vollmers *et al.*¹²³. FACS: Fluorescent activated cell sorting.

5. Quantitative metagenomics

Quantitative metagenomics aims to detect and compare the abundance of microbial taxa and genes across different environmental or host-associated communities. A common approach involves mapping sequencing reads to reference databases of genomes or gene sequences and counting how many reads are assigned to each taxon or functional category. Taxonomic composition can be estimated using tools such as Kraken³⁵ or MetaPhlAn³⁶, while

functional profiles are typically inferred by aligning reads against genic databases such as KEGG⁴³, MetaCyc⁴⁴ or eggNOG⁸⁰. Read counts are then used to estimate the abundance of taxonomic groups or gene families. However, accurate quantification is challenged by technical limitations, particularly the dependence of read counts on sequencing depth. Without proper normalization, comparisons between samples can be biased and lead to inaccurate interpretations of community structure or function.

Absolute quantification in metagenomes can be achieved using internal DNA standards known as spike-ins, which are synthetic DNA sequences added to samples in known quantities before sequencing. These standards provide a reference framework to convert sequencing read counts into absolute concentrations. Recent studies have applied spike-ins to quantify DNA viruses in wastewater samples¹²⁴, to model absolute bacterial abundances in mock and real communities¹²⁵, or to estimate absolute ARG abundance in farm-associated samples, expressed as gene copies per unit mass of sample¹²⁶. These metagenomic quantifications correlate well with control quantitative PCR measurements while avoiding primer biases, although their detection limit remains higher.

High-throughput metagenomic sequencing yields compositional data (Figure I-8). Sequencing instruments have a limited capacity to generate a fixed number of reads per sample, capturing only a random sample of molecules from the total DNA pool that is constrained by sequencing depth. As a result, read counts represent relative proportions of features (e.g., taxa or genes) within each metagenome, and not their true abundances. This compositional constraint implies that an increase in the abundance of one feature can artificially lower the observed abundance of others, even if their true amounts remain constant. Consequently, ignoring compositionality can produce spurious correlations and obscure genuine biological differences¹²⁷. This issue affects both taxonomic and functional profiling, biasing interpretations of gene and pathway dynamics¹²⁸. While it is being increasingly addressed through improved methods for assessing microbial composition and differential abundance^{129,130}, it remains less explored in functional profiling.

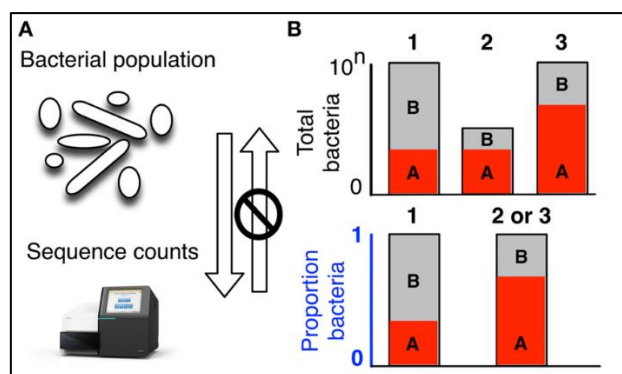


Figure I-8: Compositional problem in metagenomics.

(A) The total read count in a metagenomic library does not reflect the absolute abundance of nucleic acids in the microbial community; but instead, it provides a relative estimate constrained by sequencing depth. Read counts represent the proportion of molecules per feature (e.g. taxon or gene) in a metagenome, scaled by the total number of sequencing reads. **(B)** Bar plots illustrate the difference between absolute counts and relative proportions for two features across three samples. The top graphs show the actual counts in the community, while the bottom graphs display the relative abundances obtained after sequencing. Notably, the features in samples 2 and 3 have identical relative abundances despite differing input counts. Figure adapted from Gloor *et al.*¹²⁷.

Functional quantification of a metagenome involves cataloguing its genes and estimating their abundance to describe the functional capacity of the community. In practice, sequencing reads are aligned against gene or protein databases, and counts are aggregated per gene family. By comparing these counts, genes that differ in abundance between samples can be identified, revealing functional differences between communities. Gene abundance can be defined through several parameters, such as absolute abundance, relative abundance, or average gene copy number (Figure I-9).

The absolute abundance of a gene is the number of times it is detected in a metagenome, either by accounting the number of mapped reads against it or the number of gene copies in assembled contigs. In contrast, the relative abundance of a gene is the proportion of all genes that belong to a given gene family. A related commonly used measure is genic average copy number, which represents the expected number of copies per randomly selected microbial cell. This value is typically <1 for most gene families, ~ 1 for universal single-copy genes (USCGs), and >1 for gene families with multiple paralogs. USCGs, such as RNA polymerase subunit genes (*rpoB*) or chromosomal replication initiator protein DnaA (*dnaA*), are found in nearly all prokaryotic genomes and usually occur in a single copy, making them useful for normalization and cross-sample comparisons^{131,132}. Importantly, all these measures remain compositional, since they count only in the sequenced portion of the community.

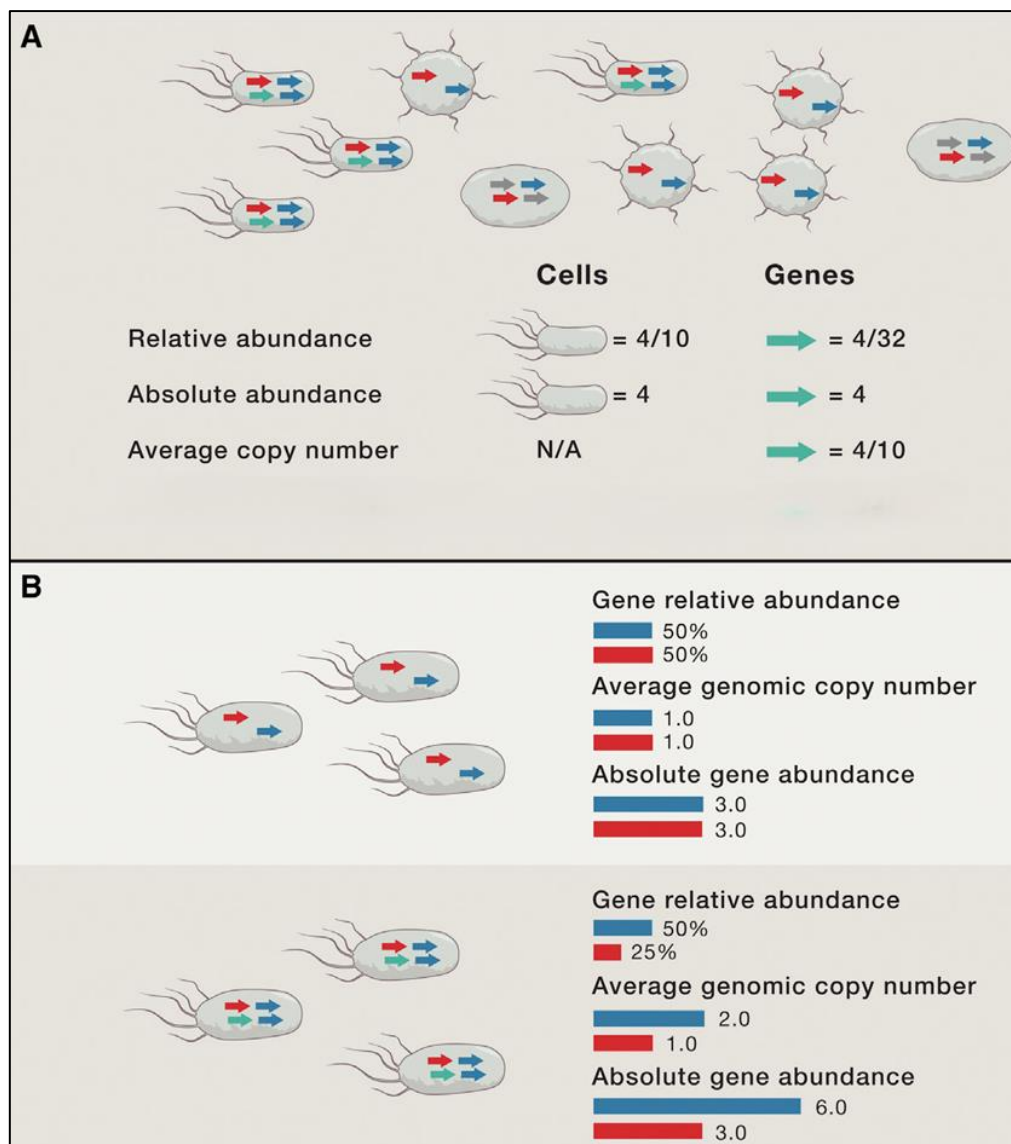


Figure I-9: Gene abundance metrics used in metagenomics.

(A) A microbial community composed of ten cells from four taxa, each carrying different combinations of four gene families (colored arrows). The abundance of the green gene is illustrated using different quantification metrics. **(B)** The red gene is present at one copy per cell, resulting in a constant absolute abundance across both communities. However, its relative abundance decreases as genome size increases due to an expansion in the copy number of the blue gene and the apparition of the green gene. Figure adapted from Nayfach *et al.*¹³³.

While gene abundance metrics offer valuable insights into the functional composition of microbial communities, their accuracy can be compromised by several technical biases. To mitigate these effects, several normalization strategies have been proposed, each tailored to correct specific sources of variability (Table I-1). Among the most common challenges to deal with is the distortion in sequencing depth across samples. A sample sequenced more deeply will naturally yield a higher number of reads for any given gene, regardless of whether its true

abundance in the community has changed. Consequently, raw read counts cannot be directly compared between samples with different library sizes. Gene length also introduces bias for an analogous reason, as longer genes inherently attract more reads during alignment. Normalization methods such as RPKM⁸¹ adjust for both sequencing depth and gene length^{134,135}.

Another important source of bias in metagenomic gene abundance estimation arises from the selection of the reference databases, which constrain the set of genes that can be detected. Since only reads that match sequences in the database can be annotated, the choice of reference directly influences the abundance profile. This limitation can be mitigated by using curated or multiple reference databases and comparing results across annotations to assess consistency^{99,136}.

Biases linked to microbial genome size further complicate interpretation. The probability of sampling a particular gene from a metagenome is inversely related to the average genome size of the community¹³⁴. As a result, communities dominated by organisms with larger genomes may appear to have lower functional potential simply due to reduced gene sampling rates. The probability of a sampling a gene (P_g) from a metagenome can be defined as the probability that a sequencing read originates from that gene, and it can be approximated as:

$$P_g \cong \frac{l_g \times C_g}{AGS}$$

where l_g is the gene length, C_g its average copy number, and AGS the average genome size in the microbiome. This formulation highlights how USCGs (for which $C_g = 1$) are overrepresented in a metagenome with small AGS compared to a metagenome with a large AGS, simply because these genes make up a larger fraction of the total genomic DNA in the community¹³⁴.

Nayfach *et al.*¹³⁷ formalized a normalization method, implemented in the MicrobeCensus tool, that accounts for average genome size to estimate gene abundance in terms of reads per kilobase per genome equivalent (RPKG):

$$RPKG = \frac{N_{aligned} / l_g}{N / AGS}$$

where $N_{aligned}$ is the number of reads mapped to the gene, l_g its length, N the total library size, and AGS the average genome size, itself further estimated as:

$$AGS = \sum_{i=1}^n \frac{R_i \times S_i}{\sum_{i=1}^n R_i}$$

with R_i and S_i denoting the relative abundance and size of genome i , respectively. This enables normalization per microbial cell rather than per unit of DNA.

GC content also influences gene detectability. Regions of low GC content tend to be underrepresented in sequencing data due to reduced nucleotide stability, leading to the underestimation of genes or organisms with extreme GC percentages. Correction strategies that explicitly model GC bias can be applied to adjust estimates and avoid distortions in comparative analyses¹³⁸. However, these approaches normally assume that representative genome references are available for most microorganisms in the sample, an assumption that may hold for well-characterized microbiomes such as GM, but not for poorly studied environments.

Biases introduced during read alignment can significantly affect gene quantification. Both the choice of aligner (e.g. BLAST¹³⁹ or DIAMOND⁷⁹), and the specific alignment parameters, such as the E-value or the minimum percentage identity thresholds, strongly influence which read-to-gene matches are retained. Permissive thresholds may inflate false positives by allowing spurious matches, while overly stringent thresholds risk missing true hits, especially in divergent genes. The treatment of multi-mapping reads, which align equally well to multiple genes or loci, leads to ambiguous quantification and adds another layer of complexity¹⁴⁰. Empirical studies suggest that optimal gene detection is achieved using read lengths of 150-200 bp and conservative E-value thresholds (e.g. 1E-10), which balance sensitivity and specificity⁹⁹.

Finally, one of the most complex normalization steps involves correcting gene counts based on the taxonomic background of each sample. Marker genes with stable copy numbers across genomes are often used to standardize gene abundance relative to microbial cell counts. The normalization proposed by Pal *et al.*¹³⁵ uses the ratio between the gene of interest and 16S rRNA gene abundance to determine the abundance of a gene in a metagenome (A_g):

$$A_g = \frac{N_{aligned} \times l_g}{N_{16S} \times l_{16S}}$$

where $N_{aligned}$ and l_g are the number of reads and average length of the target gene, and N_{16S} , l_{16S} the same for the 16S rRNA gene. While widely used, for example to quantify ARG abundance across environments¹³⁵, the variable copy number of 16S rRNA gene across taxa²⁴ limits its reliability.

A more consistent alternative is to normalize gene abundance using USCGs, which are typically present at one copy per genome across most taxa¹³². Methods such as MUSiCC¹²⁸ or MicrobeCensus, which uses a core of 30 USCGs¹³⁷, apply this principle by estimating average gene copy numbers through normalization against the median abundance of USCGs to estimate per-cell gene copy numbers, thereby enabling more accurate cross-sample comparisons.

Source of bias	Normalization	Reference
Sequencing depth	By library size (e.g. RPKM)	99,136,137
	By sum of all gene abundances	141
Gene length	By gene length (e.g. RPKM)	134,135
Genome size	By average genome size	134,137
GC-content bias	Apply GC-content correction	138
Reference database	Use of curated or multiple databases	99,136
Read alignment ambiguity	Optimize alignment tool and parameters	99
Microbial background	By 16S rRNA abundance	135
	By USCG abundance	128,137

Table I-1: Bias sources in metagenomic gene quantification and normalization strategies.

After raw sequencing reads are aligned to a reference database, multiple sources of technical bias may affect gene abundance estimates (column 1). Various normalization strategies (column 2) have been proposed or applied in previous studies (column 3) to mitigate these effects.

When estimating the complete set of differentially abundant genes between two or more metagenomes from distinct conditions, statistical methods originally developed for RNA-seq count-based data, such as DESeq2¹⁴², have shown good performance while minimizing the false discovery rate¹⁴³. Unlike targeted comparisons involving a small number of genes, ORF-wide analyses face the additional challenge that many genes may be represented by only a few, or even zero, sequencing reads. Consequently, methods for statistical inference must be robust to high levels of noise and capable of detecting true differences with limited information. In this context, both gene abundance and sample size are key factors influencing the power to detect differential abundance¹⁴³. A subsequent evaluation by the same group found that normalization strategies such as the trimmed mean of M-values and relative log expression outperformed other methods in producing unbiased p-values and effectively controlling the false discovery rate, and are therefore recommended for normalizing whole-gene abundance data in metagenomic studies¹⁴⁴.

II. OBJECTIVES

1. General hypothesis

Analyzing entire metagenomes requires substantial computational power and manual effort, often without yielding clear biological interpretations of the roles played by specific microbial protein families. A more efficient strategy is to focus on well-defined gene families with known functions. This targeted, gene-centric approach facilitates answering biologically relevant questions about key microbial functions and their relationship with host or environmental phenotypes, while reducing analytical noise and computational burden.

2. Objectives

The main objective of this thesis is to quantify selected gene families associated with specific phenotypes using a targeted, gene-centric approach to large-scale metagenomic datasets. Chapter I focuses on genes from the GM involved in microbial metabolism linked to gut disease. Chapter II examines plasmid-marker genes in several ecosystems. Gene abundance is quantified in clinical samples from healthy individuals and patients with gut disorders in Chapter I, and in marine and other environmental samples in Chapter II. By correlating gene abundance with relevant phenotypes -either disease status or plasmid prevalence- this thesis aims to reveal biologically meaningful functional patterns that may be overlooked by broader, taxonomy-based approaches. Five specific objectives were established:

- Characterize the enzymes and their coding genes responsible for the production of short chain fatty acids, alcohols and other metabolites in the human GM.
- Determine the correlation between these metabolic signatures and gut disease by quantifying target genes in fecal metagenomes from patient cohorts.
- Investigate whether candidate genes belong to the accessory genome of the GM.
- Assess the distribution and prevalence of marine plasmids by quantifying relaxase genes in marine and other environmental metagenomes.
- Explore the correlation between relaxase gene abundance and the prevalence of antibiotic-resistance genes and other plasmid-marker genes.

CHAPTER I

METABOLIC-MARKER GENES

IN MASLD

1. Introduction

Metabolic dysfunction-associated steatotic liver disease (MASLD) is the most common chronic liver disease worldwide, affecting approximately 38% of the adult population¹⁴⁵. Clinically, MASLD is defined by the presence of steatosis (i.e., accumulation of intrahepatic triglycerides) in more than 5% of hepatocytes, in association with metabolic risk factors (particularly, obesity and type 2 diabetes) and in the absence of excessive alcohol consumption (≥ 30 g per day for men and ≥ 20 g per day for women) or other chronic liver diseases¹⁴⁶. However, the term encompasses a spectrum of pathological situations, ranging from simple steatosis to metabolic dysfunction-associated steatohepatitis (MASH), which can further progress to fibrosis, cirrhosis and hepatocellular carcinoma^{147,148} (Figure III-1). Beyond hepatic complications, MASLD is strongly associated with cardiovascular disease, extrahepatic cancers and chronic kidney disease, highlighting its role as a multisystemic metabolic disorder¹⁴⁹. The pathogenesis of MASLD is complex and multifactorial, involving genetic susceptibility, environmental exposures, inflammatory and metabolic factors, with a strong association with insulin resistance¹⁵⁰.

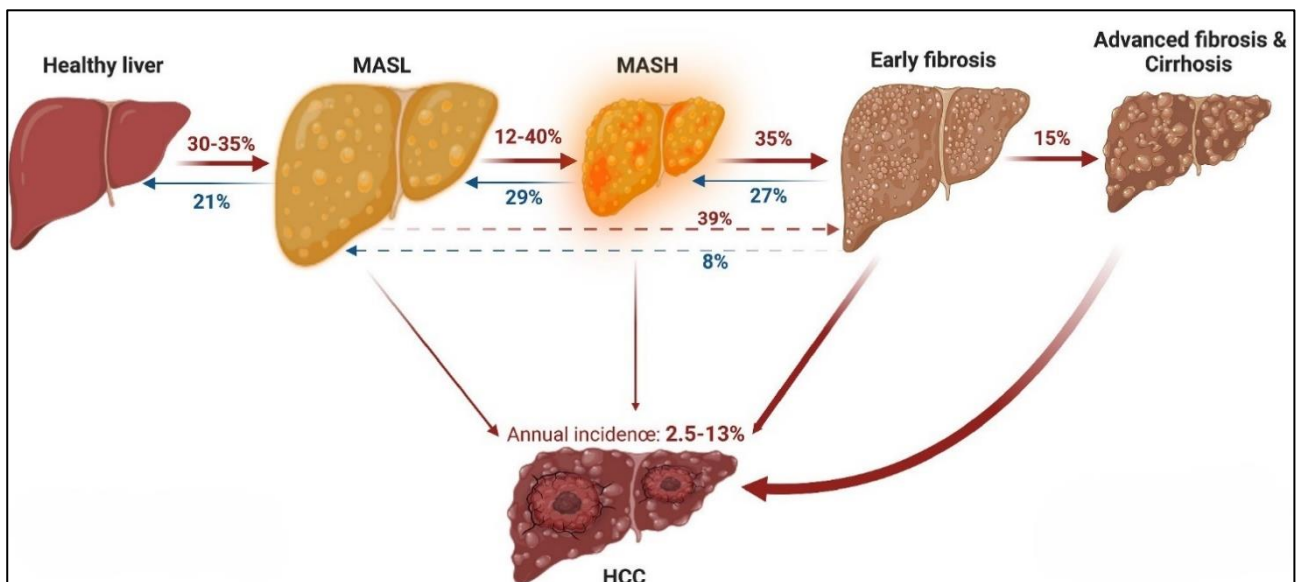


Figure 0-1: Overview of MASLD progression.

Stages of liver disease and the estimated percentage of patients progressing between them. Figure adapted from Lekakis *et al.*¹⁴⁸. MASL: metabolic dysfunction-associated steatotic liver, MASH: metabolic dysfunction-associated steatotic hepatitis, HCC: hepatocellular carcinoma.

The GM interacts with the liver via the gut-liver axis (GLA) through the portal vein, which transports GM-derived metabolites to the liver. The intestinal barrier plays a crucial role in this interaction, promoting the movement of water and nutrients into circulation towards the liver while preventing the systemic spread of microbes and toxins. It is composed of a mucus layer and an epithelial monolayer of specialized cells interconnected by tight junctions that regulate the paracellular passage. The mucus, primarily composed of large glycosylated proteins known as mucins, serves to shield the epithelial lining from direct bacterial contact. Additionally, there is a third barrier known as the gut-vascular barrier which further contributes to maintaining intestinal integrity¹⁵¹ (Figure III-2).

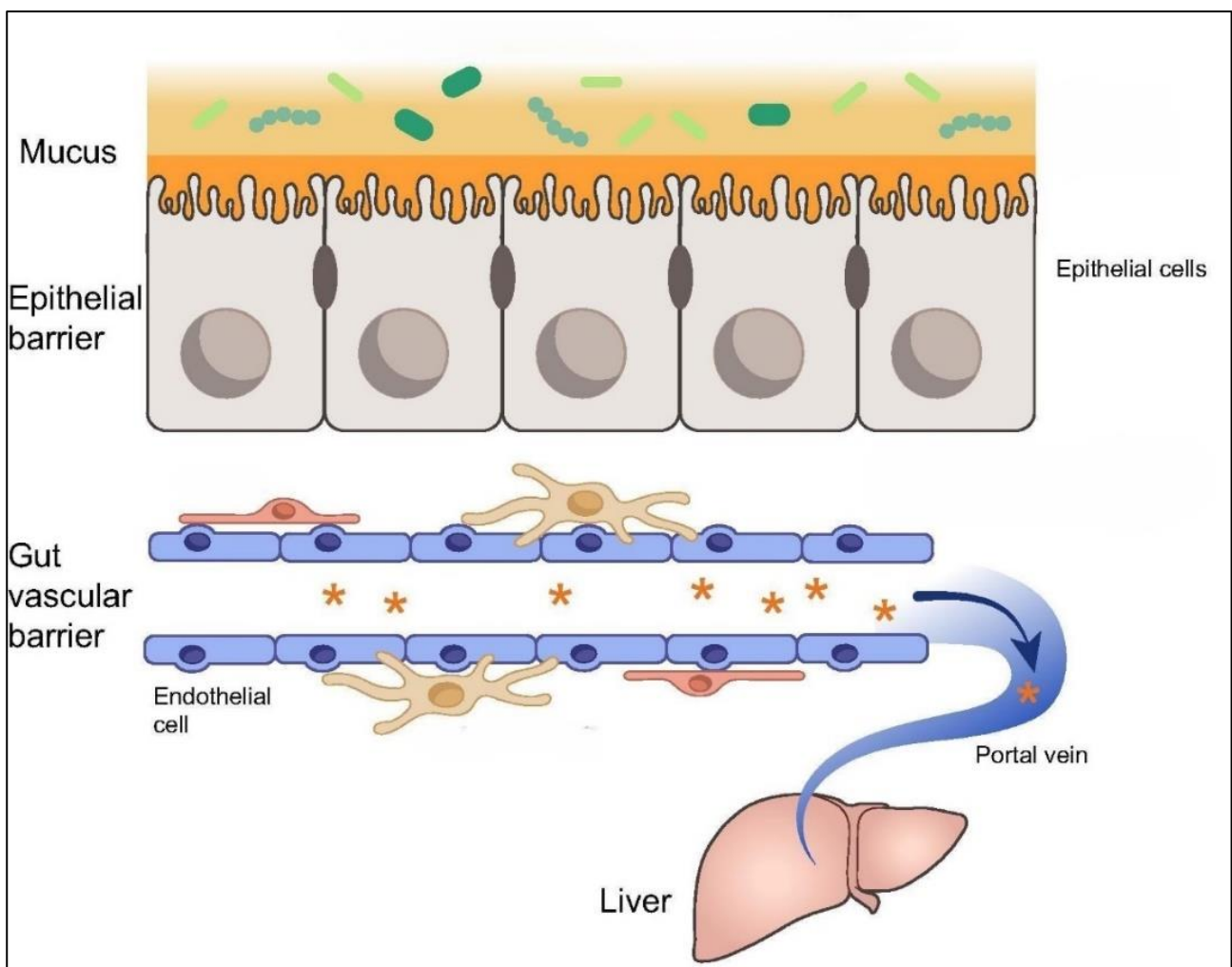


Figure 0-2: Gut-liver axis structure.

Schematic representation of GLA, illustrating the key components involved in gut-liver communication. Figure adapted from Albillos *et al.*¹⁵².

The GLA can be disrupted by multiple interconnected factors, affecting the progression of MASLD¹⁵³. Alterations of the GM composition and/or their associated metabolism can lead to an increased production of harmful metabolites such as pathogen-associated molecular patterns, which induce inflammatory responses in the liver when they reach it through the bloodstream, leading to liver injury^{154,155}. The triggering of inflammatory pathways is mediated by activation of pattern recognition receptors like Toll-like receptors in hepatic stellate cells and Kupffer cells¹⁵⁶. Transit of the aforementioned endotoxins into portal circulation is facilitated by an increased intestinal permeability caused by a disruption of tight junction proteins in the intestinal membrane¹⁵⁷. Additionally, alterations of the bile acid composition and circulation affect liver metabolism and function. Since bile acids play a role in the regulation of glucose and lipid metabolism, their dysregulation contributes to gut inflammation and MASLD^{158,159}.

Several human and preclinical studies have attempted to establish a causal link between bacterial metabolism and MASLD^{160,161}. However, metabolic levels can also be influenced by host metabolism, diet or environmental factors¹⁶². Similarly, metataxonomic analyses have explored the relationship between GM composition and MASLD^{163,164}. Although 16S rRNA gene sequencing provides taxonomic resolution at the genus or species level, it fails to capture strain-level metabolic variability. Bacterial genomic plasticity allows key metabolic genes, such as those potentially relevant to MASLD pathogenesis, to be encoded in their accessory genome, including MGEs, further decoupling taxonomy from metabolic function¹⁶⁵.

In this Chapter, we used the genes encoding metabolic enzymes involved in the final steps of microbial pathways leading to the production of butyrate, short-chain alcohols (SCAs) ethanol and propanol, methane, trimethylamine (TMA) and trimethylamine N-oxide (TMAO) as proxies to determine their abundance in MASLD. To achieve this, we isolated these gene families and quantified them in over 550 metagenomes obtained from fecal samples of MASLD and healthy patients, stratified across three independent cohorts. We also explored taxonomic signatures to gain a comprehensive view of bacterial diversity and abundance. Finally, we assessed the distribution of these genes within the human GM pangenomes and plasmids. By focusing on functionally relevant genes, we aim to move beyond correlative taxonomic observations to provide mechanistic insights into the functional reprogramming of GM in MASLD. This gene-centric framework provided insights into the role of microbial pathways in MASLD progression and highlighted new genetic targets for diagnosis and intervention.

2. Materials and methods

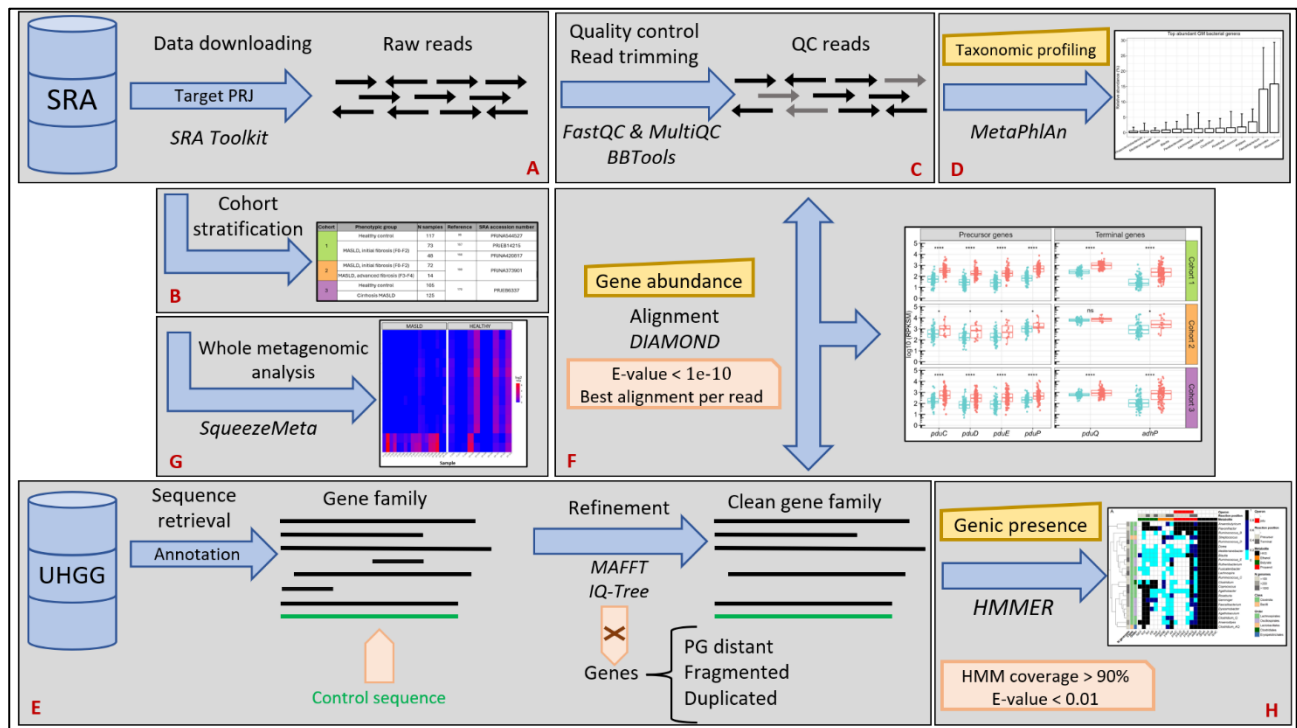


Figure 0-3: Overview of the methodology used in Chapter I.

Diagram summarizing the full bioinformatic pipeline for processing paired-end metagenomic samples: from data retrieval and read quality control to taxonomic profiling and target gene quantification. Gene families were isolated and aligned against the sequencing libraries to quantify abundance. Additionally, the presence of target genes was screened across GM genomes and annotated plasmids. Blue arrows indicate the workflow progression, and software tools are shown in italicized black beneath each arrow. Filtering steps are highlighted in pink boxes. Red labels refer to subsections (A-H) in Materials and Methods for detailed descriptions.

A) Downloading of metagenomic sequencing data

A total of 554 metagenomic sequencing libraries generated from the fecal samples of MASLD and healthy subjects were downloaded from the Sequence Read Archive (SRA) with SRA Toolkit's *fastq-dump* (version 3.2)¹⁶⁶. These samples were retrieved using the accession numbers corresponding to five different studies, and they were stratified in three cohorts, as detailed in Table III-1. Information on the collection and processing of fecal samples, bacterial DNA extraction and sequencing protocols are provided in the original publications. Additionally, we retrieved the medical metadata associated with each sample donor to verify MASLD diagnosis and classify samples into two phenotypic groups of comparison within each cohort. Samples with undefined MASLD status were excluded from this study. All data was downloaded between March 2022 and March 2024.

B) Description of MASLD patient cohorts

Cohort 1 includes 83 European MASLD patients with obesity and initial fibrosis (F0-F2), aged 20-64 years. Patients were drawn from two previously published studies: 73 individuals from Hoyles *et al.*¹⁶⁷ and 10 from Mardinoglu *et al.*¹⁶⁸. In the latter group, 48 fecal samples were collected at five time points during the study (two samples from day 1 were not obtained). MASLD diagnoses were confirmed by liver biopsy performed during bariatric surgery. None of the subjects had alcoholic liver disease, viral hepatitis or diabetes mellitus. The control group includes 117 healthy patients from Poyet *et al.*⁹⁵. Cohort 2 comprises 86 American MASLD patients from Loomba *et al.*¹⁶⁹, including 72 with initial fibrosis and 14 exhibiting advanced fibrosis (F3-F4), all diagnosed by liver biopsy and magnetic resonance imaging. No participants had liver comorbidities or diabetes. Note that in this study, patients with initial fibrosis were treated as a “control” group for comparison purposes. Cohort 3 includes 230 Han Chinese individuals from Qin *et al.*¹⁷⁰: 125 diagnosed with MASLD-related cirrhosis and 105 without liver injury. Patients with viral hepatitis or other hepatic disorders were excluded. All cohorts consist of one metagenomic stool sample per patient, except for the 48 samples collected at multiple time points from 10 patients in the study from Mardinoglu *et al.*¹⁶⁸.

Cohort	Phenotypic group	N samples	Reference	SRA accession number
1	Healthy control	117	95	PRJNA544527
	MASLD, initial fibrosis (F0-F2)	73	167	PRJEB14215
		48	168	PRJNA420817
2	MASLD, initial fibrosis (F0-F2)	72	169	PRJNA373901
	MASLD, advanced fibrosis (F3-F4)	14		
3	Healthy control	105	170	PRJEB6337
	Cirrhosis MASLD	125		

Table 0-1: MASLD patient cohorts and fecal metagenomic datasets analyzed in Chapter I.

Each cohort (column 1) was divided into two phenotypic comparison groups: control vs. disease (column 2). The cohorts include *N* MASLD patients and corresponding whole-metagenomic stool samples derived from them (column 3). The disease group in Cohort 1 consists of 73¹⁶⁷ and 48¹⁶⁸ MASLD patients with initial fibrosis. Cohort 2 was divided into initial fibrosis (control) and advanced fibrosis groups. Patients were originally recruited in the listed studies (column 4), with metagenomic data and clinical metadata retrieved from the indicated SRA accession numbers (column 5).

C) Quality control of metagenomic reads

Sequencing reads were trimmed to remove Illumina adapter remnants and low-quality regions (<Q25) using BBDuk from the BBTools suite (version 37.62)¹⁷¹. Read quality was assessed with FastQC (version 0.12.1)¹⁷² and summarized using MultiQC (version 1.22.3)¹⁷³.

D) Taxonomic signatures associated with MASLD

To determine whether any bacterial clade is systematically increased or decreased in MASLD across the three cohorts, we profiled the taxonomic composition of GM communities in fecal samples using MetaPhlAn (version 4.0.6), which uses a wide-range of clade-specific marker genes to estimate taxonomic abundances³⁶. The diversity of these markers allows MetaPhlAn to provide highly accurate taxonomic classification across a wide spectrum of bacterial clades, from the species level to broader taxonomic ranks, without being limited to 16S rRNA alone.

For this study, we used the database version *mpa_vOct22_CHOCOPhlanSGB_202212* (downloaded on August 22, 2023) and added the *--ignore_eukaryotes* flag to skip profiling eukaryotic organisms. Post-processing of the MetaPhlAn output involved a) removing low-abundant clades (i.e., those with an average relative abundance of 0 in at least one cohort), b) excluding taxa predicted as “unclassified”, and c) ensuring that no removed clades were exclusive to one of the comparison groups, in order to avoid removing what could be very clear taxonomic markers. To identify the most abundant clades present in the GM, we selected genera with an average relative abundance above 0.5% (Figure III-4C) and species above 0.2% (Figure III-4E). Notably, in this step we faced an issue regarding the status of taxonomic assignments, as several species required manual updates to their nomenclature assigned by MetaPhlAn. Specifically, *Eubacterium rectale* was reassigned to *Agathobacter rectalis*¹⁷⁴, *Ruminococcus torques* to *Mediterraneibacter torques*¹⁷⁵ and *Roseburia faecis* to *Agathobacter faecis*¹⁷⁶.

E) Isolation of target gene families in the GM

We constructed families of the human GM metabolic genes encoding the enzymes involved in the production of butyrate, ethanol, propanol, TMA, methane and their precursor metabolites, as represented in Figures III-5A, 6A and 7A. The Unified Human Gastrointestinal Genome (UHGG, version 2.0.2)⁸⁴ served as the reference database for this purpose, being the

most comprehensive collection of characterized human GM genes to date (Table I-1). Each gene family was built by retrieving a seed set of sequences based on gene and protein annotation from the UHGG, assessing their similarity with lab-validated enzymatic sequences when available. Validated gene sequences were retrieved from MetaCyc⁴⁴ and used as controls to assess their phylogenetic proximity to the retrieved gene seeds.

Gene families were aligned using MAFFT (version 7.271)¹⁷⁷ with options “*--localpair*” and “*--maxiterate 1000*” and phylogenetic trees were constructed with IQ-TREE (version 2.0.3)¹⁷⁸ with the ultrafast bootstrap option (1000 bootstraps)¹⁷⁹ to assess branch support. The best fitting model for each metabolic gene was estimated using ModelFinder Plus¹⁸⁰, according to the Bayesian information criterion. Low phylogenetic distances between control and retrieved sequences served as a criterion for defining gene families. Duplicated sequences, gene fragments and phylogenetically distant genes were excluded after inspecting the trees. To improve gene identification and classification, we then built profile hidden Markov models (HMMs) for each target gene family using HMMER *hmmbuild* and *hmpress* (version 3.4)¹⁸¹. MetaCyc⁴⁴ and KEGG⁴³ were also used to validate gene and enzyme names, KO groups, the biochemical reactions involved in the formation of each metabolite, and their EC numbers.

The same protocol was applied to isolate five USCGs used as controls. These USCGs and their encoded proteins were: *argS* (arginyl-tRNA synthetase), *dnaA* (chromosomal replication initiator protein DnaA), *rpoA*, *rpoB* and *rpoC* (alpha, beta and gamma subunits of the DNA-directed RNA polymerase, respectively).

F) Analysis of gene abundance by read-based quantification

We quantified the abundance of target genes in the metagenomic samples from the three cohorts by aligning the quality-controlled reads against the isolated gene families using DIAMOND (version 2.0.14)⁷⁹. Alignments were filtered to retain only the best match for each read and gene family sequence by using the “*--max-hsps 1*” option, with an E-value threshold of <1E-10. These parameters were based on experimentally validated recommendations for detecting genes in fecal metagenomic samples with DIAMOND⁹⁹. The total number of aligned reads for each gene family was normalized using a variation of the RPKM formula, referred to as RPKSM (reads per kilobase per size per million reads) (Equation III-1). This measure was used as an indicator of gene abundance (Figures III-5B, 6B-C and 7B-D; and Supplementary Figures S-III-3 and 4). It is defined as the number of aligned reads per kilobase per gene family size per

million reads, to account for variations in gene length, gene family size (i.e., the number of sequences within each gene family) and library size, respectively.

$$RPKSM = \frac{\#Reads\ aligned \times 10^3 \times 10^3 \times 10^6}{\#Total\ reads \times Gene\ length \times \#Genes}$$

Equation III-1: RPKSM formula.

G) Whole metagenomic analysis

We applied a whole-metagenome analysis workflow to validate our results at single-gene level. To this end, SqueezeMeta (version 1.6.0)⁷⁶ was used to predict both the taxonomic and functional composition of the metagenomic samples from the three cohorts. Due to computational limitations, it was not feasible to co-assemble all available samples, so approximately 24-30 samples per cohort were randomly selected, with equal representation from each phenotypic group. Briefly, SqueezeMeta assembled the metagenomes from each sub-cohort into a single co-assembly using MEGAHIT (version 1.2.9)⁴⁸ by pooling the sequencing reads from every sample. ORFs were then predicted from contigs⁵⁰, and reads from individual samples were mapped back against the co-assembly to quantify the abundance of each KO group per metagenome. We further processed these results with SQMtools (version 1.6.0)⁸² to identify the most differentially abundant KO groups between phenotypic groups within each cohort using DESeq2 (version 1.34.0)¹⁴², applying a log fold-change >2 and a p-adjusted value <0.05 as thresholds (Supplementary Figure S-III-5).

H) Presence of candidate genes in human GM genomes and plasmids

To determine whether candidate metabolic genes involved in MASLD were core or accessory across the human GM, their presence was quantified in both UHGG genomes and curated RefSeq200 plasmids. Curation involved eliminating sequences corresponding to partial plasmid DNA sequences, unassignable hosts, or PacBio internal control sequences, as done previously¹⁶⁵. Since many UHGG genomes originate from metagenome-assembled or single-cell-amplified genomes, they may be incomplete, potentially leading to false positive annotations. To minimize this risk, we included only UHGG genomes with >95% completeness and containing the five USCGs described above. ORFs were predicted using Prodigal (version 2.6.3)¹⁸², retaining only those covering >90% of the HMM size. We then used HMMER *hmmsearch* (version 3.4)¹⁸¹ with the following parameters: protein-pHMM coverage threshold >90%, E-value <0.01 and independent E-value (i-E-value) <0.01, strictly enforcing the 90% threshold to avoid

capturing ORFs with similar protein domains. This approach yielded 31,227 genomes classified as complete and containing the five USCGs.

To classify genes as core or accessory, we defined their presence as the percentage of strains within a given genus and species that encode the gene. Genes were categorized as core if present in >80% of congeneric (Figure III-8A) or conspecific (Figure III-8B) genomes, as accessory if present in 20-80%, and as highly accessory if present in <20% genomes. Only the most abundant clades in the GM with >100 genomes meeting the completeness criteria were included. Additionally, we assessed whether these genes could be encoded in annotated plasmids. Using the same constraints, we predicted ORFs from the RefSeq200 plasmid collection, which includes 23,309 plasmids encoding over 2 million ORFs. ORFs were retained if they met the >90% HMM size threshold, with *hmmScan* parameters identical to those applied in the UHGG genome analysis (Figure III-8C).

I) R packages and statistical analysis

Data manipulation was performed using R (version 4.1.3) and the *tidyverse* package (version 1.3.2)¹⁸³. Figures were generated with the R packages *ggplot2* (version 3.4.2)¹⁸⁴, *ggpubr* (version 0.6.0)¹⁸⁵ and *pheatmap* (version 1.0.12)¹⁸⁶. Mann-Whitney tests with a Benjamini-Hochberg adjustment for multiple testing were applied on the relative abundance of bacterial clades (Figures III-4D and E, Supplementary Figure S-III-2) and gene abundances (Figures III-5B, 6B-C and 7B-D; and Supplementary Figures S-III-3 and 4) between phenotypic groups in the metagenomic fecal samples of each cohort to assess the significance of the differences. Statistical analyses and graphical representation on the plots were conducted using the R package *rstatix* (version 0.7.2)¹⁸⁷.

3. Results

A) *Agathobacter rectalis* is consistently depleted in MASLD

Certain bacterial clades have been positively correlated with MASLD progression, including *Enterobacteriaceae*^{164,188} such as *Escherichia / Shigella spp.*^{169,188}, *Bacteroidaceae* like *Bacteroides spp.*^{169,189}, *Veillonellaceae*¹⁶⁴ or *Streptococcus spp.*¹⁹⁰. Conversely, other clades have been reported to exhibit an inverse correlation with MASLD progression, including *Ruminococcaceae*¹⁶⁴ such as *Faecalibacterium prausnitzii*^{169,191}, *Lachnospiraceae* like *Eubacterium rectale*^{169,191} or *Dorea longicatena*¹⁹¹, *Clostridiaceae* like *Clostridium spp.*¹⁸⁸, and

even *Pseudomonas spp.*¹⁹⁰. However, these studies often yield conflicting results. For example, a reported positive correlation between *Ruminococcus spp.* abundance and MASLD progression¹⁸⁹ contradicts one of the main findings in two posterior studies^{169,191}. Similarly, *Prevotella spp.* was found to decrease with MASLD development¹⁹² and fibrosis stage¹⁸⁹, yet another study reported an increase in this genus in MASLD patients¹⁹⁰. There are additional examples of discordant taxonomic signatures in MASLD, such as *Blautia spp.* or *Roseburia spp.*¹⁹³.

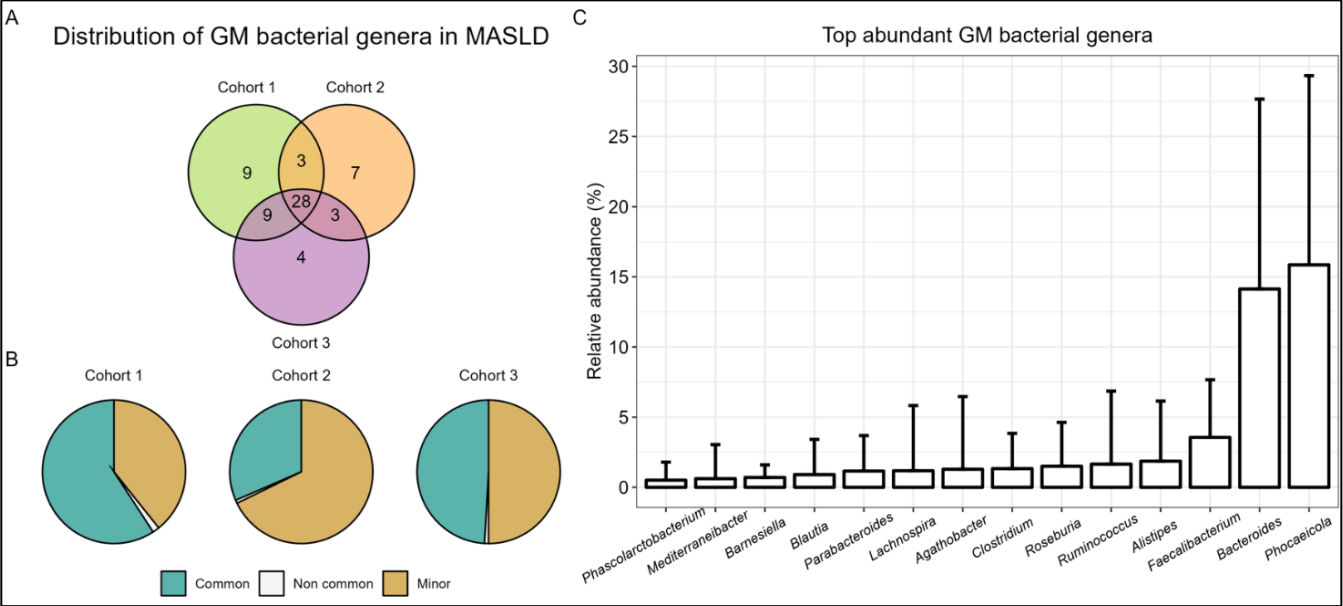
To investigate whether any bacterial clade is systematically altered in MASLD, we profiled the taxonomic composition of the GM communities across all samples from the three cohorts at both genus and species levels using MetaPhlAn³⁶, as detailed in Materials and Methods. This analysis identified 28 bacterial genera shared across all three cohorts (Figure III-4A), which accounted for 59%, 31.5% and 49% of the GM composition in each cohort, respectively (Figure III-4B). Thus, nearly half of the bacterial genera identified in at least one cohort were shared across all three (28 out of 63), representing between one-third and two-thirds of the total GM abundance.

Among these shared genera, *Gemmiger*, *Streptococcus*, *Mediterraneibacter*, *Ruminococcus* and *Agathobacter* exhibited a consistent pattern of differential abundance between MASLD and control groups across all cohorts, reaching statistical significance in at least two (Figure III-4D). This exclusion criterion was applied to enhance the reliability of the results and minimize confounding factors from cohort-specific demographic variables (e.g. geographical location, age, sex, or ethnicity) or environmental factors (e.g. diet, physical activity, comorbidities, or drug exposure), rather than being directly related to the MASLD phenotype^{162,194}. Cohort 2 was exempt from this threshold as it only included MASLD patients, and applying it could have been overly restrictive, masking potential taxonomic differences associated with the disease.

Strikingly, *Agathobacter* was the only genus to show significant depletion across all three cohorts (Figure III-4D). This effect was observed in MASLD patients with initial fibrosis (F0-F2, Cohort 1), advanced fibrosis (F3-F4, Cohort 2), and cirrhosis (Cohort 3) compared to healthy controls or earlier disease stages (Cohort 2). Note that *Agathobacter*, along with *Ruminococcus* and *Mediterraneibacter*, belongs to the fraction of the highest abundant GM

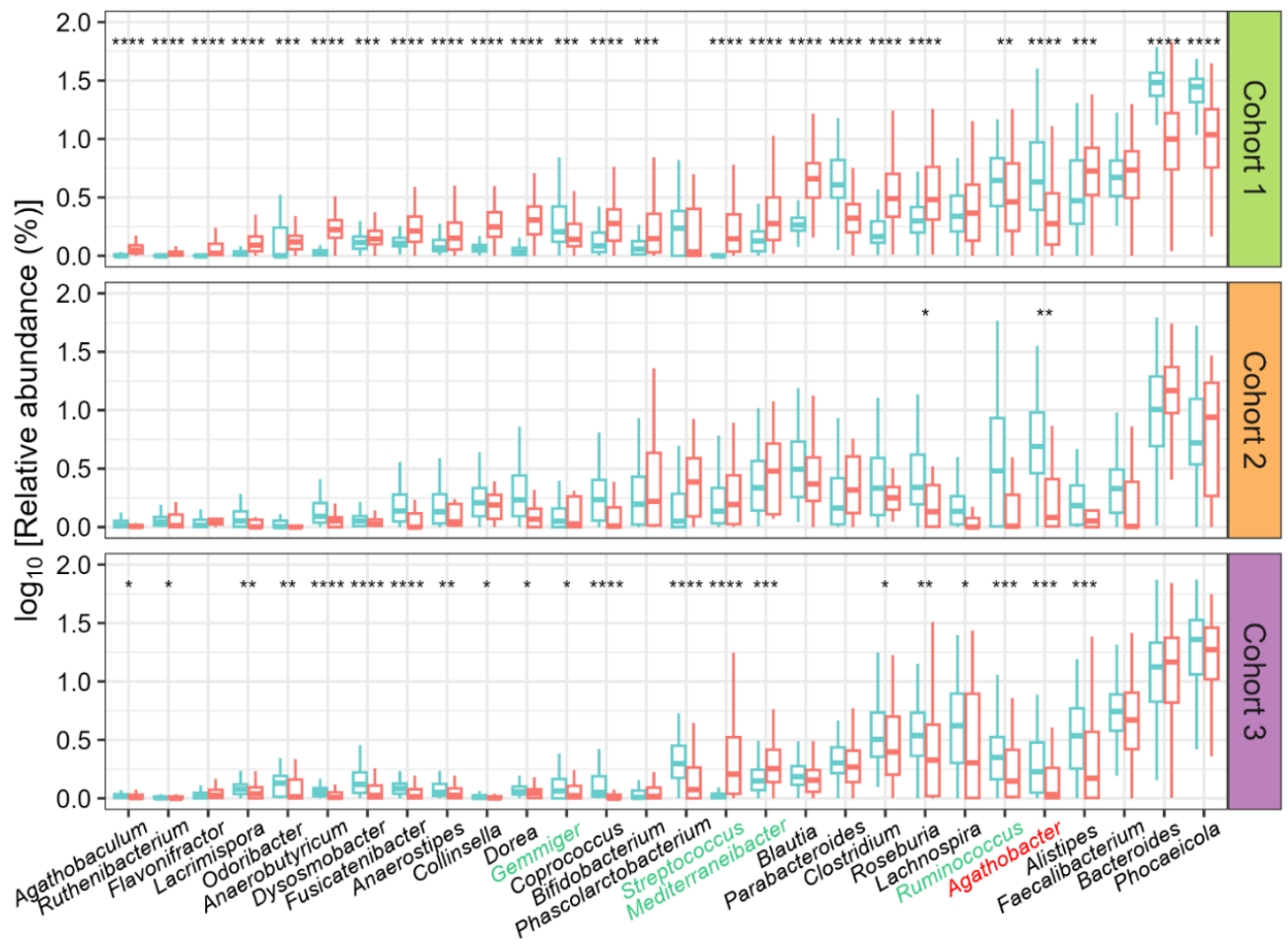
bacterial genera (Figure III-4C), highlighting their potential contribution to the phenotype in specific patient cohorts.

Analogously, the most abundant bacterial species within the GM were analyzed for differences in abundances between groups across the three cohorts. Only *Agathobacter rectalis* and *Bacteroides uniformis*, two of the most abundant GM bacterial species (Supplementary Figure S-III-1), exhibited a consistent pattern of differential abundance across all cohorts, with statistical significance in at least two (Figure III-4E). Consistently, *A. rectalis* was significantly depleted in MASLD, identifying it as a robust, cross-cohort taxonomic indicator of MASLD severity (Supplementary Figure S-III-2).



D

Abundance of GM bacterial genera in MASLD



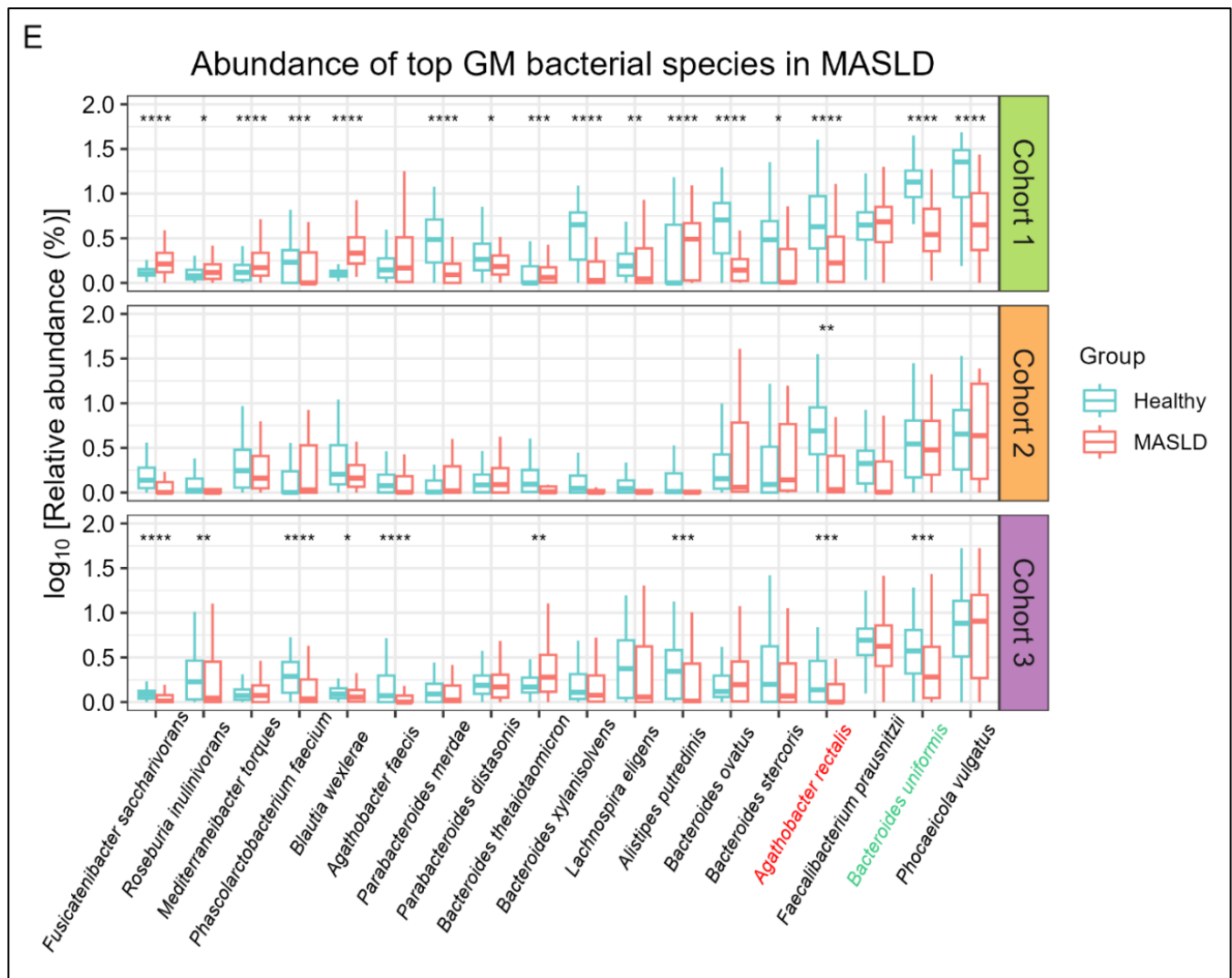


Figure 0-4: Taxonomic profiling of the GM in MASLD.

(A) Number of bacterial genera shared across the three study cohorts. **(B)** Average relative abundance of bacterial genera predicted by MetaPhlAn. The blue section of the pie charts represents the average relative abundance of genera common to the three cohorts ($n=28$), whereas the grey and green sections reflect, respectively, cohort-specific and minor genera (i.e., with an average relative abundance of 0 in at least one of the cohorts). **(C)** Most abundant GM genera (average absolute abundance $>0.5\%$) **(D-E)** Relative abundance of bacterial genera **(D)** and species **(E)** in fecal metagenomes, shown on a logarithmic scale. Blue and red boxplots indicate healthy and MASLD groups, respectively. Each facet represents one cohort. Differences in abundance were evaluated using pairwise Mann-Whitney tests with Benjamini-Hochberg adjustment for multiple testing. Horizontal bars represent median values, while boxes and whiskers represent the interquartile range and the full data spread (minimum to maximum), respectively. Clades depicted in green across the x-axis show consistent differential abundance trends between groups over cohorts. *Agathobacter rectalis*, depicted in red, is the only species with a statistically significant and consistent depletion in MASLD across all cohorts. (*) $p<0.05$, (**) $p<0.01$, (***) $p<0.001$, (****) $p<0.0001$. Non-starred genera present non-significant differences in abundance between groups.

B) Butyrate-producing genes are depleted in MASLD

The identification of *A. rectalis* as a consistent taxonomic signature across multiple MASLD cohorts underscores its potential protective role in MASLD development. However, taxonomy alone may not fully capture the functional contributions of the GM. Certain microbial metabolites regulate GLA integrity and have been suggested to influence MASLD progression^{195–197}. These compounds derive from dietary carbohydrates and proteins that are not fully digested in the upper gastrointestinal tract and are instead catabolized by the GM through fermentative pathways that are predominant in the anaerobic environment of the colon^{198,199}. Once absorbed into the bloodstream, they are transported to the liver via the portal vein, where they can exert both beneficial and detrimental effects^{152,200}.

The most well-characterized of these metabolites are short chain fatty acids, produced by the colonic fermentation of complex carbohydrates present in dietary fiber that escape digestion in the small intestine. These metabolic pathways have been relatively well understood for nearly thirty years²⁰¹ and their impact on GM homeostasis has been reviewed in several articles^{202–204}. For this reason, we interrogated the metagenomic fecal samples from the patient cohorts for metabolic signatures that could explain the MASLD phenotype. First, we focused on the genes encoding enzymes involved in butyrate formation as a proxy for the bacteria that carry them, independently of their taxa. To validate our hypothesis, we analyzed the genic abundance of the final enzymes involved in butyrate production.

We selected butyrate as a target because it is crucial for maintaining GLA homeostasis. It supports colonic epithelial integrity by upregulating the expression of mucin-coding genes²⁰⁵ and tight-junction proteins like Claudin-1²⁰⁶, which improve the gut barrier function and prevent the translocation of toxic compounds to the liver. In fact, butyrate is the primary energy source for colonocytes²⁰⁷. It also modulates the gut immune function by inducing the development and differentiation of anti-inflammatory, colonic regulatory T cells through the acetylation of histone H3 at the *Foxp3* promoter²⁰⁸ and the activation of the cell surface receptor GPR109A in colonic macrophages and dendritic cells, which promotes the secretion of IL-10 in the colon²⁰⁹. Specifically, butyrate enhances the bactericidal capacity of macrophages²¹⁰ by increasing the production of antimicrobial peptides while decreasing proinflammatory cytokines IL-6, IL-12²¹¹ and IL-17²¹². Additionally, butyrate increases energy expenditure by stimulating lipid oxidation through the activation of brown adipose tissue^{213,214}.

Specifically, this activation is modulated through two thermogenesis-related proteins: the peroxisome proliferator-activated receptor- γ and the mitochondrial uncoupling protein 2²¹⁵. Butyrate also induces hepatic gluconeogenesis, reduces total body fat accumulation²¹⁶ and prevents diet-induced insulin resistance²¹⁷, all of which are key risk factors in MASLD.

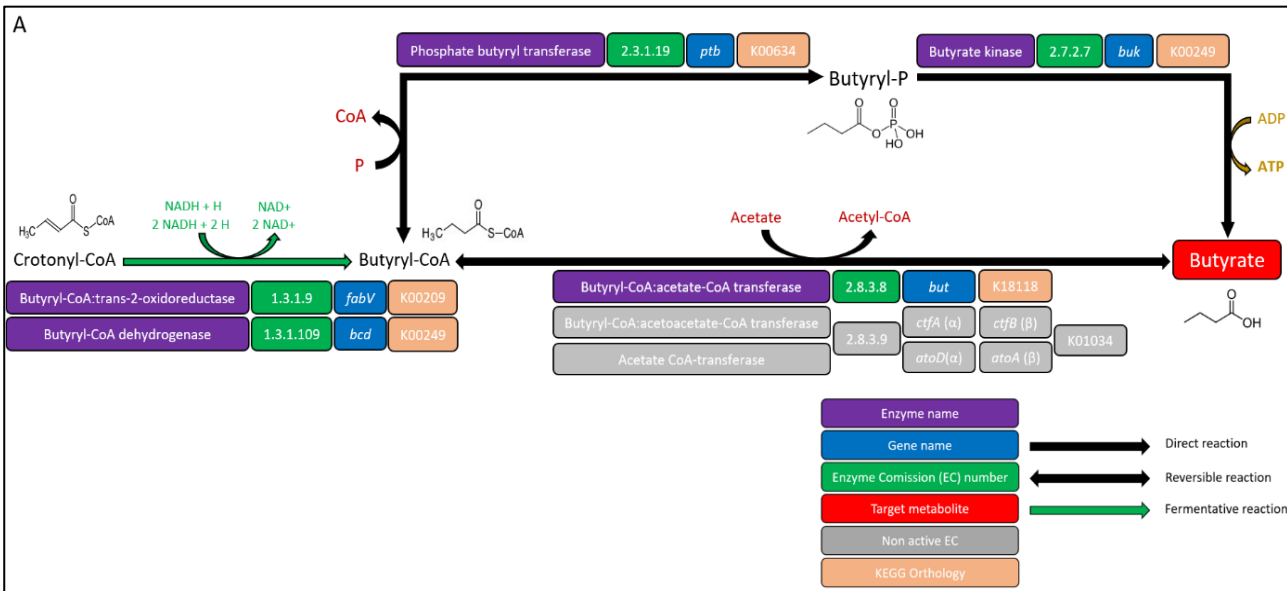
Recent *in vivo* investigations have ascribed a protective role to butyrate in MASLD. A butyrate-producing probiotic strain of *Clostridium butyricum* was shown to prevent MASLD progression in rats by restoring the expression of tight-junction proteins, as well as by decreasing steatosis, insulin resistance, serum endotoxin levels and hepatic inflammatory indexes²¹⁸. Specifically, butyrate suppresses hepatic oxidative stress by inducing the expression of transcription factors Nrf2²¹⁸ and Foxo3a²¹⁹, which regulate the production of antioxidant peptides like glutathione. Administration of sodium butyrate to mice protects them against diet-induced MASH by decreasing steatosis and hepatic inflammation²²⁰. This result replicated in a subsequent study, where authors observed that the expression of satiety hormone glucagon-like peptide-1 was downregulated in the liver of mice with MASLD, and that this effect could be reversed through the administration of sodium butyrate, concomitant with a reduction in the degree of steatosis^{221,222}.

A recent experiment with fecal microbiota transplantation (FMT) from healthy to MASH mice increased butyrate concentrations in their cecum, as well as increased levels of the intestinal tight junction protein Zonula occludens-1. This effect was accompanied by reduced steatosis, lower intrahepatic pro-inflammatory cytokines IFN- γ and IL-17, and expression of FOXP3²²³. Also, two clinical trials involving FMT from healthy human donors to obese recipients with metabolic syndrome, but without MASLD, reported increased insulin sensitivity and a rise in butyrate-producing bacterial species in the recipient's gut, such as *Roseburia intestinalis*, *Eubacterium halii*²²⁴, *Butyrivibrio* spp., *Clostridium symbiosum* and *Eubacterium* spp.²²⁵.

Bacterial butyrate can be formed through two major terminal reactions (Figure III-5A). The first involves a dephosphorylation of butyryl-P, yielding one ATP through a reaction catalyzed by a butyrate kinase encoded by *buk*. In the precursor reaction, butyryl-P is formed through a phosphorylation of butyryl-CoA catalyzed by a phosphate butyryl transferase encoded by *ptb*. Butyrate also accumulates after the transfer of a CoA cofactor from butyryl-CoA to acetate through butyryl-CoA:acetate CoA transferase, encoded by *but*. This CoA-transferase pathway also conserves the energy of the CoA bond in the formed CoA-moiety of

the co-substrate. Butyryl-CoA is produced from crotonyl-CoA through two different enzymes: butyryl-CoA trans-2-oxidoreductase, encoded by *fabV*, and butyryl-CoA dehydrogenase, encoded by *bcd*. This pathway is further reviewed in the human GM by Louis *et al.*²²⁶.

We quantified the abundance of these genes in the metagenomic fecal samples of both groups across the three MASLD cohorts, as described in Materials and Methods. Our analysis revealed that genes involved in the production of butyrate and its precursors, butyryl-P and butyryl-CoA, were significantly depleted in the GM of patients with advanced MASLD (Figure III-5B). Specifically, *bcd*, *but*, and *fabV* were consistently depleted in patients with advanced MASLD compared to those with early-stage disease in Cohort 2, and they were depleted in cirrhosis patients compared to healthy controls. Among these, the gene *fabV* was nearly 10-fold less abundant than *bcd* and *ptb* in both cohorts, indicating a minor contribution of this pathway in butyryl-CoA synthesis relative to the dehydrogenase route. Additionally, *ptb* did not exhibit significant differences in Cohort 2. Furthermore, gene *but* was significantly depleted in the control group in both cohorts, whereas *buk* did not show differences in Cohort 2 either. Notably, no depletion of butyrate genes was observed in Cohort 1, suggesting that functional disruption becomes more apparent with disease progression.



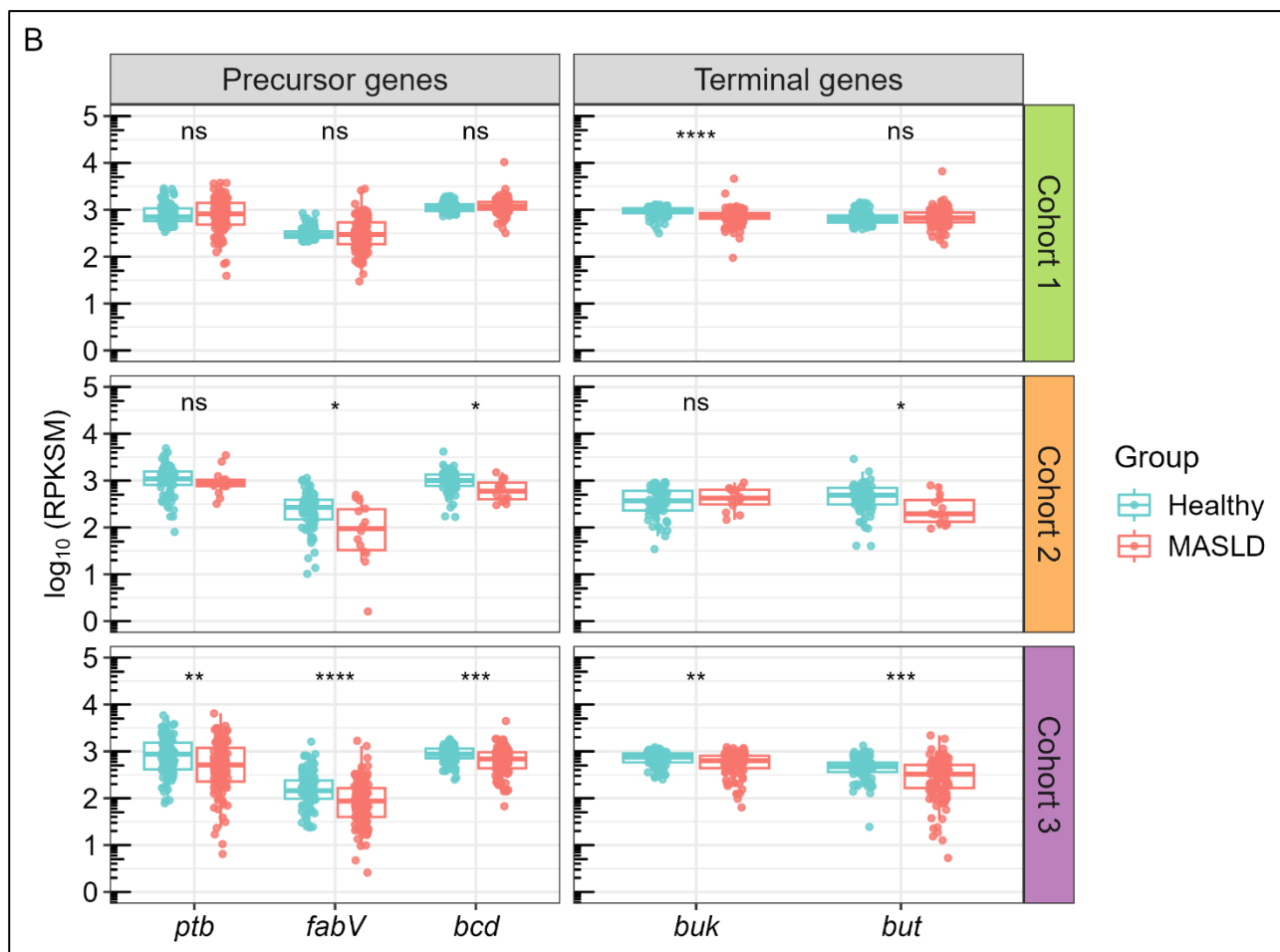


Figure 0-5: Abundance of butyrate-producing genes in MASLD.

(A) Characterization of the terminal reactions involved in the formation of bacterial butyrate from crotonyl-CoA. Enzymes, EC numbers, coding genes and KO groups, as well as reaction types, are indicated for each metabolic step according to the legend (lower part). **(B)** Abundance of butyrate-producing genes in fecal metagenomes from the three cohorts, shown on a logarithmic scale. Boxplots represent RPKSM values. Differences in abundances were evaluated using pairwise Mann-Whitney tests with Benjamini-Hochberg adjustment for multiple testing. Horizontal facets represent cohorts. Left and right facets represent abundances of precursor and terminal coding genes, respectively, as described in (A). RPKSM: reads per kilobase per genic size per million reads. (*) $p < 0.05$, (**) $p < 0.01$, (***) $p < 0.001$, (****) $p < 0.0001$, ns: non-significant.

C) SCA-producing genes are increased in MASLD

Short-chain alcohols (SCAs) are a metabolic family also suggested to regulate GLA integrity, with several studies linking GM endogenous alcohol production to MASLD. One of the first reports, published over a decade ago, described elevated peripheral blood ethanol and increased fecal *Escherichia spp.* 16S rRNA levels in patients with MASH²²⁷. Similar findings were reported in a pediatric cohort, where children with MASLD exhibited higher levels of fecal

ethanol, *Prevotella spp.*, and Gammaproteobacteria 16S rRNA²²⁸. A more recent study showed that postprandial peripheral blood ethanol and fecal *Lactobacillaceae* 16S rRNA levels increased in MASH patients after selective inhibition of host ADH²²⁹. This effect was reversed with broad-spectrum antibiotic treatment.

In 2019, Yuan and colleagues identified a high-alcohol-producing strain of *Klebsiella pneumoniae* capable of inducing MASLD through ethanol production²³⁰. Specifically, they found that FMT from a MASH patient containing this strain induced MASLD in healthy mice, whereas removing the strain before FMT prevented disease development. Furthermore, ethanol produced by this *K. pneumoniae* strain was shown to disrupt lipid homeostasis in hepatic cells and to cause mitochondrial dysfunction, oxidative stress, and the accumulation of reactive oxygen species²³¹. These factors are critical to MASH development through their role in lipid peroxidation and inflammation²³². Additionally, a recent computational study identified a positive correlation between the abundance of two fermentative pathways in certain families from class Clostridia and fatty-liver disease in patients²³³. These pathways include the production of ethanol from pyruvate and the heterolactic fermentative pathway.

With this ample evidence, we hypothesized that genes encoding enzymes involved in the formation of endogenous SCAs may represent an additional metabolic signature contributing to MASLD progression. We focused on ethanol and propanol, which are produced by the reduction of their respective aldehydes through ethanol and propanol dehydrogenases. These enzymes help to reduce the levels of toxic aldehydes in the GM^{234,235} and replenish NAD⁺ levels to maintain fermentative growth.

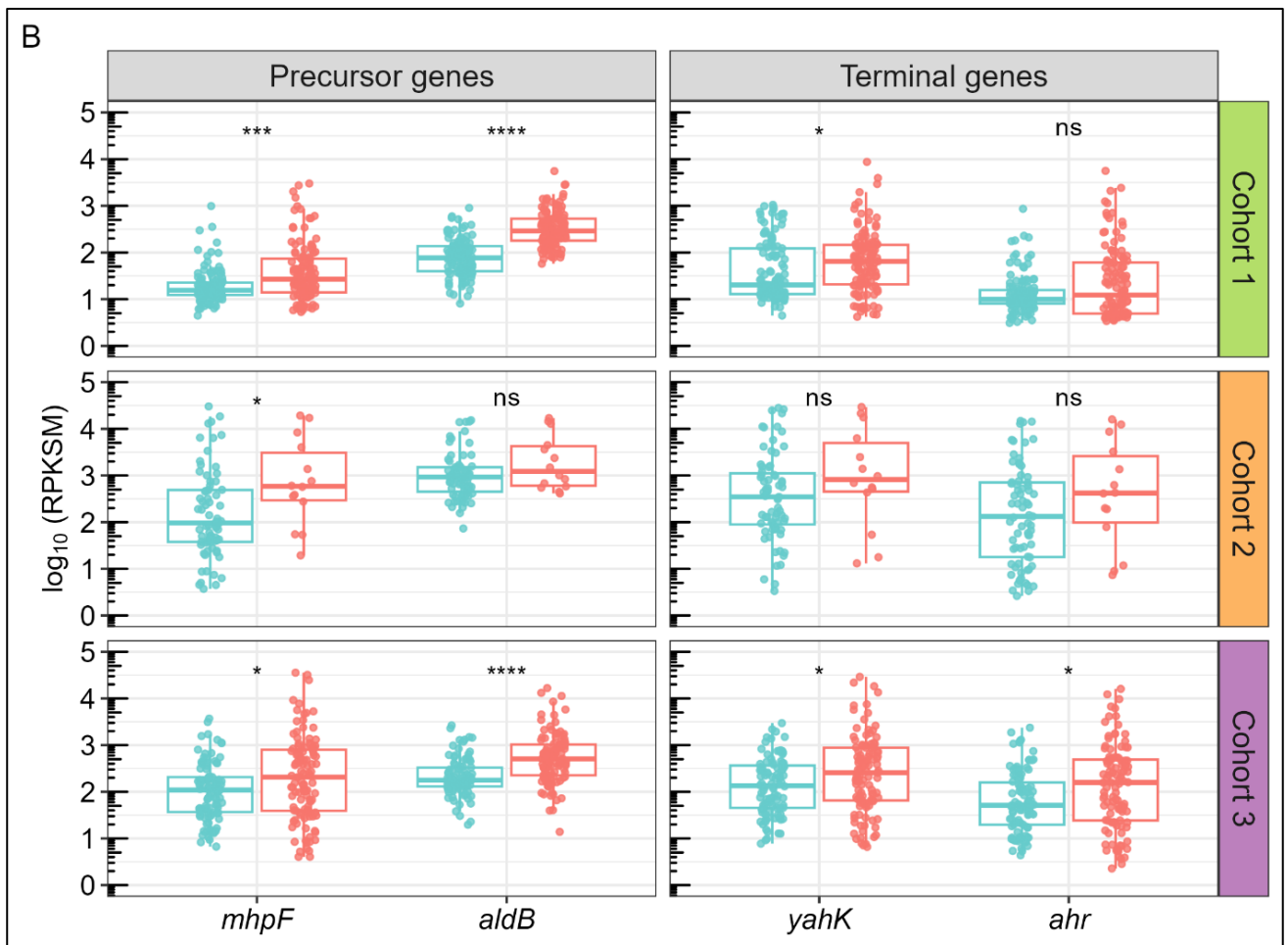
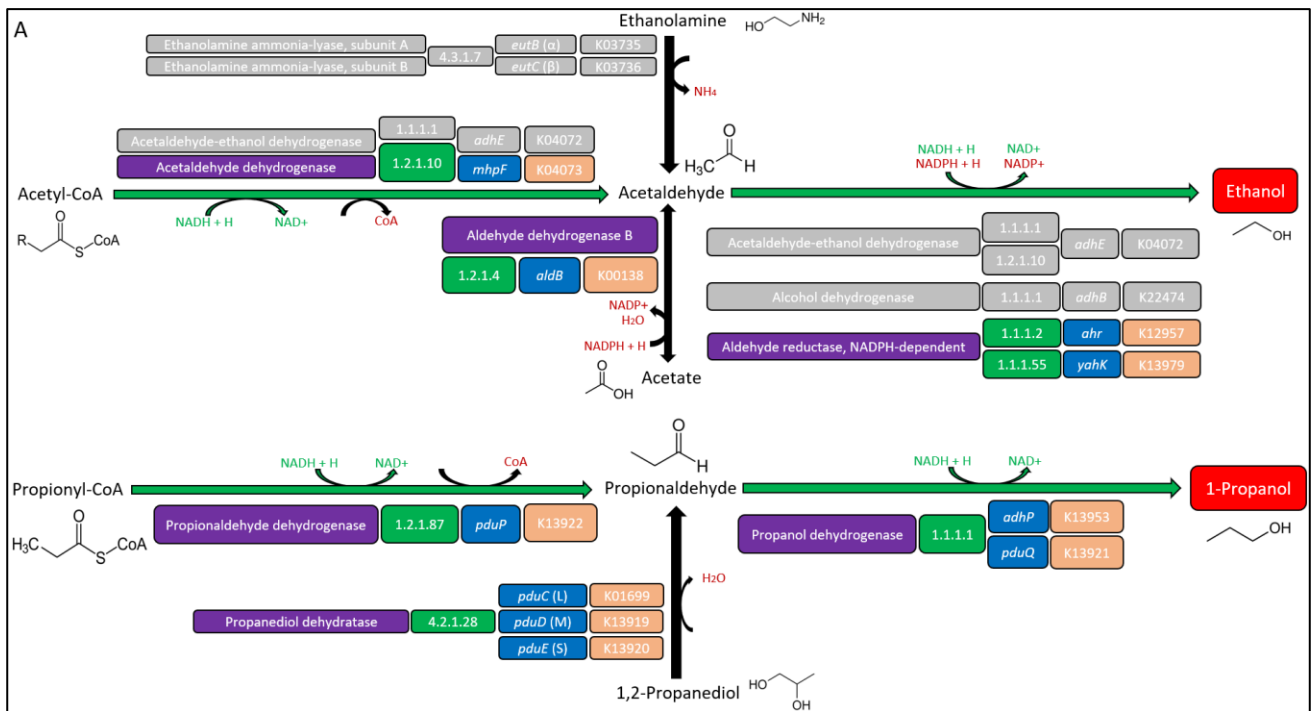
Bacterial ethanol is formed from acetaldehyde through a fermentative reaction that can be catalyzed by 1) an NADH-dependent, bifunctional acetaldehyde-ethanol dehydrogenase encoded by *adhE*, 2) an NADPH-dependent alcohol dehydrogenase encoded by *adhB*^{236,237} and 3) a NADPH-dependent aldehyde reductase encoded by two different genes: *ahr* and *yahK*, with the corresponding liberation of phosphorylated reducing power in form of NADP⁺ ^{238,239}. The precursor acetaldehyde can be formed from 1) ethanolamine degradation, through a reaction catalyzed by ethanolamine ammonia-lyase which is encoded by *eutB* and *eutC* genes and liberates ammonia as a by-product²⁴⁰, 2) acetyl-CoA, through a fermentative reaction catalyzed by bifunctional AdhE and by a CoA-acetylating acetaldehyde dehydrogenase encoded by

mhpF^{241,242}, and 3) acetate, through a dehydration catalyzed by aldehyde dehydrogenase B, encoded by *aldB*²⁴³ (Figure III-6A, top).

Bacterial 1-propanol is formed from propionaldehyde through a fermentative reaction catalyzed by a propanol dehydrogenase encoded by two different genes: *adhP* (a propanol-preferring alcohol dehydrogenase) and *pduQ*. Propionaldehyde can be generated through two distinct reactions: 1) a CoA transference from propionyl-CoA catalyzed by a propionaldehyde dehydrogenase encoded by *pduP*; and 2) a dehydration of 1,2 propanediol catalyzed by a propanediol dehydratase composed by three subunits encoded by *pduC*, *pduD* and *pduE*. These genes are co-located within the same operon and encode the corresponding enzymes responsible for the metabolism of 1,2-propanediol within the propanediol-utilizing (Pdu) microcompartment, a bacterial organelle found in certain enteric species^{244,245} (Figure III-6A, bottom).

We quantified the abundance of genes involved in bacterial SCA formation in the metagenomic samples of the three cohorts as described above. The metagenomic profiling revealed that genes *mhpF* and *aldB*, involved in ethanol production, and *yahK* and *ahr*, involved in acetaldehyde production, were significantly enriched in the GM of MASLD patients from Cohort 1 and Cohort 3 relative to healthy controls. Their abundance also increased with fibrosis progression (Figure III-6B). Notably, *mhpF* and *ahr* were depleted compared to *aldB* and *yahK* respectively in all three cohorts.

All genes involved in the formation of 1-propanol and propionaldehyde were also significantly enriched in MASLD patients by more than half order of magnitude, and they displayed a stepwise increase with fibrosis severity (Figure III-6C, Supplementary Figure S-III-5A). Together, these findings reveal a systematic and stage-associated enrichment of genes involved in the microbial production of ethanol and propanol in MASLD.



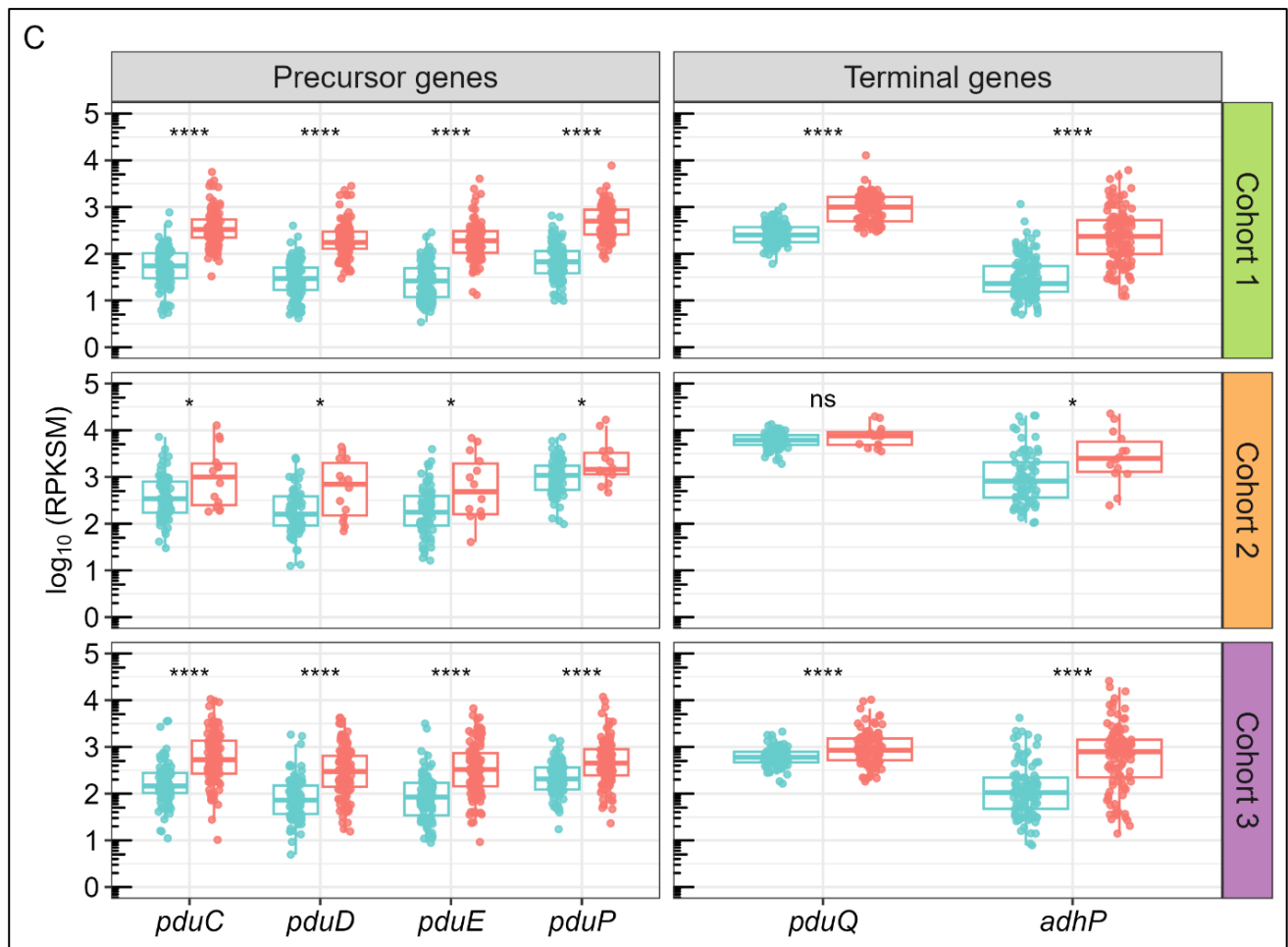


Figure 0-6: Abundance of SCA-producing genes in MASLD.

(A) Characterization of the terminal reactions involved in the formation of bacterial ethanol (top) and propanol (bottom). Enzymes, EC numbers, coding genes and KO groups, as well as reaction types, are indicated for each metabolic step according to the legend in Figure III-5A. Non-candidate genes are colored in grey. **(B)** Abundance of *mhpF*, *aldB*, *yahK* and *ahr* genes involved in final reactions from pyruvate to ethanol across the three cohorts. **(C)** Abundance of *pduC-E*, *pduP*, *pduQ* and *adhP* genes, involved in the last reactions from 1,2-propanediol and propionyl-CoA to 1-propanol. Boxplot layout, distribution in facets and statistical tests were performed as in Figure III-4B.

D) Genes involved in methane production are decreased in MASLD, whereas *tor* operons driving TMA accumulation are elevated

Trimethylamine N-oxide (TMAO) is a metabolite derived from the oxidation of trimethylamine (TMA), which is produced by the microbial degradation of dietary compounds such as choline and L-carnitine. Circulating TMAO levels have also been linked to MASLD severity²⁴⁶ and all-cause mortality in MASLD patients²⁴⁷. Higher plasma TMAO levels also increase cardiovascular disease risk -a condition associated with MASLD- in both mice²⁴⁸ and

clinical patients²⁴⁹. Therefore, we investigated whether microbial genes involved in TMA and TMAO metabolism might reveal additional functional disruptions contributing to MASLD pathophysiology.

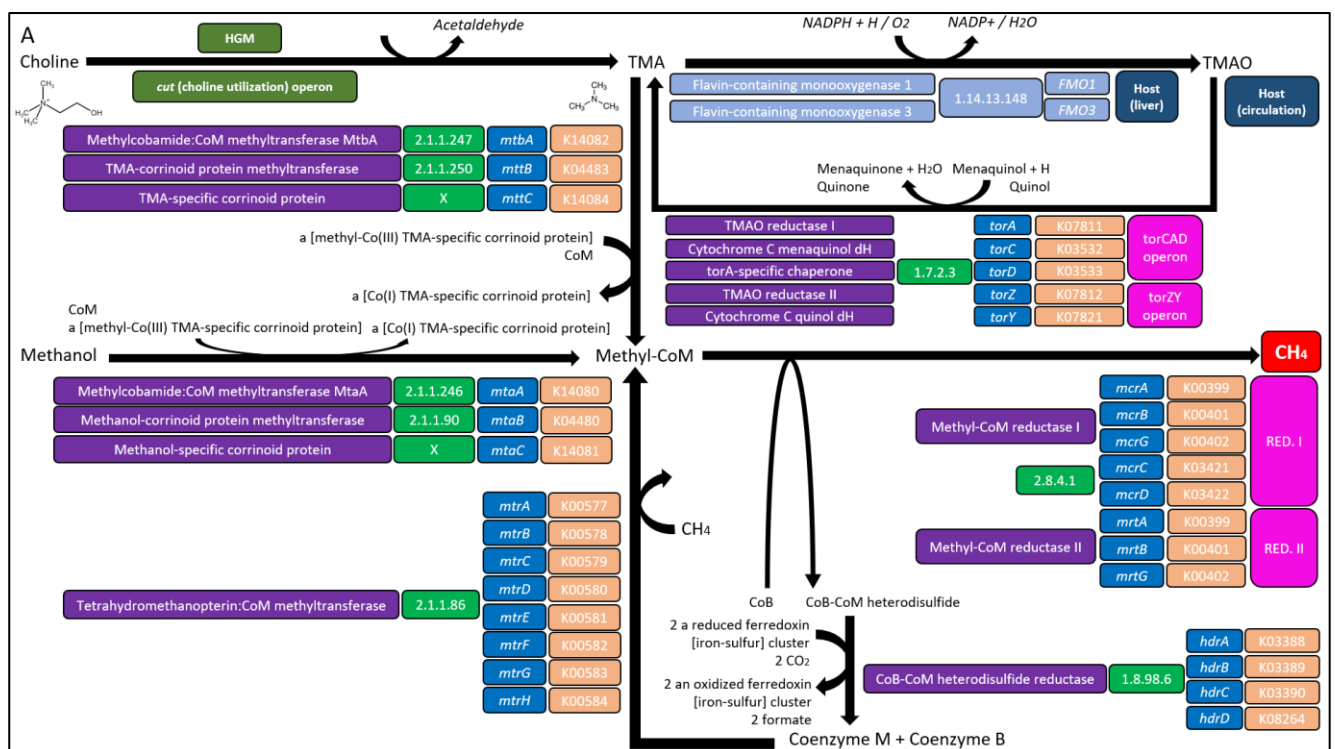
Methane production in the GM occurs through a two-step process: a simultaneous TMA demethylation and CoM methylation, followed by the reduction of methyl-CoM (Figure III-7A). TMA originates from choline degradation via the *cut* operon²⁵⁰ and is oxidized in the liver to TMAO by hepatic flavin-containing monooxygenases²⁵¹. Under anaerobic conditions, TMAO can be reduced back to TMA through TMAO reductase I, encoded by *torA*. This enzyme receives electrons from a membrane-bound cytochrome c menaquinol dehydrogenase encoded by *torC*. Together with *torD*, which encodes a chaperone that protects and matures TMAO reductase I, these genes form the *torCAD* operon, inducible by TMAO. A second anaerobic respiratory system, *torZY*, operates at a lower constitutive expression level and can also use TMAO as an electron acceptor, although it is not TMAO-inducible. In this system, *torZ* encodes TMAO reductase II and *torY* encodes a cytochrome c quinol dehydrogenase. The molecular mechanisms of *torCAD* and *torZY* operons are further reviewed by Leimkühler *et al.*²⁵².

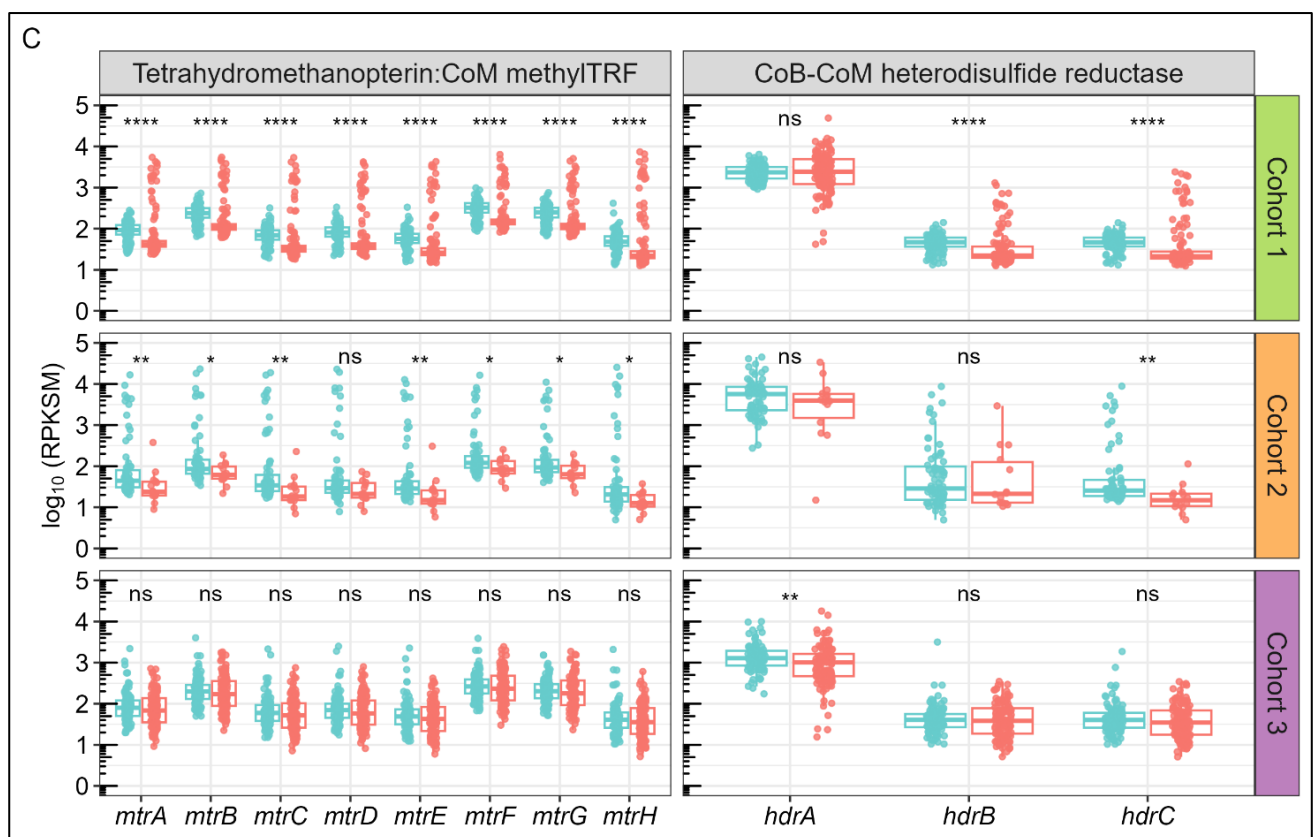
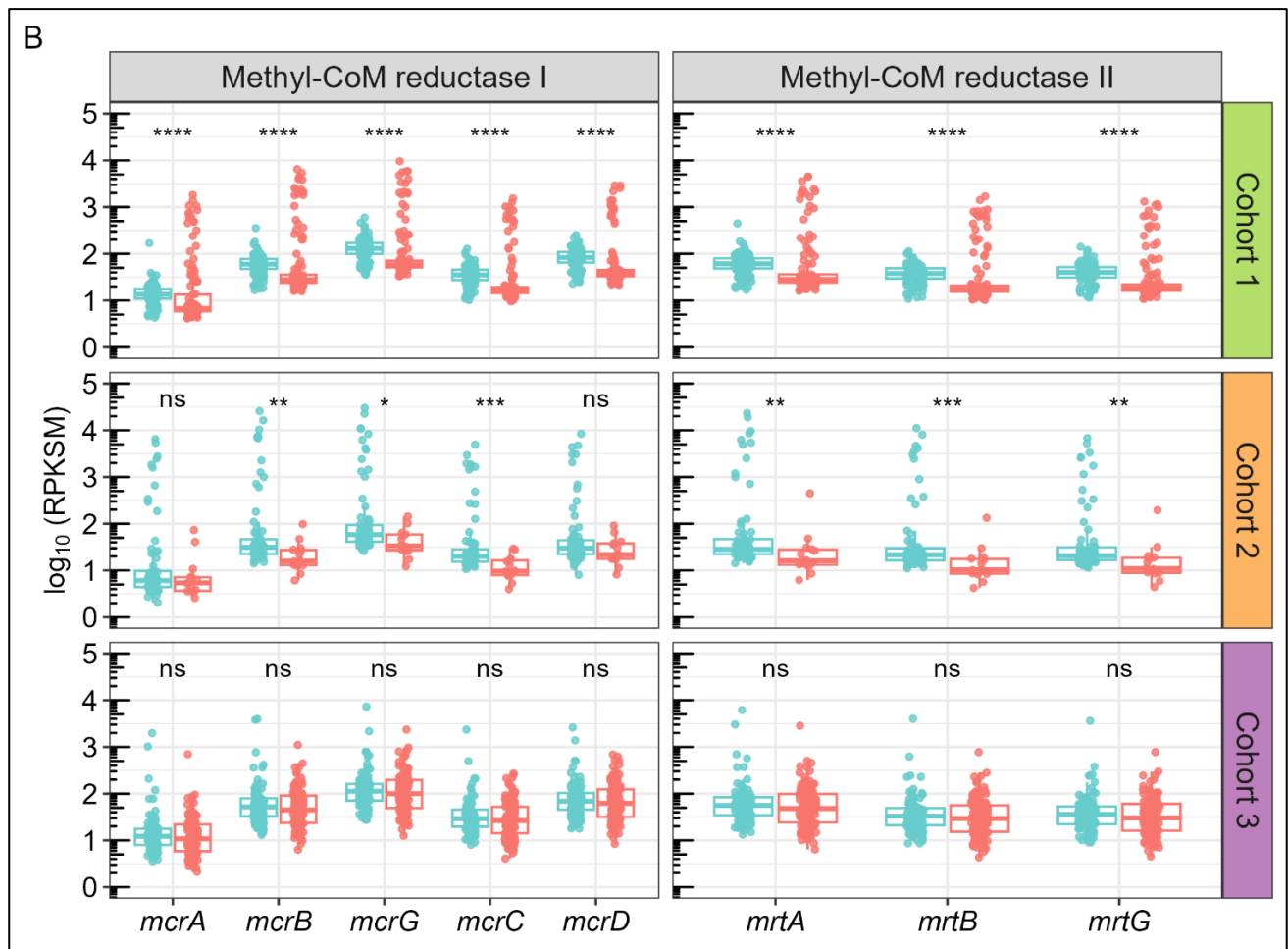
CoM methylation occurs through two well-characterized enzymatic systems: a TMA-specific (composed of *mttC*, *mtbA* and *mttB*) and a methanol-specific pathway (composed of *mtaC*, *mtaA* and *mtaB*)^{253–255}. The final reduction of methyl-CoM to methane is catalyzed by methyl-CoM reductase (MCR) I, which is encoded by the *mcrBDCGA* operon²⁵⁶. MCR I has an alternative isozyme termed MCR II, encoded by *mrtBGA* operon²⁵⁷. The resulting heterodisulfide (CoB-S-S-CoM) is then reduced to CoM and CoB by a heterodisulfide reductase encoded by operon *hdrABC*, thereby regenerating both cofactors²⁵⁸. Additionally, methyl-CoM is replenished via tetrahydromethanopterin CoM-methyltransferase (encoded by operon *mtrABCDEFGH*), which methylates CoM²⁵⁹.

Metagenomic analysis revealed that genes encoding MCR I and II, which catalyze the reduction of methyl-CoM to methane, were significantly decreased in the GM of MASLD patients. Their abundance also declined with disease progression (Figure III-7B). Concordantly, *hdr* and *mtr* genes, responsible for replenishing CoM and transferring it to a methyl group to form the methyl-CoM precursor, were significantly decreased in MASLD and showed a progressive depletion with MASLD advancement (Figure III-7C). Interestingly, *hdrA* was two orders of magnitude more abundant than the rest of the genes from the operon. Genes *mtbA*,

mttB, *mtaA* and *mtaB* were significantly decreased in MASLD. However, they did not exhibit differences between the initial and advanced stages of MASLD (Supplementary Figure S-III-3). Notably, the higher abundance of methane-producing genes observed in the healthy group was not replicated in Cohort 3. Conversely, bacterial genes involved in TMA regeneration showed the opposite trend. The *torCAD* and *torZY* operons were significantly increased in the GM of MASLD patients, and their abundance increased with fibrosis severity (Figure III-7D, Supplementary Figure S-III-5B). These findings suggest a metabolic rerouting in the GM favoring TMA regeneration over methane synthesis, which may contribute to increased hepatic TMAO load and pro-inflammatory signaling in MASLD.

Finally, we quantified the abundance of five USCGs, as described in Materials and Methods, to confirm that observed differences in genic abundance between sample groups were truly associated with the MASLD phenotype and not confounded by biological variability or spurious technical artifacts related to the quantification method. No significant differences were found in gene abundance across groups for any USCG in any of the three cohorts (Supplementary Figure S-III-4).





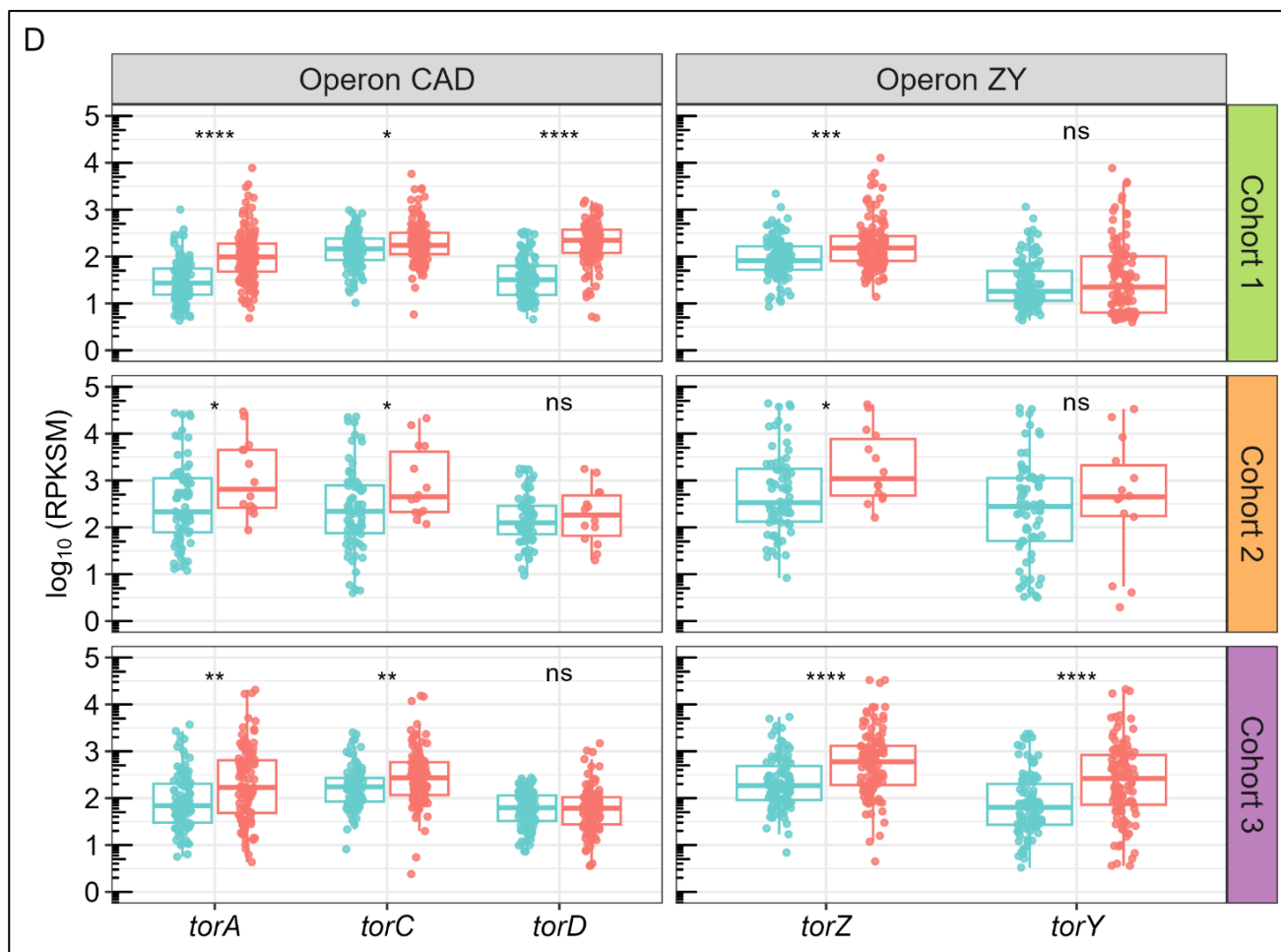


Figure 0-7: Abundance of methane and TMA-producing genes in MASLD.

(A) Characterization of the terminal reactions involved in the formation of microbial methane. Enzymes, EC numbers, coding genes and KO groups, as well as reaction types, are indicated for each metabolic step according to the legend annexed to Figure III-5A. **(B)** Abundance of genes encoding methyl-CoM reductase genes (*mcr*, MCR I; *mrt*, MCR II) across cohorts. **(C)** Abundance of genes encoding CoB-CoM heterodisulfide reductase (*hdr*) and tetrahydromethanopterin S-methyltransferase (*mtr*), involved in cofactor regeneration and methyl-CoM replenishment, respectively. **(D)** Abundance of *tor* operons (*torCAD* and *torZY*). Facets represent genic operons. Boxplot layout and statistical tests were conducted as in Figure III-4B.

E) Candidate metabolic genes are accessory in the GM

Although this analysis identifies *A. rectalis* and multiple metabolic genes involved in the formation of butyrate, SCAs, TMAO and methane as consistent taxonomic and functional signatures associated with MASLD, these markers alone are not sufficiently robust to establish a causal link between GM composition and the MASLD phenotype. GM bacterial species have pleomorphic genomes, consisting of up to thousands of genomically distinct strains that share a conserved genetic core but possess an accessory genome that is highly variable among

them^{84,260}. This accessory genome is often encoded in plasmids and other MGEs which are subject to frequent changes and can readily transfer between different bacterial strains^{165,261}.

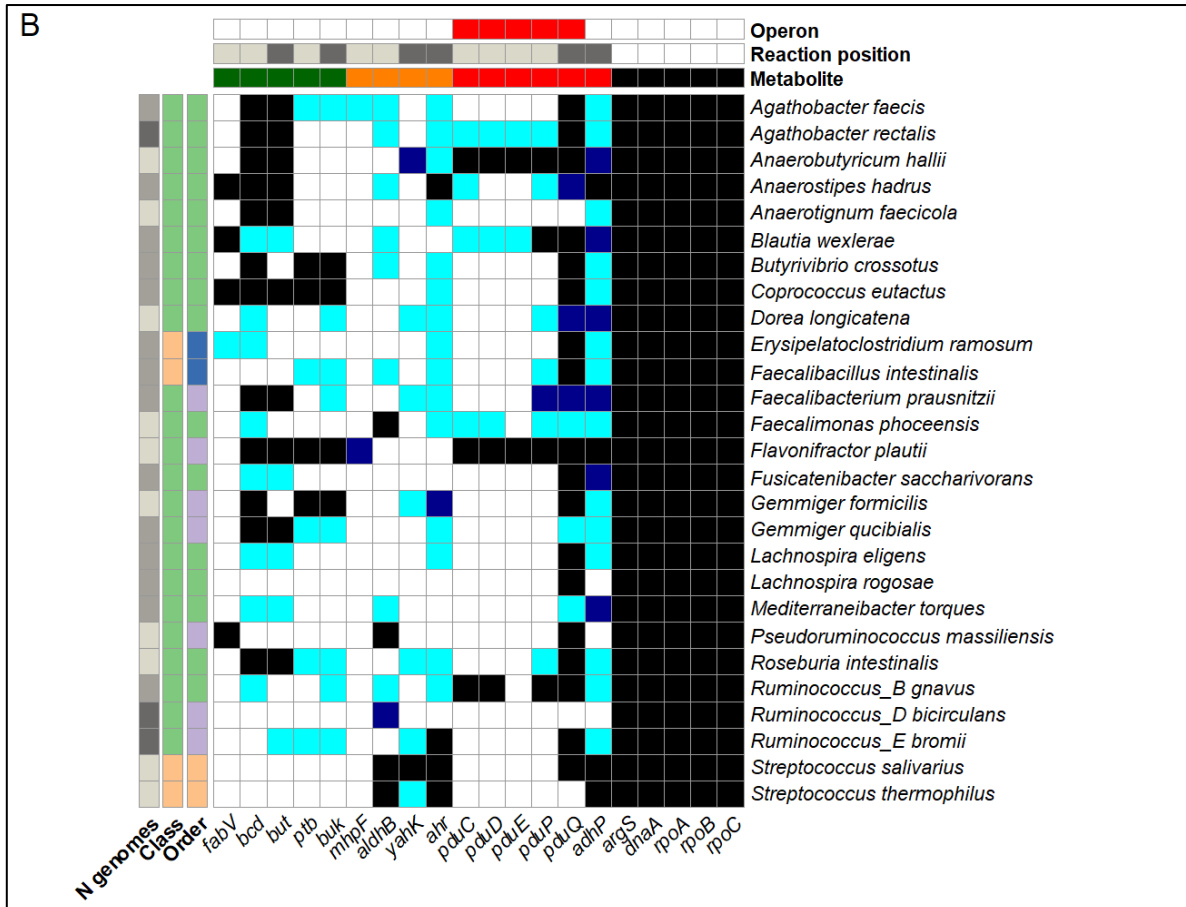
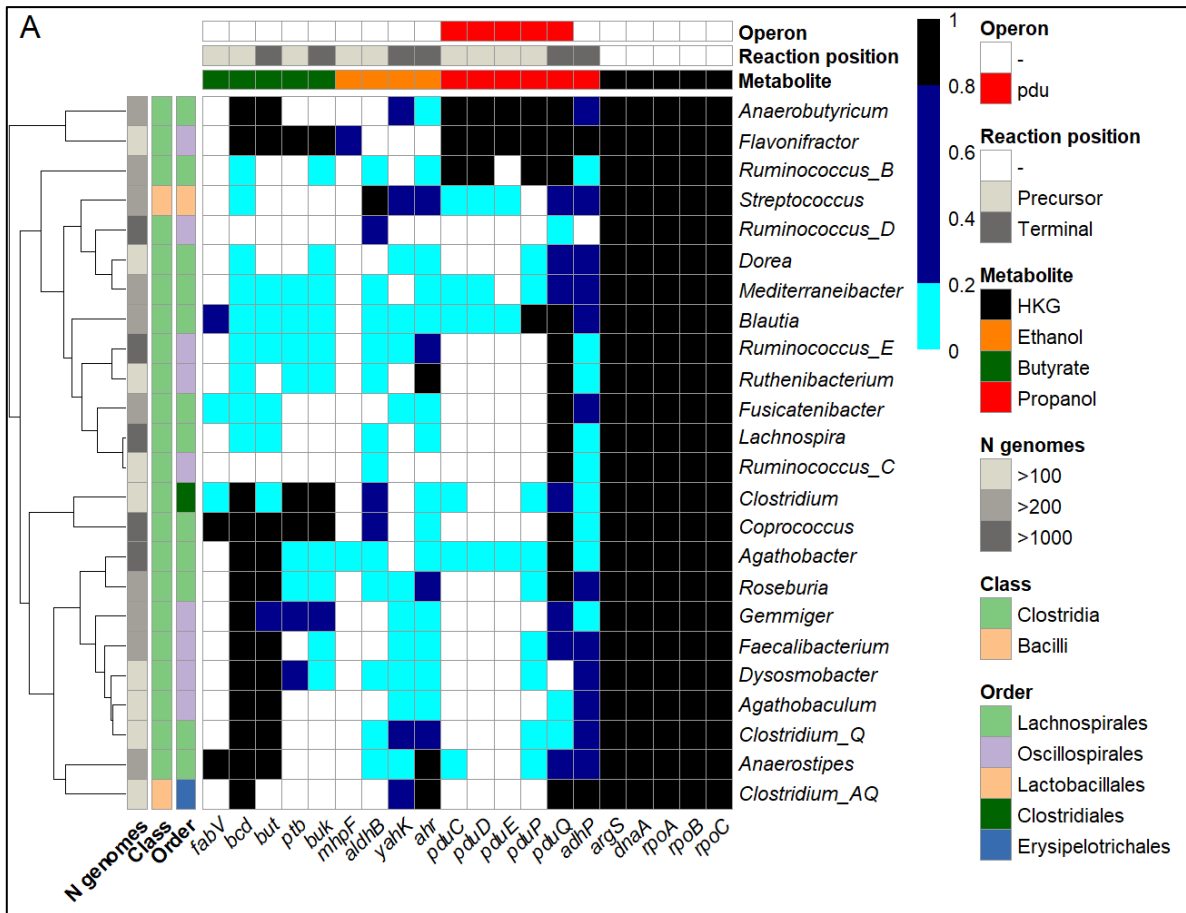
To test whether the observed metabolic shifts reflect conserved traits or strain-specific variability, we assessed whether candidate metabolic genes involved in MASLD are core or accessory across some of the most abundant human GM genomes. Genome selection and genic presence were determined as described in Materials and Methods. Briefly, the presence of a target gene in a clade was defined as the percentage of strains encoding it, classifying it as core if found in >80% of congeneric or conspecific genomes, as accessory if found in 20-80%, and as highly accessory if present in <20% genomes.

Most of the selected GM genera (n=24) belong to class Clostridia, specifically to orders Lachnospirales and Oscillospirales (Figure III-8A). Butyrate-producing genes from the butyryl-CoA pathway (*bcd* and *but*) were core in nearly half of these genera, while those from the butyryl-P pathway (*ptb* and *buk*) were core only in *Flavonifractor*, *Clostridium* and *Coproccoccus*. Notably, butyrate-producing genes were highly accessory across several genera, including *Lachnospira*, *Ruminococcus_B* and *E*, *Ruthenibacterium*, *Blautia*, *Fusicatenibacter*, *Dorea*, and *Mediterraneibacter*, and were completely absent in *Ruminococcus_C* and *D*. In contrast, acetaldehyde and ethanol-producing genes were highly accessory across all genera except *Streptococcus*, an heterolactic genus that constitutively encoded *aldhB* in all its genomes (specifically in *S. salivarius* and *S. thermophilus*, Figure III-8B) and exhibited moderated presence (20-80%) of *yahK/ahr*. The acetaldehyde-producing gene *mhpF* was nearly absent in the GM, except for a few genomes of *Flavonifractor* and *Agathobacter*. Propionaldehyde and propanol-producing genes from the *pdu* operon were core in three genera: *Ruminococcus_B* (*R_B. gnavus*), *Anaerobutyricum* (*A. hallii*) and *Flavonifractor* (*F. plautii*), although *pduE* was unexpectedly absent in *Ruminococcus_B*. These genes were highly accessory in *Blautia* (*B. wexlerae*), *Streptococcus*, *Medierraneibacter* or *Agathobacter* (*A. rectalis*). Importantly, *A. hallii* and *F. plautii* were the only species encoding both core butyrate and propanol-producing genes.

Note that the nomenclature used in the UHGG follows the Genome Taxonomy Database standards, which redefine taxonomic ranks based on genome-wide phylogenetic analyses²⁶². As a result, traditional genera such as *Ruminococcus* and *Clostridium*, which were found to be polyphyletic (i.e., composed of species that do not share a single common ancestor) are split

into multiple monophyletic clades (e.g., *Ruminococcus_A*, *Ruminococcus_B* or *Clostridium_Q*).

Beyond their presence within GM pangenomes, we also assessed the presence of candidate metabolic genes in plasmids to evaluate their potential for HGT. Genes involved in ethanol formation such as *aldhB*, *yahK* and *ahr* were among the most widely distributed, being present in hundreds of different plasmids and thus suggesting high mobility (Figure III-8C). In contrast, butyrate-producing genes (*fabV*, *but*, *ptb*, and *buk*) and propanol-related genes (*pduC*, *pduD*, *pduE*, *pduP*, and *pduQ*) were found in far fewer plasmids, further supporting the idea that these pathways are more strain-specific and chromosomally encoded across GM genomes. A notable exception was *bcd*, which was broadly present across both chromosomal and plasmid-encoded sequences. These findings reinforce the idea that non-essential metabolic functions, especially ethanol production, are often encoded in MGEs, contributing to their accessory nature within the GM.



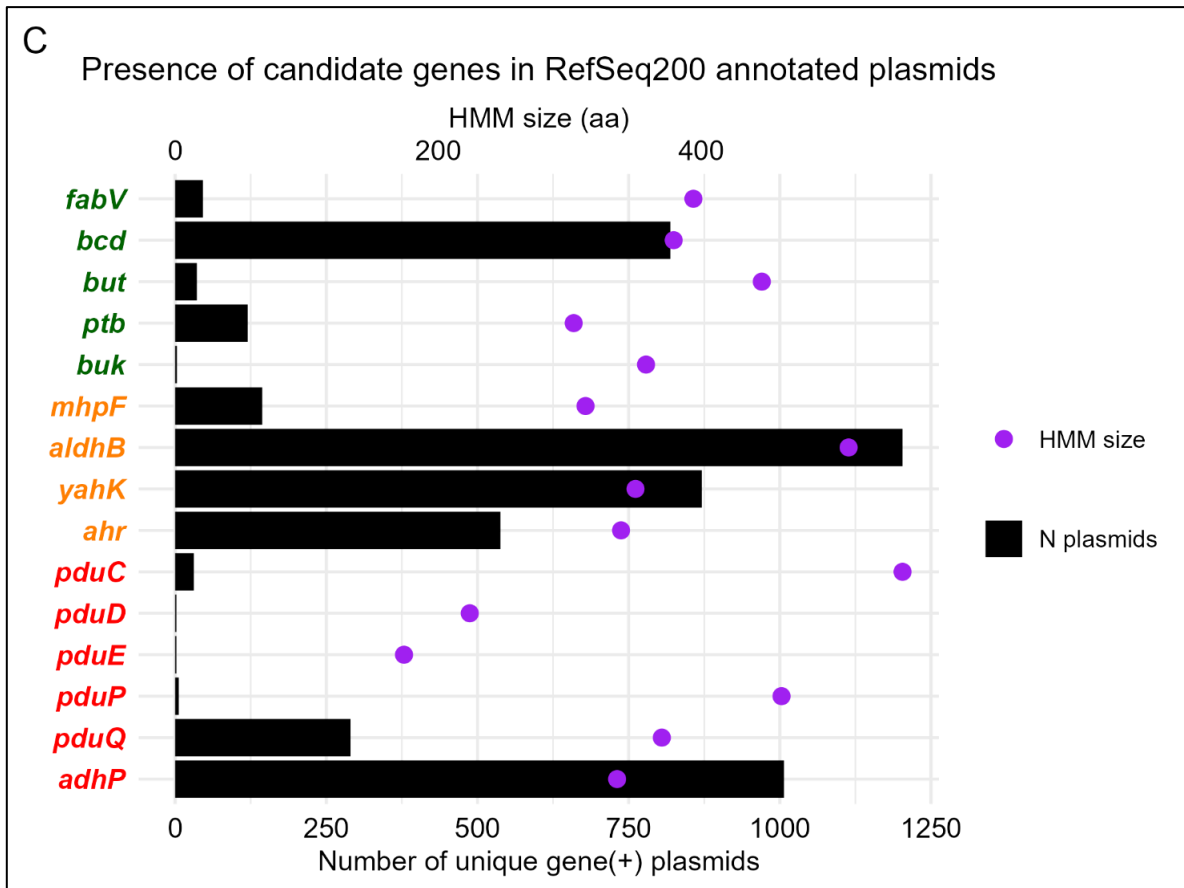


Figure 0-8: Presence of candidate metabolic genes in the human GM.

(A-B) Heatmap showing the presence of candidate genes associated with MASLD across some of the most abundant clades in the human GM. The x-axis represents genes associated with butyrate and SCAs production, as well as USCGs. The y-axis lists the bacterial clades inspected. Gene presence is expressed as the percentage of **(A)** congeneric or **(B)** conspecific genomes encoding each target gene. Colors indicate genic presence: black for core genes (>80% of genomes), dark blue for accessory genes (20-80% genomes), light blue for highly accessory genes (<20% genomes) and white for genes completely absent in the clade. Hierarchical clustering was applied to the y-axis to group genera (top) based on gene-presence similarity. **(C)** Distribution of candidate genes across RefSeq200 annotated plasmids. Bars represent the number of distinct plasmids encoding at least one copy of each gene. Purple dots indicate the HMM size in amino acids.

4. Discussion

MASLD is a liver pathology influenced by multiple contributing factors, and the GM has been long suspected to play a role in its progression. Numerous studies have tried to ascribe a causality between GM composition and MASLD, but taxonomic signatures in key GM clades, such as *Blautia* and *Roseburia spp.*, have been inconsistently associated with MASLD progression¹⁹³. These issues may be linked to the use of 16S rRNA-based profiling in metataxonomic studies, which frequently lacks resolution at species and strain level²³.

However, the heterogeneity concomitant to bacterial genomes poses another possible cause for the lack of association between disease progression and the taxonomic distribution. Bacterial genomes are highly plastic, and strains from the same species often harbor different genes, many of which are involved in different metabolic pathways⁸⁴. Fermentative pathways in lactobacilli, for example, are often strain-specific²⁶³, and the products of the enterobacterial mixed acid fermentation also exhibit wide variability between strains²⁶⁴. If MASLD progression is linked to the metabolic functions of the GM, then the association between the disease and the taxonomic profile may be obscured by the inherent metabolic diversity of bacterial species. To study these potential alternatives, we followed two distinct strategies.

To identify species associated with MASLD and overcome the limitations of 16S rRNA analysis, we utilized multiple marker genes for taxonomic classification, in an attempt to provide a more comprehensive view of microbial abundances in health and disease states. Using MetaPhlan, we profiled the taxonomic composition of the GM in metagenomic fecal samples of healthy and MASLD donors across three independent cohorts. We identified 28 bacterial genera shared across all cohorts (Figure III-4A), collectively accounting for 30-60% of the total GM abundance (Figure III-4B). Among these, five candidate genera -*Gemmiger*, *Streptococcus*, *Mediterraneibacter*, *Ruminococcus* and *Agathobacter* (Figure III-4D)- and two species, *Agathobacter rectalis* and *Bacteroides uniformis* (Figure III-4E), exhibited consistent differences in their relative abundance between groups across all the cohorts. These taxa rank among the most abundant clades in the GM (Figure III-4C, Supplementary Figure S-III-1). However, only *Agathobacter rectalis* showed statistically significant differences. Given its high abundance in the GM, its consistent abundance trend at both genus and species level, and its significant depletion in MASLD, *A. rectalis* emerges as a robust taxonomic marker for the disease (Supplementary Figure S-III-2).

Although this observation has been previously reported^{169,191}, this result should be interpreted with caution due to ongoing debates in bacterial nomenclature, as *A. rectalis* is a newer designation that has replaced *Eubacterium rectale*¹⁷⁴. Updates in taxonomic assignments pose challenges, as newer species or genera may not be consistently reflected across different databases or studies^{262,265}. Moreover, taxonomic classifications themselves, whether based solely on 16S rRNA or on multiple markers, limit our ability to reliably infer GM changes. Bacterial genome plasticity, driven by MGEs, can blur the lines of taxonomy-

phenotype associations, as they contribute to a vast strain-level diversity⁸⁴ that remains undetected by taxonomic profiling methods. In fact, we have previously reported that genomic shifts in GM propionate production associated with inflammatory bowel disease do not align with species-level variations²⁶⁶. These findings are further extended in Annex I.

GM associated metabolism contributes to MASLD by disrupting GLA integrity. For this reason, we focused on identifying abundance differences in the genes responsible for producing metabolites potentially implicated in MASLD across several patient cohorts, i.e., butyrate, SCAs, TMA and methane. We isolated the GM gene families encoding the enzymes involved in the biosynthesis of these metabolites and aligned them against the metagenomic samples to quantify their abundance. Butyrate-producing genes are depleted in the GM of MASLD patients (Figure III-5B), particularly those involved in the crotonyl-butyryl CoA axis: *fabV* (butyryl-CoA oxidoreductase), *bcd* (butyryl-CoA dehydrogenase) and *but* (butyryl-CoA:acetate-CoA transferase) (Figure III-5A). These differences were more pronounced in MASLD-related cirrhosis.

While *ptb* and *buk* were also higher in control groups, their levels did not differ significantly between initial and advanced MASLD groups within Cohort 2. This finding is consistent with the notion that the CoA transfer pathway is the predominant route for butyrate production in the GM, as previously reported²⁶⁷. The subtler differences in *ptb* and *buk* in Cohort 2 may reflect a consistent MASLD phenotype across different fibrosis stages, potentially influenced by the smaller sample size relative to Cohort 3. Additionally, two other enzymes -structured as two-subunit systems-, can transfer CoA from different cofactors to butyrate to produce butyryl-CoA: butyryl-CoA:acetoacetate-CoA transferase (encoded by *atoD* and *atoA*, EC 2.8.3.9) and acetate CoA-transferase (encoded by *ctfA* and *ctfB*, EC 2.8.3.9) (Figure III-5A). These were not included in this study since they catalyze the reverse reaction, consuming butyrate to form butyryl-CoA²⁶⁸.

Contrarily, genes involved in the formation of ethanol and propanol through connected fermentative reactions are generally increased in MASLD (Figure III-6). Specifically, genes *mhpF* and *aldB* -encoding two aldehyde dehydrogenases that produce ethanol-, and *yahK* and *ahr* -encoding two NADPH-dependent aldehyde reductases that form acetaldehyde- are significantly increased in the GM of MASLD patients compared to the healthy groups (Figure III-6B). Genes involved in the production of 1-propanol structured in operon *pdu* are also

increased in MASLD (Figure III-6C, Supplementary Figure S-III-5A). Specifically, *pduCDE* - encoding a propanediol dehydratase-, *pduP* -encoding a propionaldehyde dehydrogenase- and *pduQ* and *adhP* -encoding two distinct propanol dehydrogenases-. The pronounced differential abundance of these propanol-producing genes, organized in the *pdu* operon and *adhP*, which have been largely unexplored in the context of liver disease, suggests that 1-propanol produced in bacterial microcompartments may play a more significant role than ethanol in MASLD onset and progression.

The abundance of genes in the TMA-methane metabolic axis is also disrupted in MASLD. Genes involved in methane formation (Figure III-7A) are depleted in the GM of individuals with the pathology. This reduction includes both MCR I and II coding-genes (Figure III-7B), which reduce methyl-CoM to methane and *mtr*, *hdr* (Figure III-7C), and Mtb / Mta-coding genes (Supplementary Figure S-III-3), dedicated to regenerating methyl-CoM. Conversely, TMAO-reductive genes, organized in the *torCAD* and *torZY* operons and responsible for regenerating TMA under anaerobic conditions, are increased in MASLD (Figure III-7D, Supplementary Figure S-III-5B). These results suggest that methane production is impaired in MASLD. The opposing trend in the abundance of *tor* operons further indicates that this impairment stems from a metabolic bottleneck that leads to TMA accumulation instead of its conversion to methyl-CoM, effect that is driven by TMAO reductases I and II.

Taken together, these results reveal a geno-metabolic alteration associated with MASLD, characterized by an increase in genes involved in the production of SCAs and TMA, and a concurrent depletion of genes responsible for butyrate and methane production (Figure III-9). We confirmed that differences in gene abundance were driven by the MASLD phenotype by showing no differences in USCG abundance between groups (Supplementary Figure S-III-4). These findings support the hypothesis that the accumulation of endogenous alcohols may represent a key metabolic feature of MASLD, occurring alongside the displacement of butyrate-producing taxa, as previously proposed²⁰⁰. Moreover, the observed decrease in *A. rectalis* abundance partially explains the reduction in butyrate levels, as this organism is a known GM butyrate producer¹⁷⁴.

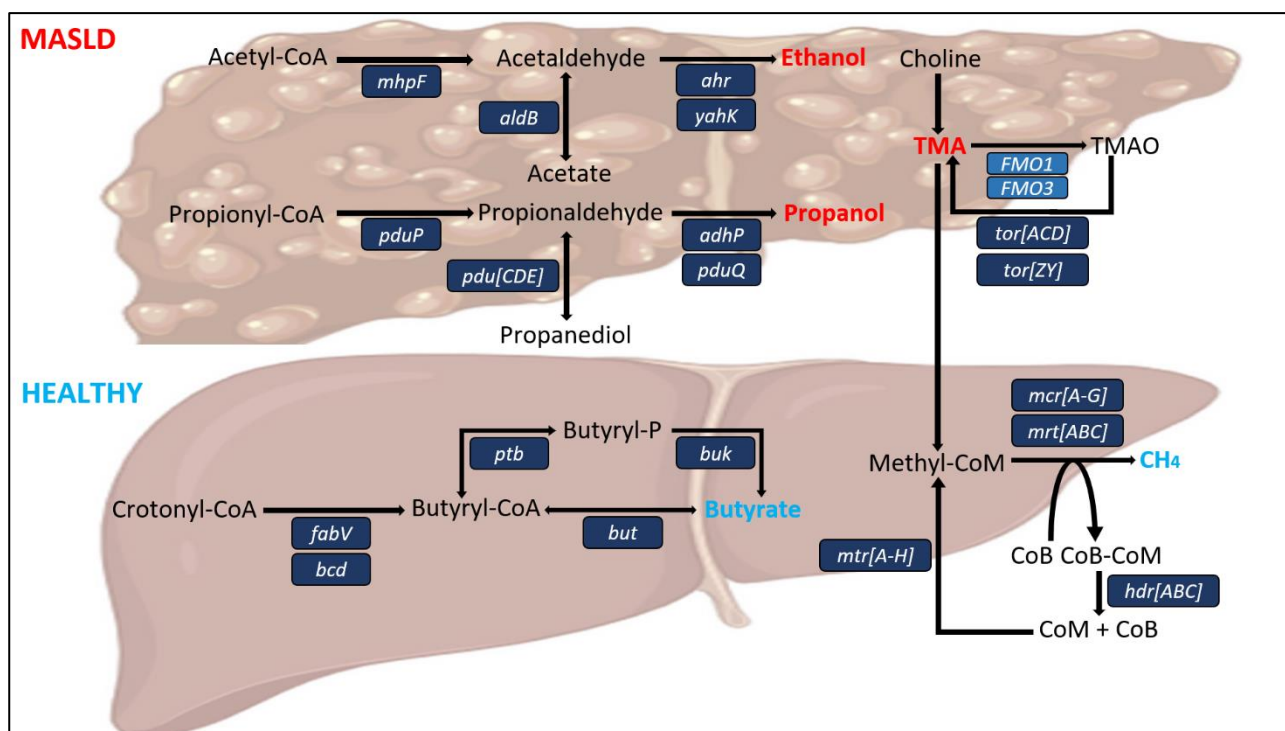


Figure 0-9: Geno-metabolic alterations associated with MASLD.

Schematic representation of GM gene shifts in MASLD, highlighting pathways for butyrate, SCAs, TMA and methane production. Genes involved in the production of ethanol, propanol and TMA are increased in MASLD (top), whereas those associated with butyrate and methane are decreased (bottom). Dark blue boxes indicate the catalytic genes with altered abundance in each metabolic step.

Notably, results obtained at the single-gene level using curated gene families were partially reproduced with the whole-metagenome approach applied to random sub-cohorts, particularly for genes involved in the TMA-methane metabolic axis and propanol production (Supplementary Figure S-III-5). This partial overlap supports the accuracy of the single-gene analysis. However, the lack of complete concordance may be attributed either to the reduced sample size in the co-assemblies or to mismatches between our custom gene families and the broader KO annotations required to detect the same functional signals.

We found that metabolic genes associated with MASLD exhibit a highly accessory profile within the GM (Figure III-8). Butyrate-producing genes are accessory in nearly half of the most abundant GM genera, while SCA-producing genes display an even more pronounced accessory pattern (Figure III-8A). Among the two main butyrate biosynthesis routes, genes from the butyryl-CoA pathway are more pervasive in the GM than those from the butyryl-P pathway. *Phocaeicola* and *Bacteroides* -the two most abundant GM genera- were excluded from this analysis due to a limited number of high-quality genomes passing our completeness filters,

despite their extensive genomic representation in the UHGG. Interestingly, the two key genes in the butyryl-CoA pathway (*bcd* and *but*) are core in *A. rectalis* (Figure III-8B). Nonetheless, some strains of this butyrate-producing species also harbor genes for ethanol and propanol production, which our findings suggest may contribute to MASLD onset. This highlights a critical point: taxonomic signatures alone are insufficient for associating bacterial clades with disease. Genome plasticity introduces substantial strain-level metabolic variability in the GM, leading to distinct phenotypic traits even within the same species: an effect well-documented in both pathogenic and commensal *E. coli* strains^{269,270}.

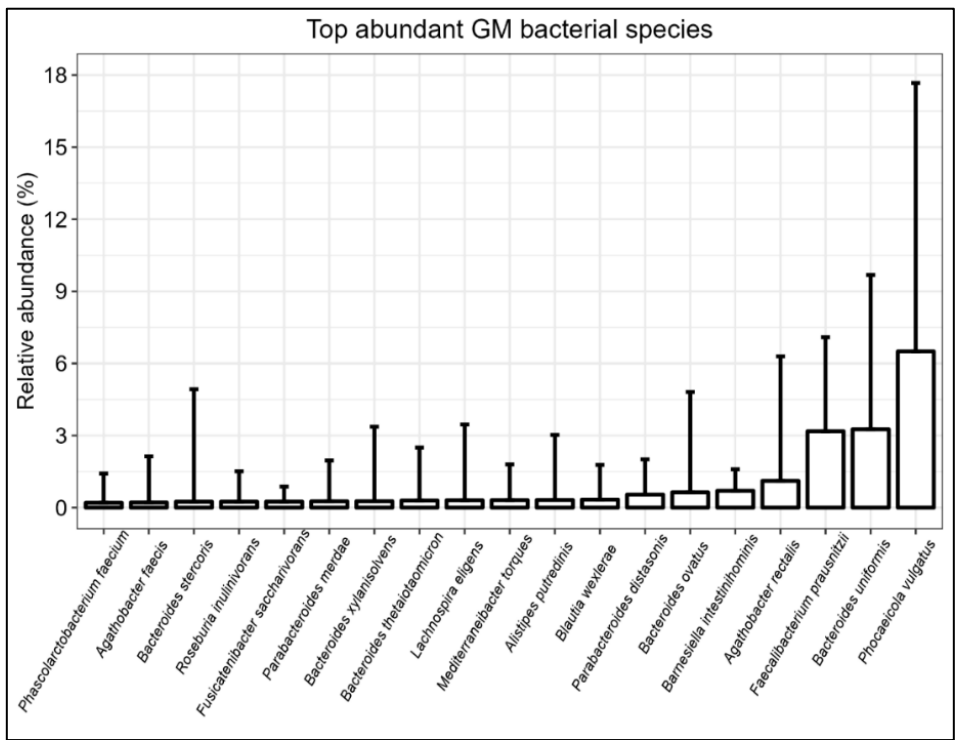
Further supporting their mobility, metabolic genes are frequently encoded in plasmids (Figure III-8C). Among these, ethanol formation genes (*aldhB*, *yahK*, and *ahr*) are the most widely distributed, while butyrate- and propanol-related genes are found in plasmids less frequently. Since non-essential, adaptive metabolic pathways are often encoded in plasmids across many bacterial species^{270–272}, their presence likely drives phenotypic differences in MASLD without altering the overall species composition detected by conventional metataxonomic marker genes. If MASLD is actually driven by bacterial metabolism, strains from different species, but harboring similar metabolic pathways, may produce similar liver-impacting metabolites, potentially contributing to MASLD pathogenesis regardless of their specific taxonomic position.

This study has several limitations. First, although gene abundance provides a useful proxy for functional potential, it does not directly measure metabolic activity. Integration with metatranscriptomic or metabolomic data would provide complementary insight into gene expression and metabolite production. Second, while our findings provides cross-cohort evidence of functional shifts in MASLD, it does not establish causality. Experimental validation through microbial isolation, gnotobiotic animal models, or gene knockout approaches, would be needed to clarify mechanistic roles. Finally, host-related factors such as diet, medication use, and genetic background -each of which may influence the abundance and activity of candidate metabolic genes- were not fully accounted for and warrant further investigation.

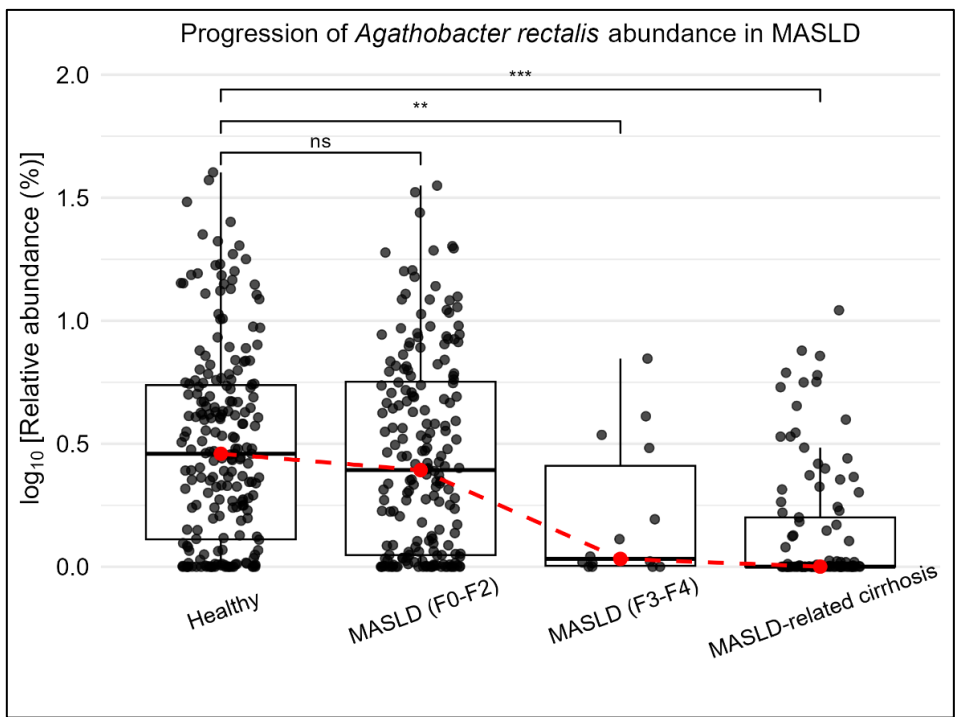
5. Conclusions

Our study reveals that MASLD is marked by distinct geno-metabolic alterations in the GM that are not captured by conventional taxonomic profiling. Specifically, we observed an increase in genes involved in the production of SCAs and TMA, coupled with a depletion of genes responsible for butyrate and methane production. These results confirm previous findings and suggest a broader metabolic shift that may contribute to the pathogenesis of MASLD, with the displacement of beneficial butyrate-producing taxa such as *Agathobacter rectalis*. Importantly, our work underscores the value of analyzing metagenomes at the gene level to extract phenotypic signatures beyond what 16S rRNA-based or marker gene-based approaches can offer. By integrating metabolic pathway analysis with the isolation of specific gene families, and by examining gene abundance, accessory gene profiles and plasmid-mediated gene mobilization, we have significantly advanced our understanding of bacterial metabolism in MASLD. This comprehensive approach not only enhances our understanding of the microbial etiology of MASLD, but also paves the way for improved bacterial analyses in clinical settings, offering new avenues for diagnosis and therapeutic intervention in liver pathologies.

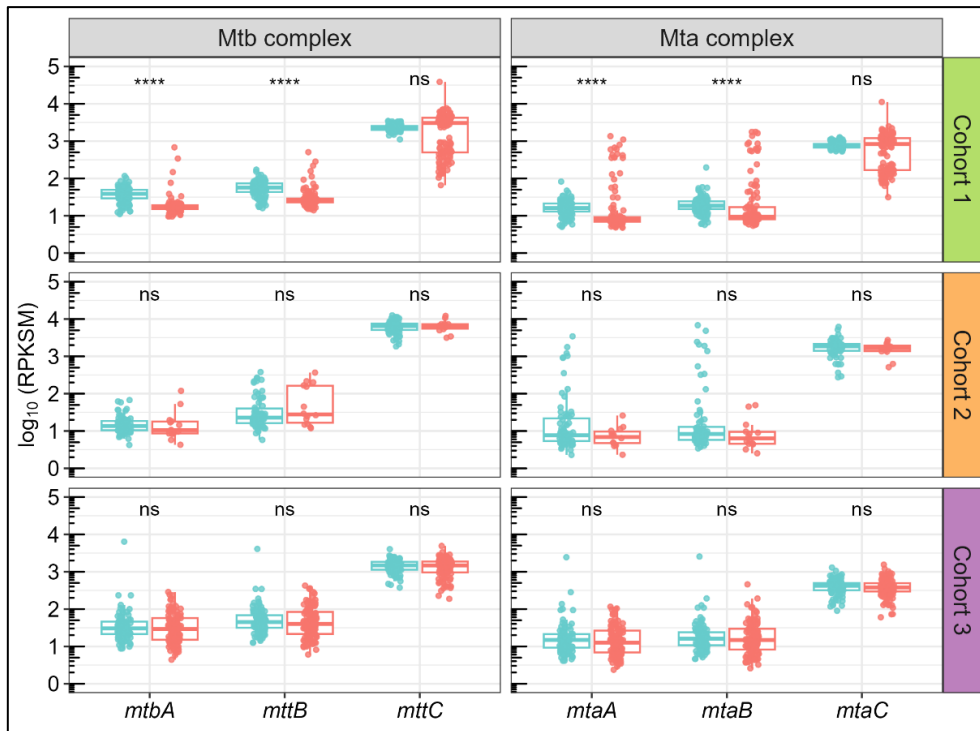
6. Supplementary material



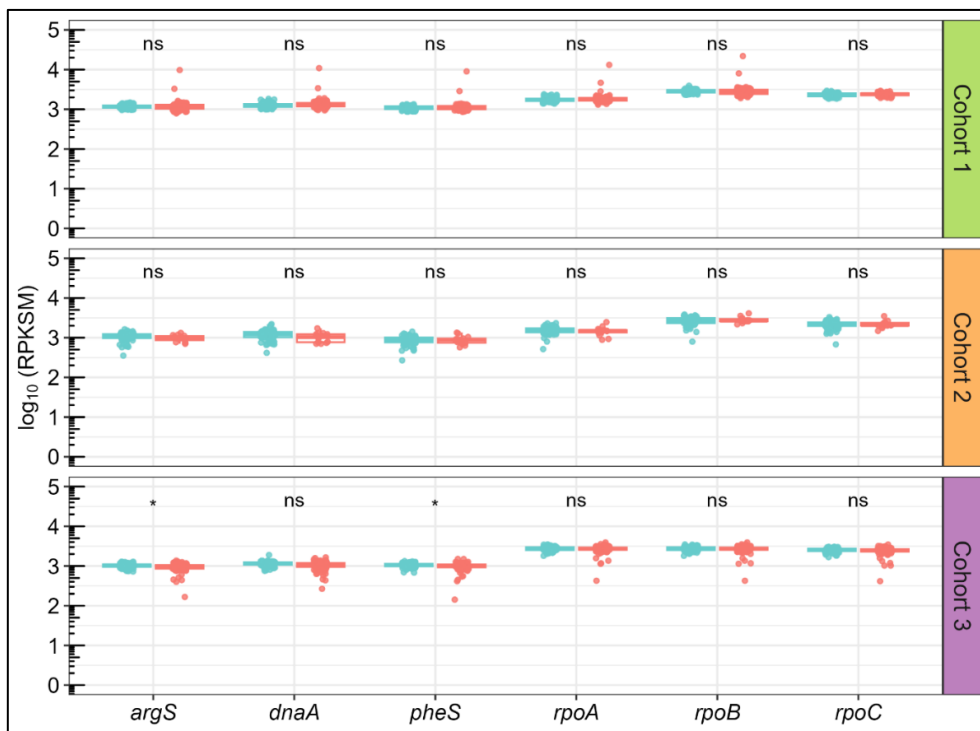
Supplementary Figure S-III-1. Top abundant GM species (average absolute abundance >0.2%).



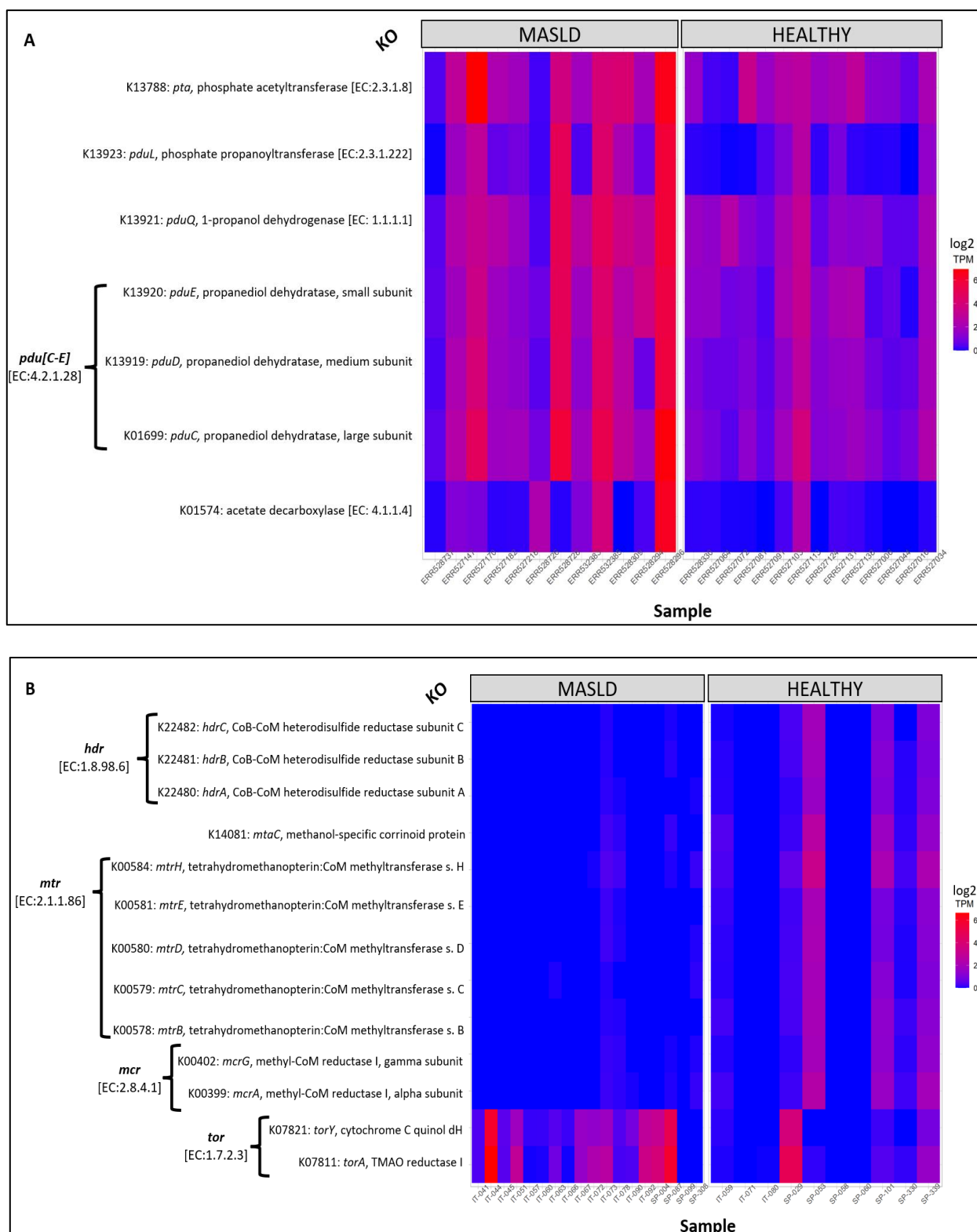
Supplementary Figure S-III-2. *Agathobacter rectalis* abundance across MASLD stages. Abundances are shown on a logarithmic scale. Median values per group are connected by a red dashed line. Differences in abundance were evaluated using pairwise Mann-Whitney tests with Benjamini-Hochberg adjustment. Samples from all three cohorts were pooled, with individuals grouped into four clinical stages.



Supplementary Figure S-III-3. Abundance of Mtb and Mta-complex genes in MASLD. Boxplot organization and statistical tests were conducted as in Figure III-4B.



Supplementary Figure S-III-4. USCG abundance in MASLD. Abundance of universal, single-copy marker genes in the three cohorts. Boxplot organization and statistical tests were conducted as in Figure III-4B. *argS* encodes for the arginyl-tRNA synthetase, *dnaA* encodes for the chromosomal replication initiator protein DnaA and *rpoA/B/C* encode, respectively, the alpha, beta and gamma subunits of the DNA-directed RNA polymerase.



Supplementary Figure S-III-5. Abundance of candidate KO groups in MASLD. Heatmaps represent KOs involved in the **(A)** propanol-formation and **(B)** TMA-methane metabolic pathway in sub-cohorts 3 and 1, respectively. KO abundance is expressed in tags per million (TPM), in log2 scale. x-axis indicates samples from both comparison groups and y-axis indicates KO groups, EC numbers, and associated gene and enzyme names according to KEGG. Only KOs with log fold-change >2 and a p-adjusted <0.05 are represented. KOs were predicted from the co-assembled sub-cohorts with SqueezeMeta.

CHAPTER II

PLASMID-MARKER GENES

IN THE OCEAN

1. Introduction

ARGs have emerged as a critical global health threat due to their role in disseminating antibiotic-resistant pathogenic and commensal bacteria, which, in turn, causes failures in treating infectious diseases and contributes to a significant percentage of human deaths worldwide²⁷³. Their widespread occurrence in human and farm settings leads to a heavy load of ARGs in the waste products of human activity²⁷⁴. Antibiotic misuse and overuse in clinical settings, combined with the contamination of natural environments through human effluents, are major drivers for the accumulation of these resistance determinants^{275,276}. Residual wastewater from anthropogenic environments serves as a reservoir for these contaminants, making aquatic ecosystems particularly vulnerable to ARG pollution^{277,278}. Consequently, ARGs have come to be regarded as emerging pollutants, with their presence documented even in areas remote from direct human activity, such as pristine rivers²⁷⁹ and Antarctica²⁸⁰.

ARGs spread occurs through HGT, mainly by the dissemination of conjugative plasmids between bacteria that belong to different taxonomic groups. This mechanism has been documented across different environments such as soil^{281,282}, wastewater treatment plants²⁸³, farms²⁸⁴, or the human GM²⁸⁵. A critical factor in assessing the risk posed by ARG contamination in global ecosystems is the mobile potential of these genes through HGT. It is widely recognized that ARGs found in conjugative plasmids and integrative and conjugative elements (ICEs) represent a greater risk of dissemination. However, the plasmid content of oceans, rivers and terrestrial ecosystems remains far less studied than that of the GM. Recent findings have revealed the existence of marine plasmids (MAPs) with global distribution and a broad host range^{286,287}. These MAPs are capable of transmitting ARGs over intercontinental distances and may even reintroduce them into the human food chain via marine products, mirroring the mobilization of seaweed-degrading genes from oceans to the human GM^{288,289}.

Although ARGs have been suggested to accumulate in the ocean²⁹⁰, their dissemination routes, contamination levels in indigenous marine bacterial populations, and overall severity of this contamination remain unclear^{291,292}. While ultrasensitive methods like PCR have been extensively employed to detect specific ARGs, these approaches have several limitations. First, the typically small number of samples may not adequately represent entire ecosystems²⁹³. Additionally, uneven sampling across geographic regions can affect the generalizability of results and must be considered when making global comparisons²⁹³. PCR-based studies also

tend to focus on a limited set of ARGs, overlooking the broader resistome present in a given environment^{122,126}. In contrast, metagenomic samples are difficult to compare quantitatively within and between studies, given the vast differences in microbial complexity and sequencing coverage^{133,136}.

Plasmids can be studied in environmental samples through *de novo* assembly of metagenomic sequencing data derived from diverse biomes, such as the GM²⁹⁴, groundwater²⁹⁵ or marine environments²⁹⁶. Computational tools like Recycler²⁹⁷, metaplasmidSPAdes²⁹⁸ or SCAPP²⁹⁹ facilitate the recovery of plasmid sequences from metagenomic samples. However, the recovery of complete plasmids -particularly large ones- from short-read based metagenomic assemblies remains challenging³⁰⁰. First, plasmid DNA typically constitutes only a minor fraction of the total genetic material in environmental samples. This reduces the probability of recovering full plasmids from sequencing libraries with insufficient coverage or from metagenomic assemblies with short average contig lengths, although plasmid DNA enrichment protocols can partially overcome this limitation³⁰¹. Second, plasmids often share gene synteny and homology with other MGEs and bacterial chromosomes. This phylogenetic and molecular complexity leads to the formation of convoluted assembly graphs and increases the risk of generating chimeric sequences during *de novo* assembly.

Alternatively, plasmid sequences can be identified from metagenomic contigs through reference-based classification using tools such as PlasFlow³⁰², PlasClass³⁰³, or PlasX²⁶¹. These supervised machine-learning approaches, however, face challenges due to potential bias in training sets, which are often under-represented in non-clinical plasmids. This bias limits classification accuracy -particularly for short plasmids, which contain fewer genes-, as it defines molecular signatures in the database that may differ from those of candidate plasmids. Additionally, the presence of chromosomes with signatures similar to those of plasmids in the training set can result in false positives. ICEs within genomes further complicate the differentiation between chromosomes and plasmids, as they also encode plasmid-signature genes involved in functions like mobilization or conjugation.

The relaxase (RLX) is the protein that recognizes and cleaves the origin of transfer in a plasmid, initiating DNA transfer by conjugation (Figure IV-1). It is the only essential component of the mobilization machinery (MOB), that is shared by all transmissible elements³⁰⁴. This way, the RLX serves as a proxy for a given conjugative system and its abundance, and it can be used

to classify mobilizable and conjugative plasmids^{305,306}. Currently, nine MOB classes encompass the RLX phylogenetic diversity³⁰⁷. Additionally, plasmid taxonomic units (PTUs) are characterized by a particular, conserved MOB class. Specifically, RLXs within the same PTU share >95% average nucleotide identity¹⁶⁵.

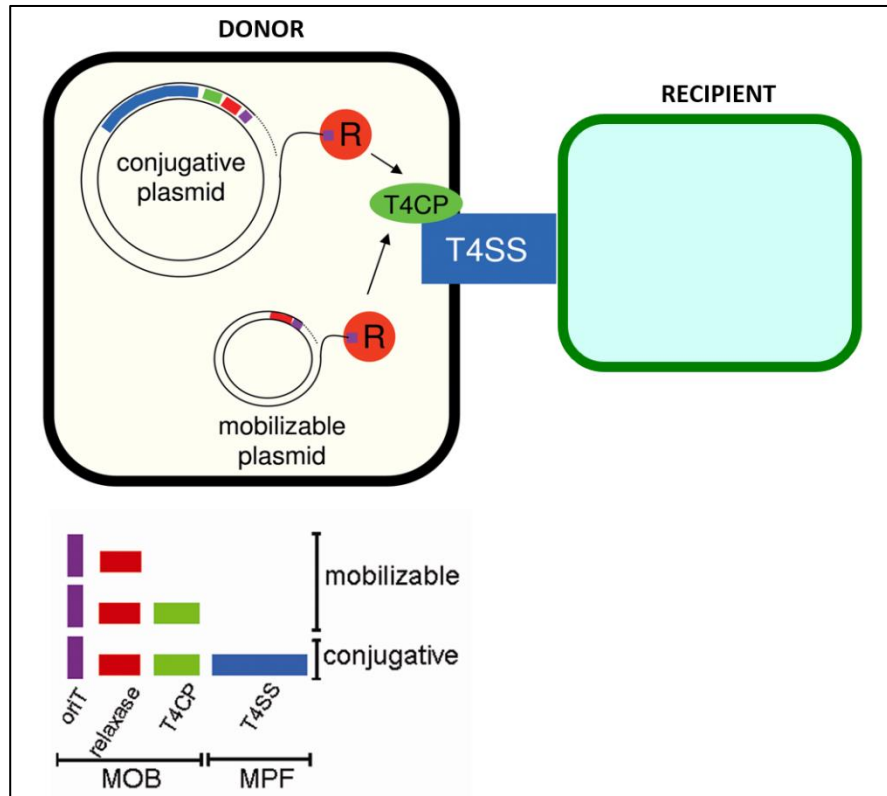


Figure 0-1: RLX organization and function in transmissible plasmids.

The RLX is the only element in common among all conjugative transmissible plasmids. R: RLX; T4SS: type IV secretion system; T4CP: type IV coupling protein; *oriT*: origin of transfer. MOB: mobilization typing. MPF: mating pair formation system. Figure adapted from Smillie *et al.*³⁰⁴.

Although the aforementioned computational tools have substantially improved our ability to identify plasmids in metagenomic contigs, their accuracy is insufficient for performing quantitative analyses. In contrast, measuring RLX abundance in environmental metagenomes can provide insight into the relevance of conjugation as a mechanism for HGT in these settings. In this Chapter, we used the RLX genes as proxies to determine the distribution and prevalence of MAPs and other elements transmissible by conjugation, thereby circumventing the limitation of identifying complete plasmids in metagenomes. We compared ocean microbiomes worldwide with those from fluvial systems, agricultural land and sewage ecosystems, as well as those of the gastrointestinal tracts from both marine and terrestrial mammals. To do so, we

analyzed the abundance and distribution of RLXs in almost 51,000 marine MAGs and in a large collection of whole-metagenomic sequencing samples sourced from diverse databases. We developed unbiased estimators of gene abundance that account for variations in sequencing depth and quality. We also interrogated these samples for the abundance of ARGs and plasmid content by using recently published HMM protein profiles of MGEs.

2. Materials and methods

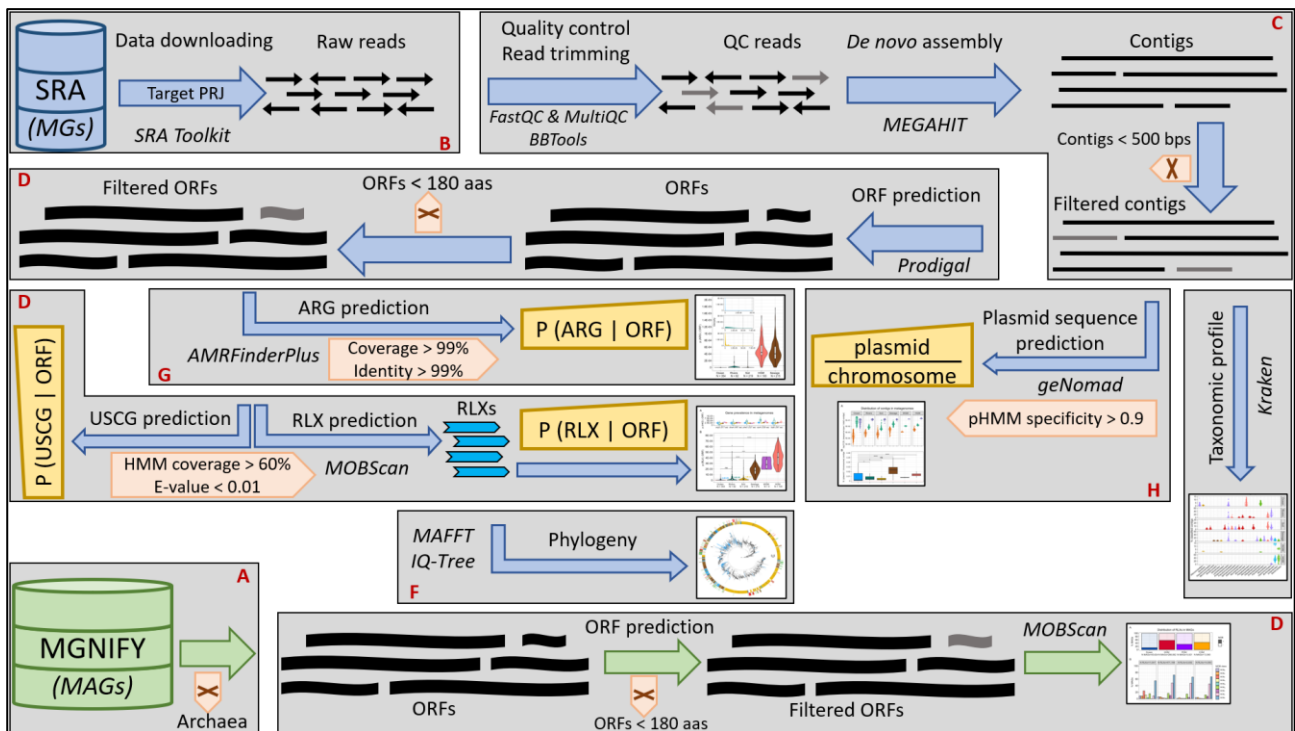


Figure 0-2: Overview of the methodology used in Chapter II.

Diagram summarizing the full bioinformatic pipeline for processing paired-end metagenomic samples: from data retrieval, read quality control and *de novo* assembly of contigs to downstream analysis of gene abundance and MGEs. Blue arrows indicate the workflow progression, and software tools are shown in italicized black beneath each arrow. Filtering steps are highlighted in pink boxes. Red labels refer to subsections (A-H) in Materials and Methods for detailed descriptions.

A) Metagenome-assembled genomes

We retrieved a collection of 50,866 marine MAGs generated from seawater samples via the MGNify FTP repository³⁰⁸. As the aim of our study was to understand the prevalence of bacterial RLXs, we removed 5,242 archaeal MAGs from the initial dataset. In addition, we obtained two separate sets of GM MAGs: 3,972 from the pig GM and 13,386 from the chicken GM. Furthermore, we incorporated the 289,232 human GM MAGs from the UHGG⁸⁴. All MAGs had >50% genome completeness and <5% contamination (Supplementary Figure S-IV-3),

according to defined quality standards⁵⁸. All data were downloaded between October 2022 and June 2024. The following numbers of archaeal MAGs were removed from each batch: 51 (pig GM), 26 (chicken GM) and 1,170 (human GM). Metadata for these MAG collections is provided in their corresponding “*genomes_all_metadata.tsv*” files, located in their individual MGnify sub-folders. Taxonomic annotation of all MAGs was performed using the Genome Taxonomy Database²⁶⁵. This integrated collection of MAGs was subsequently used to examine and compare the prevalence of marine RLXs (Figure IV-3A), the distribution of their MOB classes across marine environments and the GM of higher vertebrates (Figure IV-3B), and the taxonomic distribution of MOB+ MAGs in each biome (Figure IV-3C).

B) Metagenomic sequencing libraries

A total of 354 marine metagenomic sequencing libraries from the *Tara* Oceans project³⁰⁹ were downloaded to quantify the abundance of marine RLXs (Figures IV-4 and IV-5) and ARGs (Figure IV-8), and to classify contigs by their molecular origin (Figure IV-10). These samples, collected between 2009 and 2013 at various locations and depths across every ocean (Supplementary Figure S-IV-4), were obtained from the supplementary materials of Paolli *et al.*⁹¹.

To compare marine data with that from other environments, we downloaded 319 metagenomic fluvial samples previously employed for constructing the GMGC, a recently published global prokaryotic gene catalogue³¹⁰. These samples, encompassing both pristine and agriculturally/urban-impacted sites, were filtered to retain only 82 libraries with a minimum of 0.5 million reads and an average read length >251 bp, while discarding the rest due to low library size and/or inadequate quality. Geographic metadata was not available for these samples.

Additionally, 219 cultivated-land metagenomic samples collected worldwide from different sampling sites and depths (Supplementary Figure S-IV-4) were downloaded from the SMAG catalogue⁹³. These data were filtered to include only samples labeled as “Agricultural Land” and the selection was limited to three samples per unique longitude/latitude to avoid the over-sampling in regions such as China, the United States or Australia, that was present in the original article. Finally, we downloaded 215 urban, untreated sewage samples collected in multiple cities worldwide as part of the Global Sewage study³¹¹ (Supplementary Figure S-IV-4). For comparative analysis with the GM of non-marine and marine mammals, we also retrieved

100 metagenomic samples from the feces of healthy patients recruited in the IBDMDB project (<https://www.ibdmdb.org/>)⁹⁷ and a set of 4 metagenomic samples from the whale colon³¹², excluding 7 non-colonic samples to ensure a consistent comparison of microbial compositions. All samples were downloaded using the corresponding accession numbers indicated in the original articles (Table IV-1) with *fastq-dump*¹⁶⁶, as detailed in the Materials and Methods section of Chapter I. All data were downloaded between March 2022 and June 2024.

Biome	N samples	Reference	Project	SRA accession number
Ocean	354	91	Tara Oceans	PRJEB1787, PRJEB1788, PRJEB9740
Rivers	82	310	GMBC	PRJNA287840
Soil	215	93	SMAG	PRJNA983538
Sewage	219	311	Global Sewage study	PRJEB27621, PRJEB40798 PRJEB40815, PRJEB40816
Whale GM	4	312	NA	PRJEB23642
Human GM	100	97	IBDMDB	PRJNA398089

Table 0-1: Environmental metagenomic datasets analyzed in Chapter II.

Each biome (column 1) comprises *N* metagenomic samples (column 2), derived from various studies (column 3) and organized in specific named projects (column 4). Metagenomic data and geographical metadata were retrieved from the indicated SRA accession numbers (column 5).

C) Quality control of metagenomic reads and *de novo* assembly

Sequencing reads were trimmed to remove Illumina adapter remnants and low-quality regions ($Q < 25$) using BBDuk from the BBTools suite (version 37.62)¹⁷¹. Read quality was checked with FastQC (version 0.12.1)¹⁷² and summarized using MultiQC (version 1.22.3)¹⁷³. High-quality reads were *de novo* assembled with MEGAHIT (version 1.2.9)⁴⁸. Contigs shorter than 500 bp were removed from further analysis, as recommended in similar studies^{93,313}. The sequencing coverage of each contig was recorded as a proxy for its abundance.

D) Relaxase and USCG prediction in MAGs and contigs

Prodigal (version 2.6.3)¹⁸² was used to predict the ORFeome of each MAG, identifying proteins longer than 180 amino acids. The total number of such proteins in each MAG collection was as follows: 75,084,856 (marine), 473,416,701 (human GM), 5,254,668 (pig GM) and 19,944,528 (chicken GM). For metagenomic contigs, we also applied MetaProdigal (i.e., using the “-p meta” option) to leverage pre-trained models for ORF prediction⁵⁰. This resulted in the

following number of proteins >180 amino acids per biome: 272,179,219 (ocean), 1,487,797 (rivers), 250,173,651 (agricultural soil), 144,333,274 (sewage), 791,183 (whale GM) and 6,968,311 (human GM). The coverage of each contig was assigned to all derived ORFs them as a measure of their absolute abundance within the metagenome.

Then, we extracted the RLXs encoded in the predicted ORFeomes using MOBscan (version 1.0.0), which compares sequences to a set of HMMs corresponding to nine MOB classes³¹⁴. ORFs were classified as putative RLXs when HMM coverage was >60%, E-value was <0.01 and the i-E-value was <0.01. The 180-residue threshold corresponds to approximately 60% of the size of the HMMs integrated into MOBscan, ensuring that a hit on any protein reaches at least the minimum RLX length, which is approximately 300 residues. The same protocol was applied to five USCGs of similar size to RLXs, used as controls, with their corresponding HMMs retrieved from TIGRFAMs³¹⁵ (Table IV-2).

USCG	Encoded protein	Protein size (aa)	TIGRFAM HMM
<i>argS</i>	arginyl-tRNA synthetase	577	<i>TIGR00456.1</i>
<i>dnaA</i>	chromosomal replication initiator protein DnaA	467	<i>TIGR00362.1</i>
<i>pheS</i>	alpha subunit of the phenylalanyl-tRNA synthetase	327	<i>TIGR00468.1</i>
<i>rpoA</i>	alpha subunit of the DNA-directed RNA polymerase	329	<i>TIGR02027.1</i>
<i>trpB</i>	beta chain of the tryptophan synthase	397	<i>TIGR00263.1</i>

Table 0-2: USCGs analyzed in Chapter II.

E) Validation of the ORF quantification method

To evaluate the accuracy of the method used for quantifying ORFs in metagenomes, four synthetic communities of diverse complexity were constructed by randomly selecting and downloading closed genomes from RefSeq200²¹ (Table IV-3). For each community, metagenomic paired-end Illumina reads were simulated using InSilicoSeq (version 2.0.1)³¹⁶, which reproduces the quality profiles and error patterns (substitutions, insertions and deletions) of HiSeq and NovaSeq platforms through kernel-density estimators and precomputed error models. A corresponding abundance file was generated for each community, specifying the relative abundance (scaled from 0 to 1) of each genome. Both sequencing platforms were included to assess the impact of platform-specific error profiles on assembly quality and gene recovery.

Synthetic community	N genomes	N chromosomes	N plasmids
A	500	611	434
B	250	264	221
C	100	109	69
D	25	28	19

Table 0-3: Genomic composition of the synthetic communities.

Each mock community was constructed with a decreasing number of randomly selected genomes from RefSeq200. Individual genomes may contain multiple chromosomes and/or plasmids, contributing to the total counts reported.

Simulated reads were quality-controlled and *de novo* assembled into contigs >500 bp, as described in Section IV-2C. The sequencing coverage of each contig was recorded as a proxy for its abundance. ORFs were predicted on the assembled contigs using MetaProdigal⁵⁰, retaining those >180 residues, and RLXs were further identified as described in Section IV-2D (Supplementary Figure S-IV-16). Each ORF was assigned the coverage of its parent contig, representing its absolute abundance in the metagenome.

To establish the ground truth for ORF abundance, Prodigal¹⁸² was used to predict ORFs and RLXs >180 amino acids from the original genome sets used to simulate the reads. Each ORF predicted in a genome was then assigned the relative abundance of its source genome, as defined in the InSilicoSeq abundance files. This provided a measure of genomic relative abundance for each ORF (Figure IV-9).

F) Relaxase clustering and phylogeny

We merged all the metagenomic RLXs with those annotated in RefSeq200²¹, extracted their first 300 amino acids corresponding to the RLX domain, and clustered them using CD-hit (version 4.8.1)⁵¹ with >90% identity. The protein sequences of the RLXs from each class were aligned with MAFFT (version 7.271)¹⁷⁷ with option “*--maxiterate 1000*”, and the multiple alignments were trimmed using trimAl (version 1.5.0)³¹⁷, removing sequences with >50% gaps by using option “*-gt 0.5*”. The phylogenetic analyses were performed on the resulting multiple alignments using IQ-TREE (version 2.0.3)¹⁷⁸ with the ultrafast bootstrap option (1000 bootstraps)¹⁷⁹ to assess branch support. The best fitting model for each class of RLX was estimated using ModelFinder Plus¹⁸⁰, according to the Bayesian information criterion (Figure IV-7, Supplementary Figures S-IV-9 to 15).

G) Antibiotic-resistance gene prediction

NCBI AMRFinder (version 4.0.3) was applied to identify antimicrobial resistance in the metagenomic contigs using the nucleotide ARG database from NCBI³¹⁸. A restrictive threshold of 99% minimum identity and coverage was applied for ARG detection, considering only predictions that met these criteria as valid ARGs. This approach was designed with the purpose of investigating only the presence of known, annotated ARGs prevalent in clinical isolates, aiming to identify proteins with a confirmed role in resistance against specific substrates. The goal was to exclude proteins from families with divergent functions that do not provide resistance to the same substrates. Additionally, we retained the ARG family annotation information to facilitate downstream analyses and ensure consistent classification of these resistance determinants.

H) Detection of plasmid-specific sequences

The geNomad tool (version 1.5.2)³¹⁹ was used to classify the metagenomic contigs from all the microbiomes according to their molecular origin. Briefly, this tool catalogues biological sequences based on gene content by comparison against a marker set of 227,897 protein profiles specific to plasmids, viruses, or chromosomes. Some of these profiles contain hallmark markers related to core functions, such as conjugation genes for plasmids and capsid proteins for viruses. We used an average profile-specificity measure of 0.9 as threshold to assign a molecular origin to a contig (Figure IV-10).

I) Detection of prevalent marine plasmids

We aimed to assess whether the pLA6_012 and pP72_e plasmids, pervasive in the ocean and previously described^{286,287}, could be identified in the marine metagenomes from the *Tara* Oceans Expedition. To do this, we used BLAST+ (version 2.15.0)¹³⁹ to align the oceanic contigs against the pLA6_012 plasmid (GenBank accession number CP031597.1) and the pP72_e plasmid (GeneBank accession number CP010740.1). We filtered the matching contigs to retain only those aligning with >99.9% identity, an E-value=0 and an alignment length >4,000 bp between plasmid and contig. We used BRIG (version 0.9.5)³²⁰ to visualize the resulting aligned contigs as ringed, circular graphs (Supplementary Figure S-IV-8).

J) R packages and statistical analysis

Data manipulation was performed using R (version 4.1.3) and the *tidyverse* package (version 1.3.2)¹⁸³. Figures were generated with the R packages *ggplot2* (version 3.4.2)¹⁸⁴ and *ggpubr* (version 0.6.0)¹⁸⁵. Mann-Whitney tests with a Benjamini-Hochberg adjustment for multiple testing were applied on the global P_{rlx} (Figure IV-4B) and the *rpc* from the five environments (Figure IV-10B) to assess the significance of the differences. Statistical analyses and graphical representation on the plots were conducted using the R package *rstatix* (version 0.7.2)¹⁸⁷. Phylogenetic trees were rooted using the “*midpoint.root*” function from *phytools* package (version 2.1.1)³²¹ and represented using *ggtree* (version 3.2.1)³²².

3. Results

A) Relaxases are infrequent in marine bacterial genomes

To inspect the prevalence of RLXs in marine genomes and their distribution in the different MOB classes, we used Prodigal¹⁸², as described in the Materials and Methods section, to predict the ORFs encoded in a filtered collection of 45,624 marine bacterial MAGs and in three additional bacterial MAG datasets from the GM of superior vertebrates, which were used for comparison: 288,062 MAGs from humans, 3,921 from pigs and 13,360 from chickens. From the ORF collections, we predicted the RLXs by using MOBscan³¹⁴, as detailed in Materials and Methods. Only 5,515 marine MAGs (12.1%) contained at least one RLX, while 163,779 (56.9%), 1,296 (33.1%) and 5,971 (44.7%) human, pig and chicken GM MAGs were, respectively, MOB+ (Figure IV-3A). This represents a large difference in abundance that is unlikely to be due to chance. Additionally, we obtained the MOB-class distributions in the marine and the GM MAGs and observed remarkable differences between them. Namely, although class MOB_P was ubiquitous across both biomes, marine MAGs were enriched in classes MOB_B and MOB_F compared to GM MAGs, as well as MOB_C and MOB_H to a lesser extent. On the contrary, they were depleted in classes MOB_V and MOB_T (Figure IV-3B).

As recently suggested, plasmid-driven gene transfer occurs mainly within bacterial orders¹⁶⁵. Hence, we investigated the taxonomy of these MAGs and MOB distribution at order level to explore differences between both biomes. MAG taxonomic composition was significantly different between the oceans on one side, dominated by orders Flavobacteriales,

Pseudomonadales and the SAR86 clade, and the GM on the other, colonized by anaerobic Oscillospirales, Lachnospirales and Bacteroidales (Figure IV-3C).

The few marine RLXs were mainly distributed across MAGs from Rhodobacterales, Flavobacteriales and Pseudomonadales, while they were present to a lesser extent in Enterobacterales and Actinomycetales. MOB+ MAGs from each order were encompassed within a single bacterial family. However, those from order Pseudomonadales and Enterobacterales were distributed across multiple families (Supplementary Figure S-IV-1). We also observed that, with the exception of Pseudomonadales and Enterobacterales, each bacterial order was predominantly associated with a specific MOB class, in addition to the ubiquitous MOB_F and MOB_P classes (Supplementary Figure S-IV-2).

To ensure a fair comparison between the four collections of MAGs, we verified their average similarity in genomic length, completeness and contamination, following standardized metrics⁵⁷. MAGs from the four datasets had an average length of 2.2 Mbp and were <1% contaminated (Supplementary Figure S-IV-3). However, whereas average completeness of GM MAGs was 90%, that of marine MAGs dropped below 85%. These differences in genomic completeness may translate to an underestimation of the real number of marine RLXs, leading to an imprecise inference of mobilization by conjugation in the sea³²³.

Another uncertainty that could further contribute to this imprecision is the potential sampling bias. In fact, there are numerous challenges that complicate plasmid recovery by MAG reconstruction from metagenomic contigs through short-read binning strategies. Metagenomic binning groups contigs from the same source genome based on their similar relative abundance and nucleotide composition. However, plasmids may have different copy number and sequence distribution compared to chromosomes, making them difficult to recover in a MAG³²⁴.

Long-read sequencing methods can be used as an alternative to extract them from microbial communities³²⁵, although these methods have been reported to miss small plasmids (i.e., <10 kilobases) unless used in hybrid-assembly strategies with Illumina libraries³²⁶. Notably, some of the RLXs detected in the MAGs could originate from ICEs, as stated before.

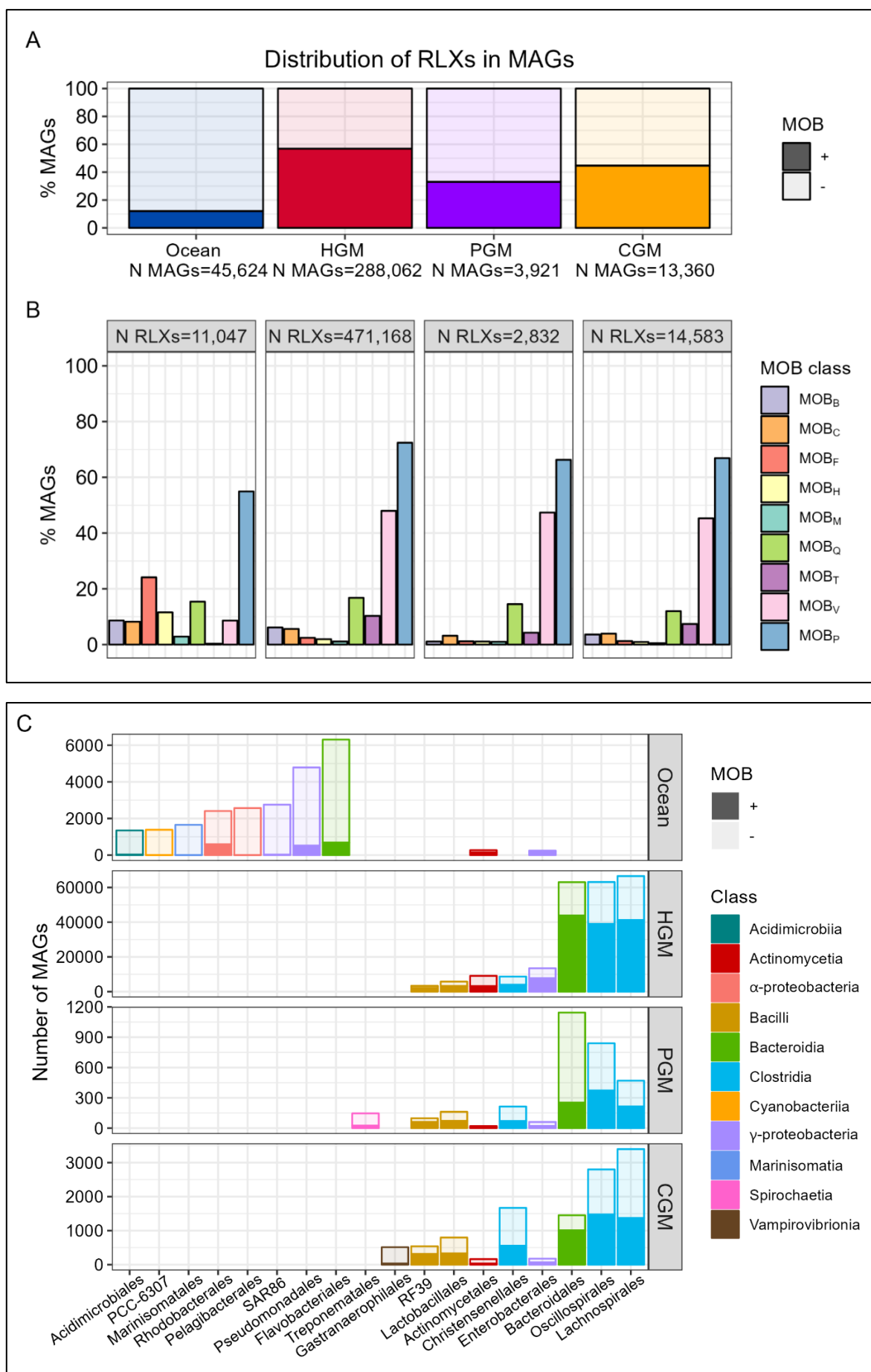


Figure 0-3: Distribution of RLXs in marine and vertebrate GM MAGs.

(A) Proportion of MAGs colored by biome and MOB presence. Bars represent the total number of MAGs from each dataset. The shaded part of the bars corresponds to the proportion of MOB+ MAGs (i.e., MAGs encoding a RLX) in each biome, whereas the clearer part of the bar corresponds to the proportion of MOB- MAGs. The total number of MAGs from each biome is indicated below the x-axis. **(B)** Proportion of MOB+ MAGs in each biome split and colored by MOB class. MAGs with multiple RLXs from the same or different MOB classes were counted only once per class. **(C)** Taxonomic composition of the bacterial MAGs analyzed in this study as annotated in MGNify. Bars represent the absolute abundance of MAGs from each bacterial order, colored by class. As in panel A, the shaded part of the bars corresponds to the proportion of MOB+ MAGs in each order. Orders encompassing <3% of the total MAGs were removed for visualization purposes.

B) Relaxases are scarce in aquatic ecosystems

To establish a precise measure of RLX abundance in the ocean and bypass the limitations encountered in MAG-based analyses, we studied 354 marine metagenomic sequencing libraries from samples collected across all oceans by the *Tara* Oceans Expedition³⁰⁹ (Supplementary Figure S-IV-4). We performed quality control of the sequencing reads, assembled them *de novo* into contigs, and then predicted their complete ORFeomes and RLXs as described in Materials and Methods. To accurately capture the absolute abundance of each ORF within a metagenomic sample, we weighted each ORF by the sequencing coverage of the contig on which it was encoded.

We represented RLX abundance as the probability of an ORF being a RLX, i.e., $P_{rlx}=p(\text{RLX} \mid \text{ORF})$, and we compared these results with the same probability in rivers, agricultural soils, sewage and in the GM of marine and non-marine mammals (i.e. whales and humans) (Table IV-1). The following numbers of RLXs were predicted in each dataset: 8,062 (ocean), 157 (rivers), 21,155 (soil), 113,945 (sewage), 1,084 (whale GM) and 11,942 (human GM). Results revealed that RLXs are markedly depleted in the ocean ($P_{rlx}=2.4\text{E-}05$) compared to other environments. Specifically, RLX levels in the ocean were over 50-fold lower than in the sewage ($P_{rlx}=1.7\text{E-}03$), and over 100-fold lower than in both whale ($P_{rlx}=2.8\text{E-}03$) and human GM ($P_{rlx}=4.1\text{E-}03$) (Figure IV-4B). In addition, RLX abundance in cultured lands ($P_{rlx}=1.5\text{E-}04$) was only about 6-fold higher than in the sea, whereas fluvial ecosystems ($P_{rlx}=3.6\text{E-}05$) exhibited nearly the same abundance as marine environments. Differences in RLX abundance between ocean, sewage and human GM were mainly driven by the MOB_P, MOB_V and MOB_Q classes (Supplementary Figure S-IV-5). In the human GM, $P_{MOBP}=4\text{E-}03$ and $P_{MOBV}=8.5\text{E-}04$, that is, two and three orders of magnitude higher than in the ocean, respectively.

To confirm that abundance differences between environments were not due to technical issues derived from the metagenomic assemblies, we analyzed the abundance of multiple USCGs, applying the same HMM coverage and E-value thresholds as described above (Table IV-2). Analogously, we defined their abundance as the probability of an ORF being a USCG, i.e., $P_{\text{USCG}} = p(\text{USCG} | \text{ORF})$. As expected, we observed minimal variation in P_{USCG} across all USCGs and environments (average $P_{\text{USCG}} = 2.4\text{E-}04$, Figure IV-4A). Protein prediction in short-read-based metagenomic assemblies is constrained by both the sizes of the recovered contigs and the ORFs detected within them, so the number of proteins predicted in contigs is directly proportional to their size. To evaluate whether our detection method was biased towards detecting short RLXs in the metagenomic data, we predicted P_{RLX} in bins of increasing ORF size. We determined that the average size of the RLXs predicted by MOBscan in the metagenomes analyzed in this study was 353 residues, whereas the average ORF size was 289 (Figure IV-5), confirming the precision of the methodology.

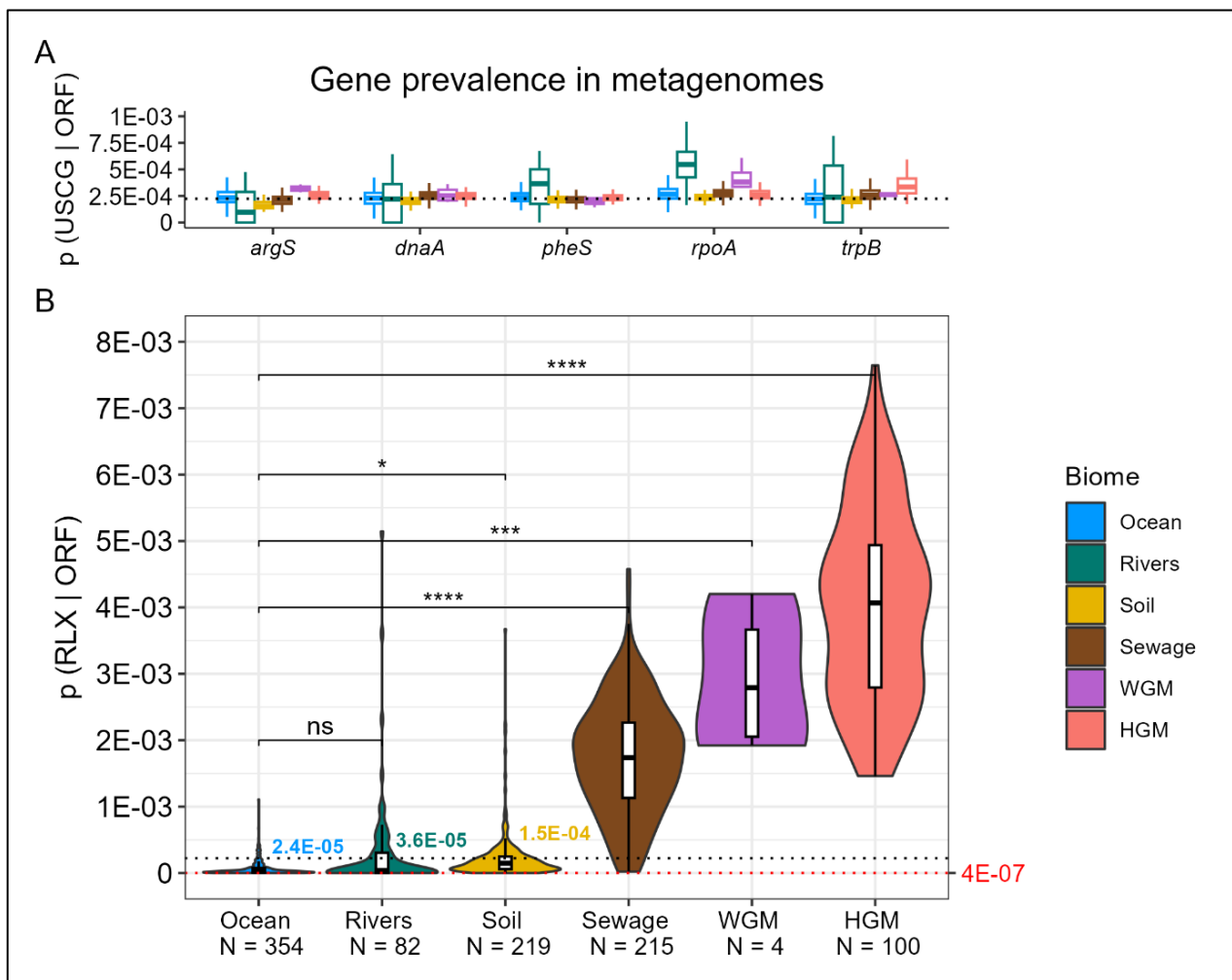


Figure 0-4: RLX and USCG prevalence in the metagenomes.

Violin-boxplots represent **(A)** P_{uscg} and **(B)** P_{rlx} in each biome. Differences in P_{rlx} were evaluated using pairwise Mann-Whitney tests. The horizontal, black-dotted line across both plots indicates the average P_{uscg} across all USCGs and environments. In panel B, the horizontal, red-dotted line marks the minimum, non-zero P_{rlx} in the metagenomic samples. Horizontal black bars represent median levels, colored to match the corresponding violin plots for ocean, river and soil biomes. Boxes and whiskers indicate the data from first to third quartiles and from the quartiles to the minimum and maximum, respectively. Sizes of the USCG-encoded proteins and the total number of samples analyzed from each biome are indicated below the x-axes. (*) $p < 0.05$, (***) $p < 0.001$, (****) $p < 0.0001$, ns: non-significant.

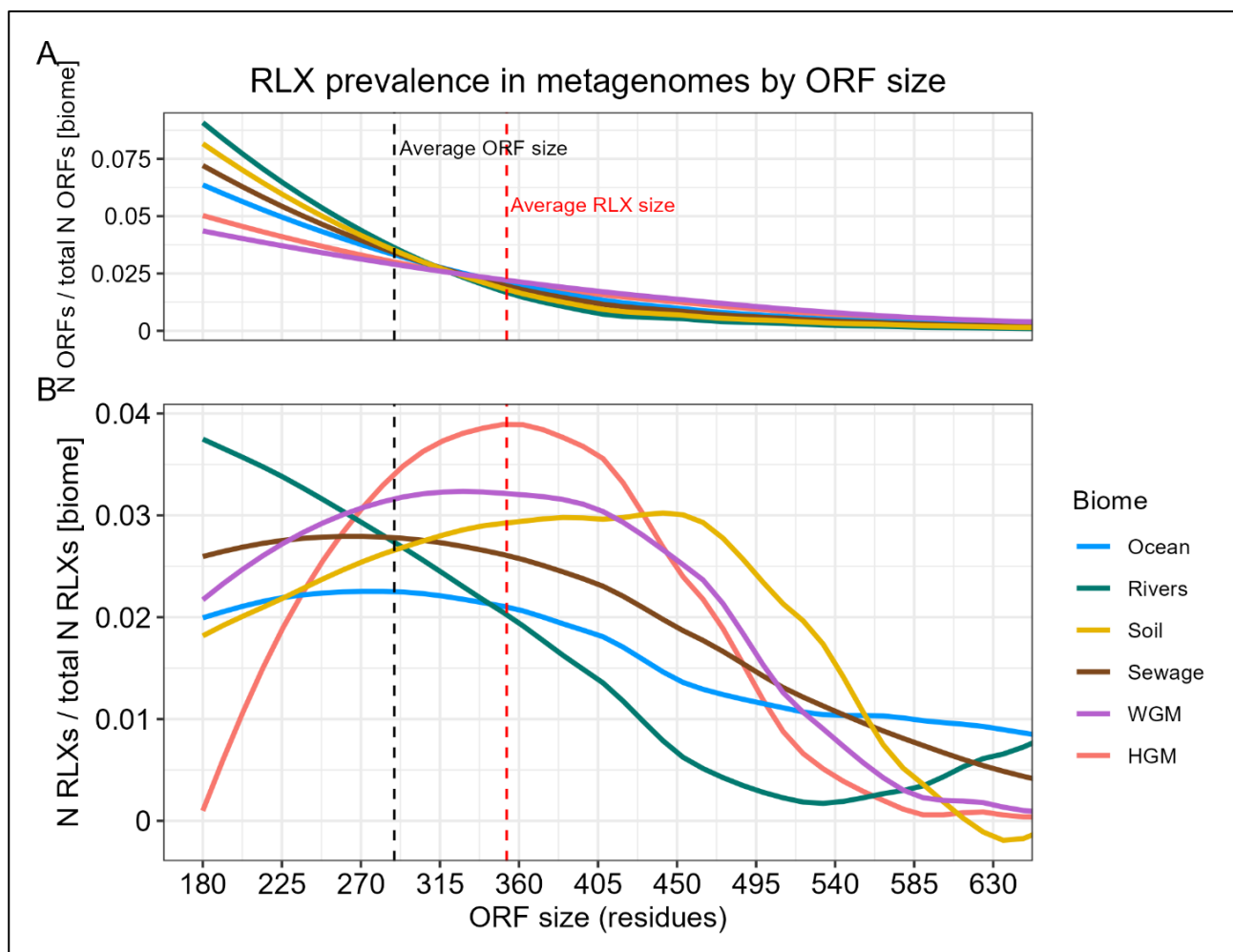


Figure 0-5: RLX prevalence in the metagenomes, by ORF size.

Line plots represent **(A)** the total number of ORFs and **(B)** RLXs predicted in each bin from every biome, normalized by the total number of ORFs and RLXs from all biomes, respectively. x-axis is split in bins of size $= 180 + 10n$ residues, where n is the bin number (starting by bin0=180). Data points were smoothed through a LOESS regression, and confidence intervals were removed for better visualization. Vertical dashed lines indicate the average ORF (black) and RLX (red) sizes, respectively. Over 95% of both the ORFs and the RLXs detected in this study were < 630 residues, so we chose this threshold as the upper ORF-size limit also for visualization purposes. LOESS: locally-estimated scatterplot smoothing.

Similarly to the MAG analysis, we investigated which bacterial orders were predominant in the metagenomic samples. To do this, we used Kraken (version 2.1.2)³⁵ to align the contigs against the “standard” Refseq database collection, including Archaea, Bacteria, virus and plasmids (version 06/05/2023). In agreement with the taxonomic analysis at MAG level, orders Flavobacteriales, Pseudomonadales and Pelagibacteriales were predominant in the ocean. In contrast, the sewage was populated by several Gammaproteobacteria orders, while fluvial and soil ecosystems were colonized by multiple orders from Actinomycetia, as well as some Gammaproteobacteria. As observed previously, the order Bacteroidales and class Clostridia were predominant in both the whale and human GM, which, not surprisingly, were also found in sewage (Figure IV-6).

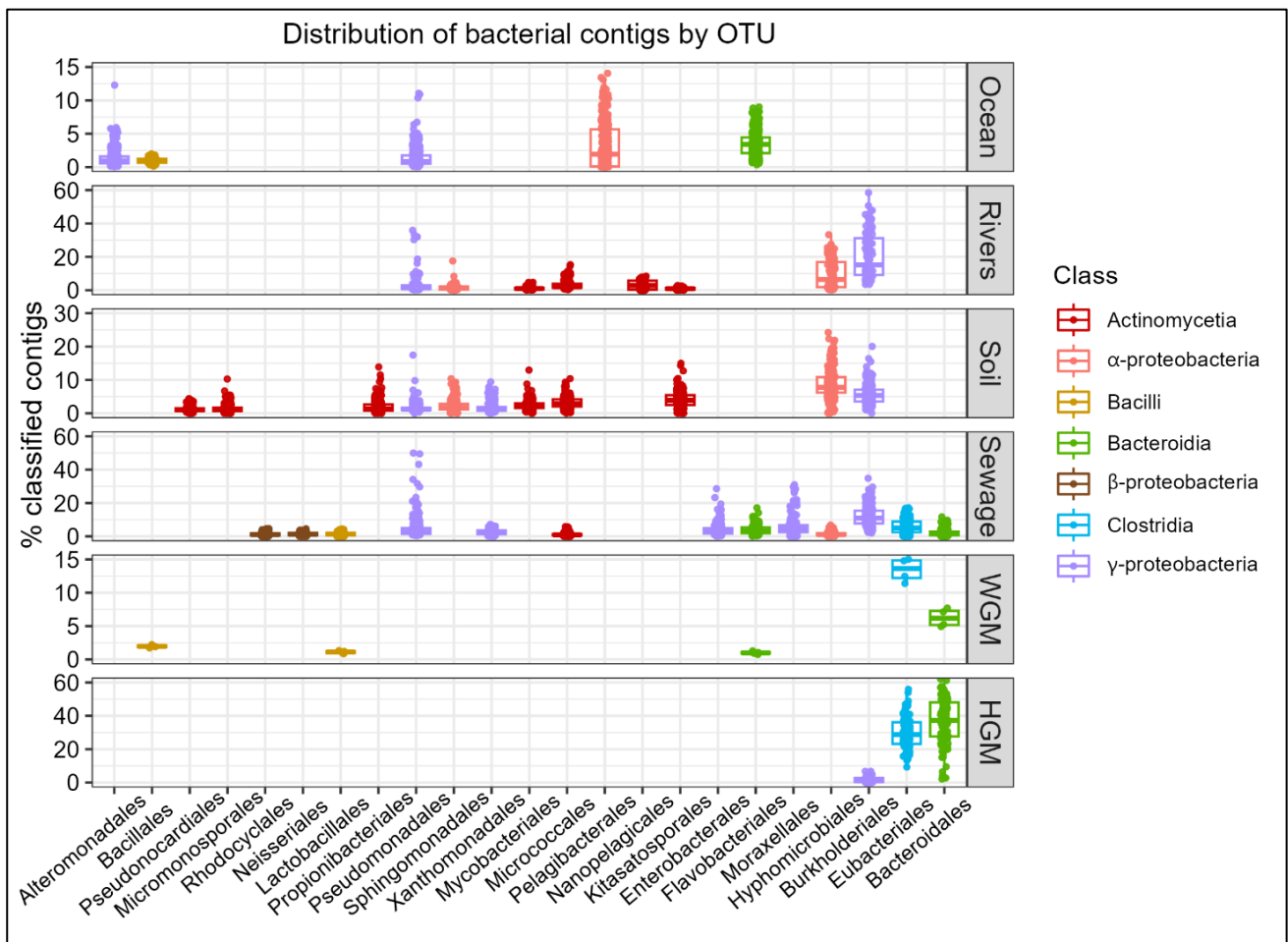


Figure 0-6: Taxonomic composition of bacterial contigs.

Boxplots represent the percentage of classified bacterial contigs from the metagenomic assemblies attributed to each order, colored by class. Orders ascribed with <1% of the total classified contigs, in average, were removed for visualization purposes. OTU: operational taxonomic unit.

C) Oceanic relaxases are phylogenetically diverse

To further investigate the evolutionary relationship of oceanic RLXs with those of the other biomes, we clustered the metagenomic RLXs from all MOB classes but MOB_T with those annotated in RefSeq200^{21,165} and performed phylogenetic analyses as detailed in Materials and Methods. The MOB_B RLX tree showed two differentiated clades, one enriched in GM MOB_B RLXs and the other one in marine MOB_B RLXs (Figure S-IV-7). This clear phylogenetic separation between both samples contrasts with the broader distribution of MOB_B RLXs from sewage and agricultural soil. The tree also showed that oceanic MOB_B RLXs are mainly located in chromosomes. Similarly, oceanic MOB_M RLXs primarily clustered with chromosomal representatives, but were sparsely distributed throughout the tree and intermixed with RLXs from soil and sewage (Supplementary Figure S-IV-12).

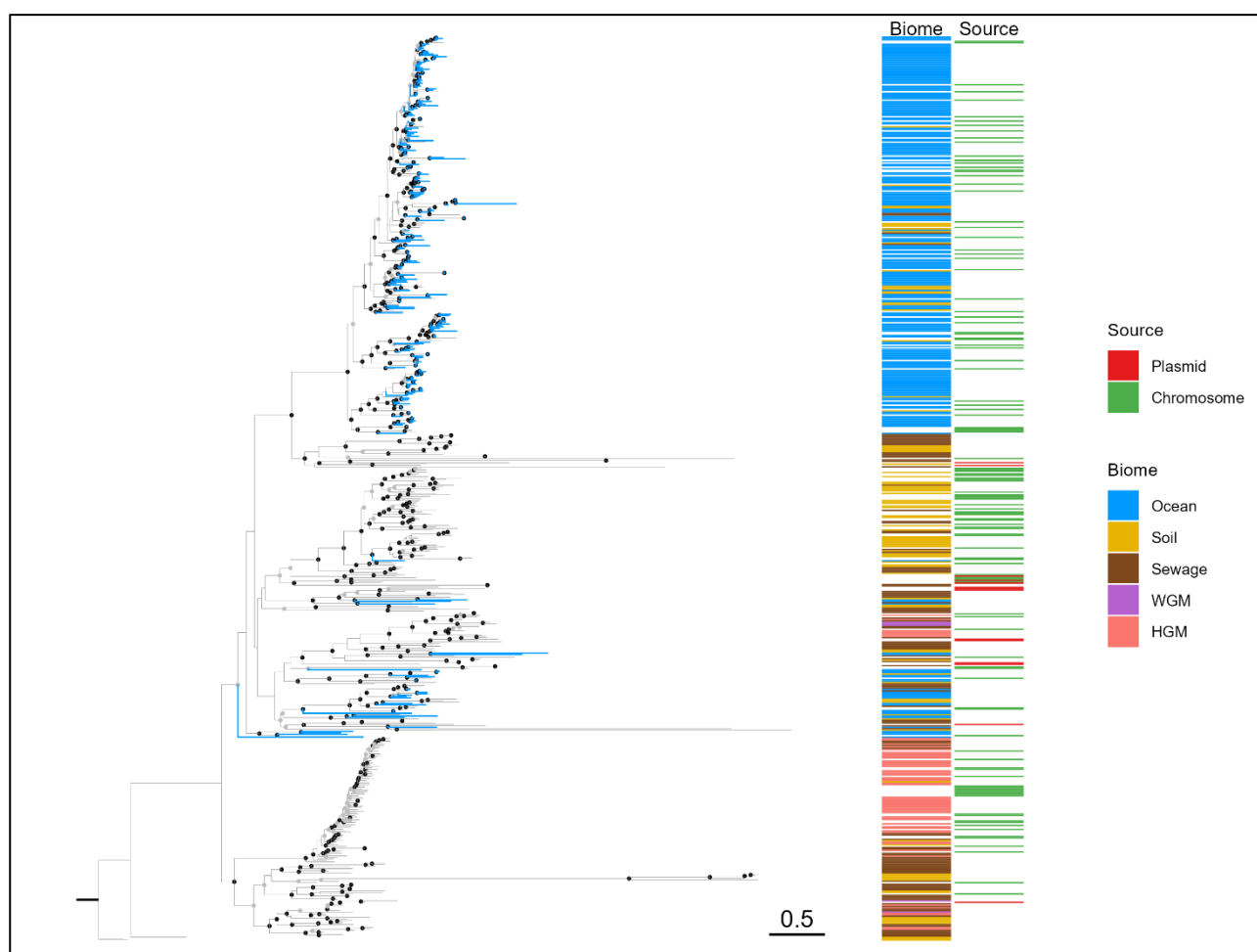


Figure 0-7: Phylogenetic tree of MOB_B RLXs.

The tree was built using 646 MOB_B RLXs with maximum likelihood with IQ-TREE¹⁷⁸ (model Q.yeast+R8 according to BIC and 1000 ultrafast bootstraps). RLXs were retrieved with the MOB_B HMM profile from MOBScan³¹⁴ from the metagenomic samples as detailed in Material and Methods, and those annotated

in RefSeq200¹⁶⁵. Ultrafast bootstrap values superior to 75 are shown with a light grey circle and values superior to 90 with a black circle. Blue branches indicate RLXs from marine origin, whereas grey branches indicate the rest of RLXs. The trees were rooted using the midpoint root, as detailed in Material and Methods. The inner heatmap indicates the biome where each RLXs was detected, and white tiles correspond to RLXs from RefSeq200. The outer heatmap indicates the genomic source where each RLX has been annotated in RefSeq200, and white tiles from this heatmap correspond to RLXs detected in metagenomes. Evolutionary distance of the branches is indicated next to the heatmap.

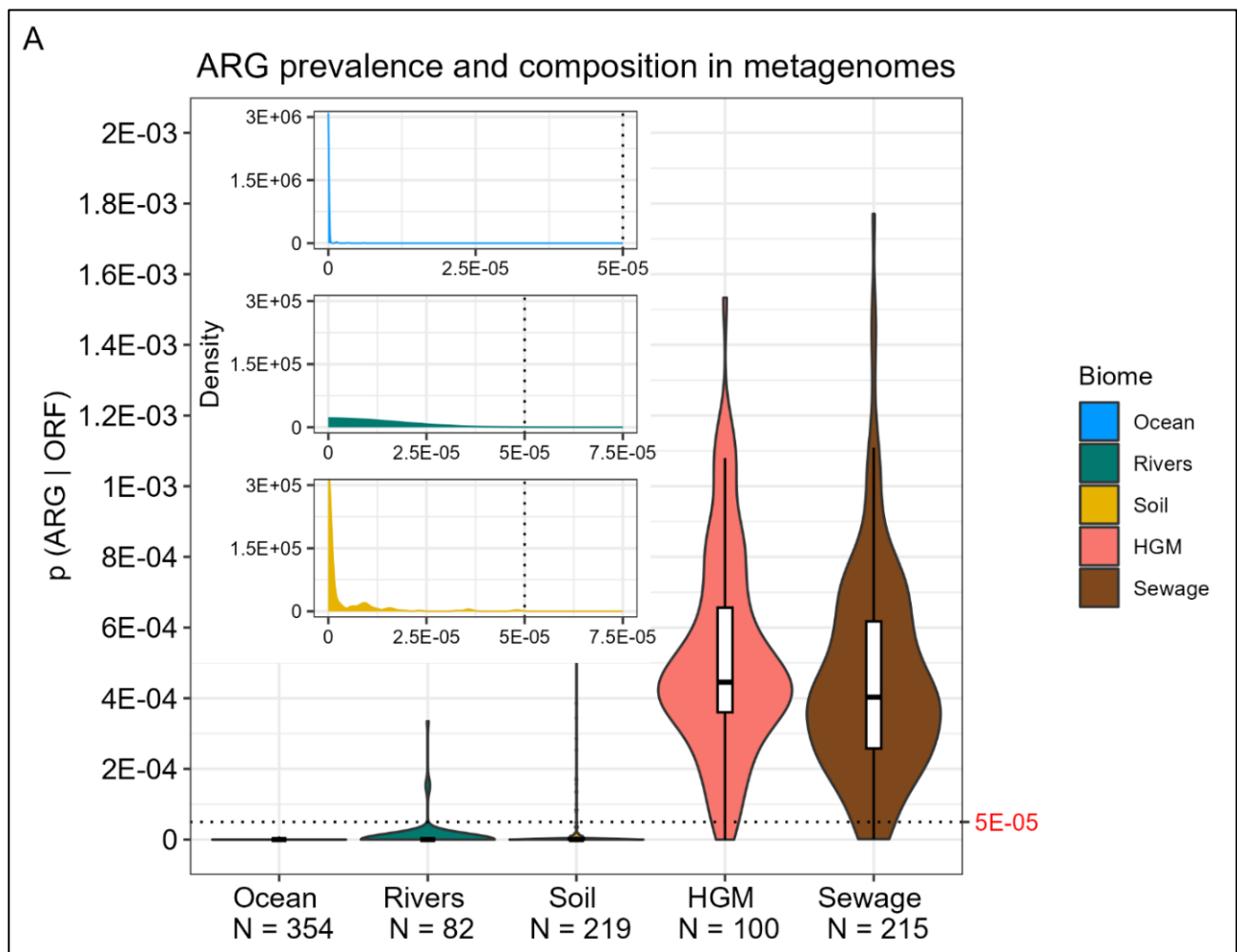
Contrary to the specialization of oceanic MOB_B RLXs, MOB_C, MOB_F, MOB_H, MOB_P, MOB_Q, and MOB_V from marine samples were widely distributed in their respective phylogenetic trees (Supplementary Figures S-IV-9 to 11 and Supplementary Figures S-IV-13 to 15). Most of MOB_C, MOB_H, and MOB_F RLXs were located in clades where chromosomal representatives are more prevalent, while oceanic MOB_Q, MOB_V, and MOB_P RLXs were more evenly distributed across both chromosomal and plasmid contexts. Although marine RLXs from these six MOB classes were generally grouped in mixed clades with RLXs from other biomes, some clades were relatively enriched in oceanic RLXs. Interestingly, clades containing plasmid-derived oceanic MOB_F and MOB_H from a plasmid origin also included RLXs from sewage. MOB_P, MOB_Q, and MOB_V trees showed a pronounced abundance of RLXs from sewage sources, consistently with Supplementary Figure S-IV-5, resulting in an intermingled grouping of oceanic and sewage RLXs.

D) Antibiotic resistance genes are depleted in the ocean

The RLX analysis described above allowed us to obtain a broad picture of the relevance of MAP mobilization dynamics. To further understand the dissemination of ARGs in the ocean, we studied ARG abundance in the metagenomic samples with AMRFinder³¹⁸ as detailed in Materials and Methods. Analogously as P_{rlx} , ARG abundance was represented as the probability of an ORF being an ARG, i.e., $P_{arg}=p(\text{ARG} \mid \text{ORF})$. We observed that ARGs were, in average, depleted in ocean, rivers and soil ($P_{arg}=0$), while similarly abundant in sewage ($P_{arg}=4E-04$) and the human GM ($P_{arg}=4.5E-04$) (Figure IV-8A). No ARGs were detected in the whale GM samples.

Additionally, we profiled the ARG family composition in each biome by determining the absolute number of ARGs per family based on AMRFinder annotations. For each sample, we recorded whether it contained at least one representative of an ARG family. To focus on the most common ARG families, we retained only those present in at least 20% of samples in at least one biome, and we observed environment-specific ARG family patterns (Figure IV-8B). In

the human GM, predominant ARG families were *tet* (tetracycline-resistance), *mef* (macrolide-resistance), *lnu* (lincosamide-resistance), *erm* (macrolide–lincosamide–streptogramin B resistance), *dfr* (trimethoprim-resistance), *cfxA* and *cbIA* (β -lactamases), and *ANT(3)* (aminoglycoside-resistance). Notably, *cbIA*, although prevalent in human GM, was almost absent in sewage. In contrast, *blaTEM* and *blaOXA* were present at only low levels in marine, river, and soil samples but were highly prevalent in sewage. Furthermore, several ARG families, including *tet*, *sul1* and *sul2* (both sulfonamide-resistance genes), *blaTEM*, *APH(6)* and *APH(3)* (aminoglycoside phosphotransferases), and *ANT(3)*, were commonly detected in both soil and sewage samples. Overall, while ARGs were present to some extent in the human GM but almost depleted in aquatic environments, the findings demonstrate that sewage samples are particularly enriched with a diverse array of ARG families. This underscores the significant anthropogenic impact on the dissemination of clinically relevant antimicrobial resistance determinants across different biomes.



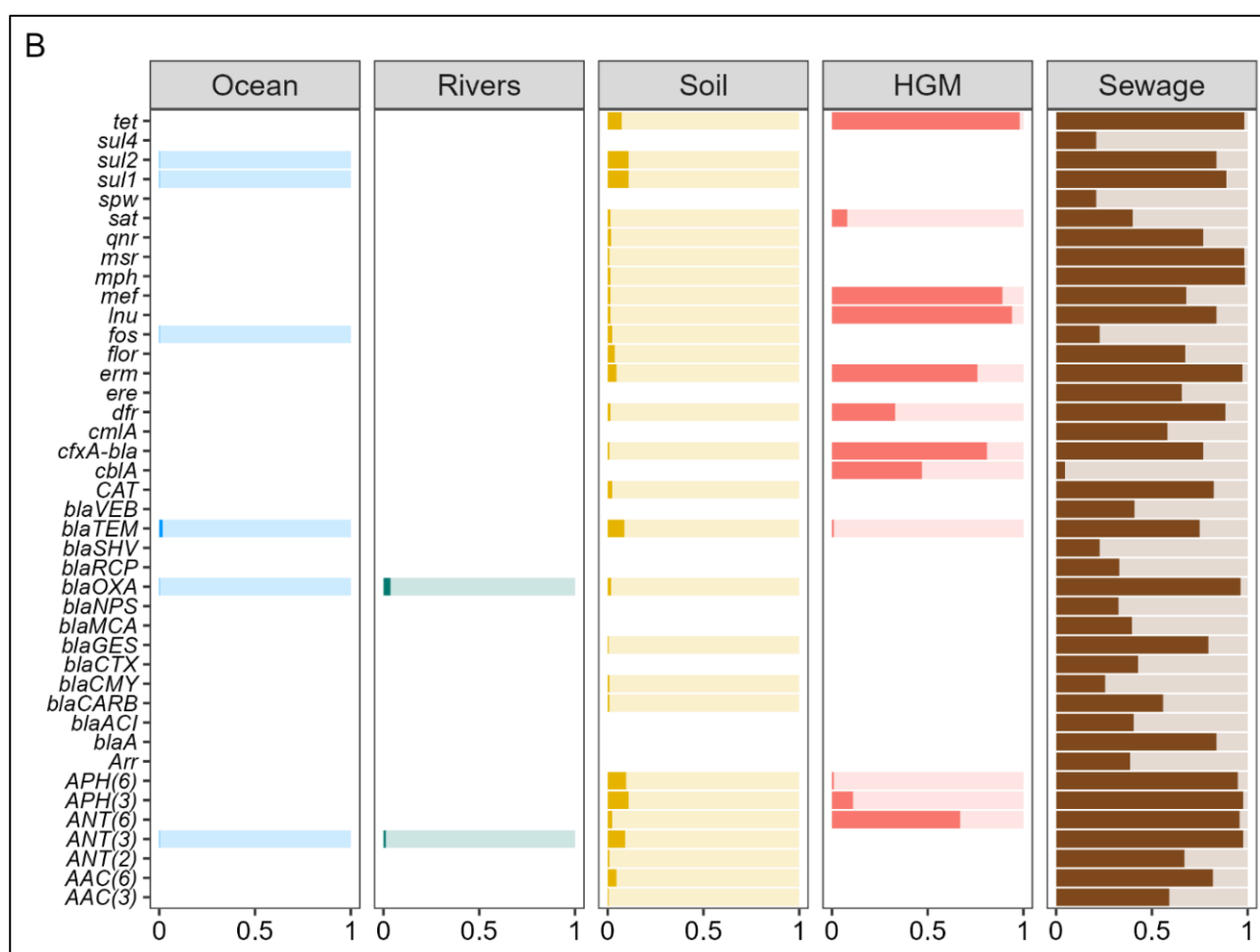


Figure 0-8: ARG prevalence and composition in the metagenomes.

(A) Boxplots represent ARG prevalence, expressed as P_{arg} , in each biome. A close-up panel displays kernel density estimates of P_{arg} values for oceans, rivers and agricultural soils. In the close-up, the x-axis represents $p(ARG|ORF)$ -the proportion of predicted ORFs annotated as ARGs-, and the y-axis shows the estimated density, reflecting the relative frequency of P_{arg} values across samples within each biome. The horizontal, black-dotted line across both the main plot and the close-up indicates the y-axis limit included in the close-up. No ARGs were detected in the whale GM samples. **(B)** ARG family composition in each biome. Bars represent the fraction of metagenomic samples encoding at least one ARG from each of the families indicated on the y-axis. The shaded portion of each bar corresponds to the proportion of ARG+ samples from each family in each biome. Only ARG families present in at least one biome with >20% ARG+ samples were kept for visualization purposes.

E) Genomic and metagenomic ORF abundances are correlated

To determine how accurately ORF abundances were estimated in a microbiome, we identified reciprocal best hits between genomic and metagenomic ORFs in synthetic microbial communities using the *easy-rbh* function from MMseqs (version 15.6)⁵². ORFs predicted by Prodigal from each reference genome were compared against ORFs predicted by MetaProdigal from assembled contigs. Alignments were filtered to retain only high-confidence matches with 100% identity and alignment lengths between 180 and 630 amino acids, corresponding to the expected size range for RLXs (Figure IV-5). Additionally, only ORF pairs with identical amino acid lengths were kept. This stringent filtering yielded a set of one-to-one perfect matches, pairing theoretical genomic ORF abundances with observed metagenomic ORF quantities (i.e., estimated from the coverage of their parent contigs). Because genomic ORF abundances were expressed as relative values and metagenomic ORF counts as absolute values, the latter were rescaled to the [0,1] range using min-max normalization (Equation IV-1) to enable direct comparison between the two abundance estimates.

$$X_{scaled} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad \text{and} \quad \Delta_{abundance} = \left| x^{(genomic)} - x_{scaled}^{(metagenomic)} \right|$$

Equation IV-1: Min-max normalization and abundance difference calculation.

Metagenomic ORF abundance values were scaled to the [0,1] interval using min-max normalization (left). The absolute difference ($\Delta_{abundance}$) between each matched ORF pair was then computed to quantify the discrepancy between genomic and metagenomic abundance estimates.

Figure IV-9 shows the distribution of $\Delta_{abundance}$ across all reciprocal best-hit ORF pairs. Most ORFs cluster around $\Delta = 0$, indicating strong concordance between metagenomic abundance estimates based on contig coverage and the known genomic relative abundances. A minor tail of slightly larger Δ values likely reflects cases of assembly fragmentation or coverage bias, particularly in low-abundance genomes. Overall, the narrow distribution centered around zero supports our ORF quantification approach, suggesting that contig coverage serves as a valid proxy for true genomic ORF abundance. Unexpectedly, the error distribution broadens for the lowest-complexity community (D), suggesting reduced estimation accuracy under these conditions.

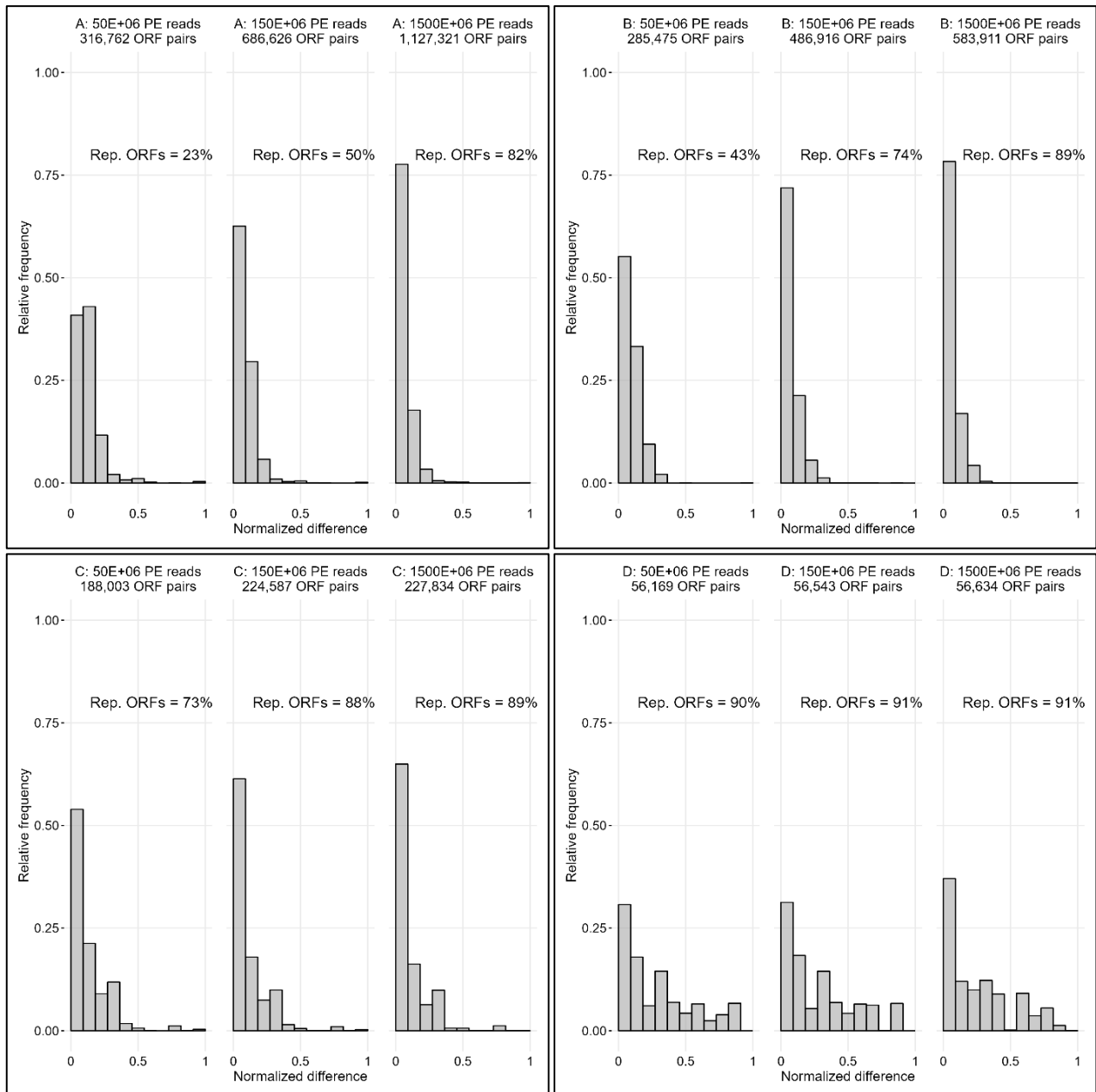


Figure 0-9: Comparison between genomic and metagenomic ORF abundances.

Distribution of $\Delta_{\text{abundance}}$ values, defined as the absolute difference between normalized genomic and metagenomic ORF abundances, across all reciprocal best-hit ORF pairs in the mock communities. Genomic ORF abundances were derived from the known relative abundance of the source genomes, while metagenomic ORF abundances were estimated from assembler-reported contig coverage. To allow direct comparison, metagenomic ORF abundances were rescaled to the [0,1] range using min-max normalization (Equation IV-1). Each panel represents a synthetic community, and each facet shows the sequencing coverage and the number of matched ORF pairs. The percentage of original genomic ORFs that were successfully recovered from the community is also indicated.

F) Plasmid-specific sequences are limited in the sea

The relative lack of RLXs and ARGs in the marine assemblies led us to hypothesize that MAPs were infrequent in the sea. To test this, we classified the metagenomic contigs of all the microbiomes according to their molecular origin with geNomad³¹⁹. Plasmid content was half an order of magnitude lower in the ocean compared to the rest of the environments (Figure IV-10A). Proportions of molecular sequences were constant across every biome except for viral load, which was unsurprisingly high in the ocean^{327,328}. The ratio plasmid-contig/chromosome-contig (*rpc*) is significantly lower in the ocean (*rpc*=2.1E-03) than in the rivers (*rpc*=5E-03), the soil (*rpc*=3.5E-03), the human GM (*rpc*=0.013), the sewage (*rpc*=0.024) and was also lower, although not significantly, in the whale GM (*rpc*=7.1E-03) (Figure IV-10B).

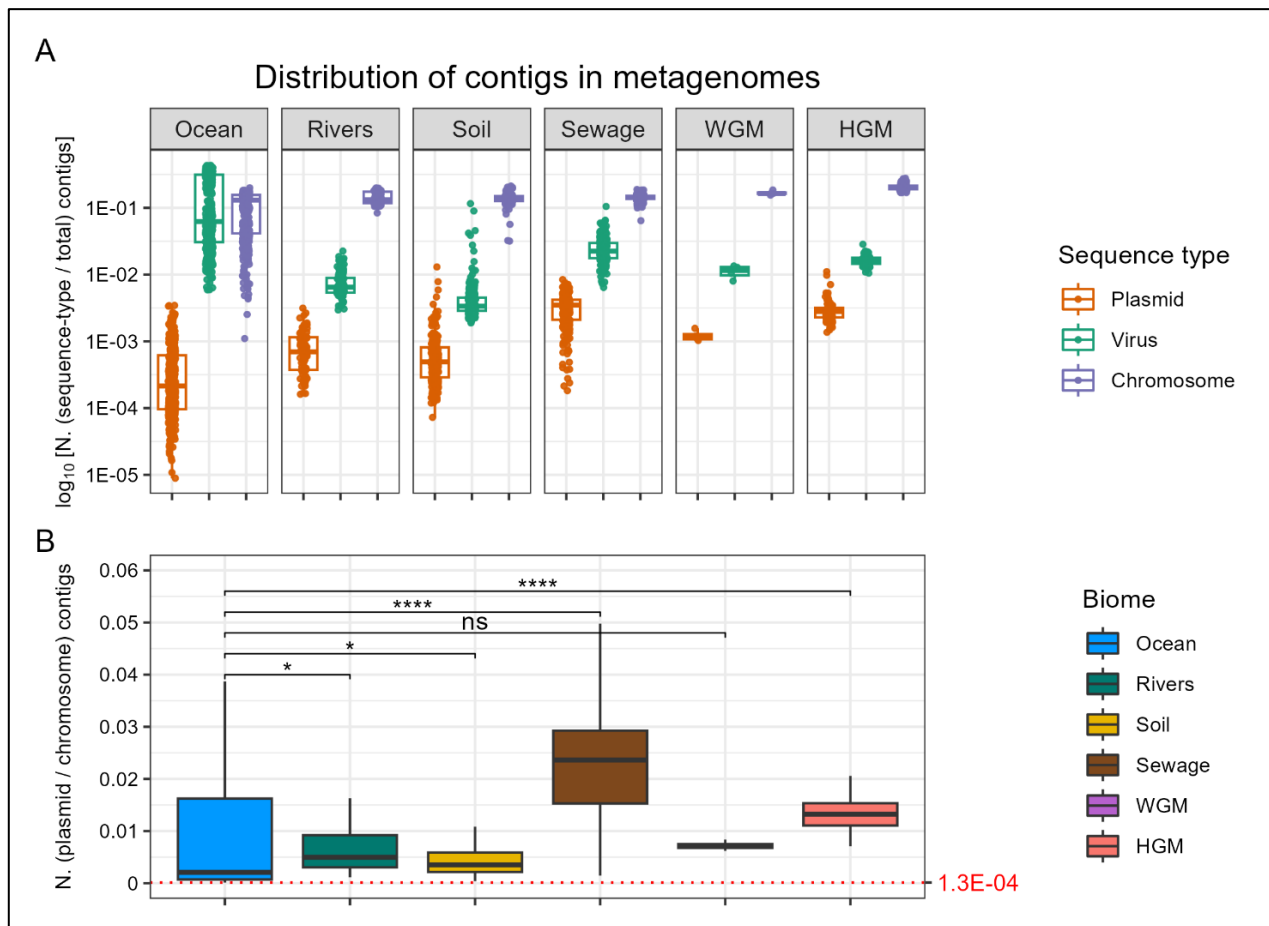


Figure 0-10: Genomic content classified by molecular origin in the metagenomes.

(A) Boxplots represent the ratio between the absolute abundance of metagenomic contigs predicted as plasmid, viral or chromosomal, and the total number of contigs in each biome, shown on logarithmic scale. **(B)** Boxplots represent the ratio between the absolute abundance of plasmid and chromosomal contigs (*rpc*). The horizontal, red-dotted line across panel B indicates the minimum, non-zero *rpc* value in the metagenomic samples. Differences in *rpc* between biomes were evaluated using a pairwise Mann-Whitney test. (*) $p < 0.05$, (****) $p < 0.0001$, ns: non-significant.

4. Discussion

In this work, we describe the abundance of plasmids in the oceans, their specificities compared to terrestrial plasmids, and the associated risk of ARG dissemination. To this end, we studied a set of marine MAGs to examine the distribution of RLXs (the gene proxy for plasmids) in oceanic bacteria. Only 12% of them encode a RLX, a notably lower fraction compared to 57%, 33% and 45% MOB+ human, pig, and chicken GM MAGs, respectively (Figure IV-3A). Specifically, MAPs are highly enriched in MOB classes MOB_B, MOB_F, MOB_C and MOB_H; and depleted in classes MOB_V and MOB_T (Figure IV-3B). A novel MOB_F MAP was recently discovered in a computational reconstruction of plasmids from marine metagenomic samples²⁹⁶, enhancing the relevance of this MOB class in the sea. Additionally, MOB+ MAGs from both biomes show a similar proportion of MOB_P RLXs.

Their bacterial composition at order level is also entirely different (Figure IV-3C). The sea is enriched in genomes from Flavobacteriales, Pseudomonadales and SAR86, which is an abundant, non-photosynthetic Gammaproteobacteria in the global surface ocean³²⁹ which remained uncultured until very recently³³⁰. Marinisomatales, Rhodobacterales, Pelagibacterales, photosynthetic PCC-6307 and Acidimicrobiales are also abundant marine clades absent in the GM. On the contrary, this environment is mainly populated by anaerobic Oscillospirales, Lachnospirales and Bacteroidales that thrive as commensal bacteria of the GM from superior vertebrates, as described in Chapter I (Figures III-4 III-8).

The low amount of RLXs found in marine MAGs is distributed across Flavobacteriales, Pseudomonadales and Rhodobacterales, as well as Actinomycetales and Enterobacterales (Figure IV-3C). These results indicate that the low RLX prevalence and differential MOB distribution in the ocean are strongly associated with a distinct bacterial taxonomic composition. Marine MOB+ genomes within each bacterial order are generally restricted to a single family, with the exceptions of Pseudomonadales and Enterobacterales, whose MOB+ MAGs are distributed across multiple families (Supplementary Figure S-IV-1). MOB+ genomes from these orders are scattered across multiple families, which indicates that marine, conjugative clades are diversified in these orders. Additionally, most bacterial orders are associated with a specific MOB class, except MOB_F and MOB_P, which are present across many orders (Supplementary Figure S-IV-2). An overview of NCBI bacterial database RefSeq200 further revealed that there are almost no annotated MOB_B plasmids (Supplementary Figure S-

IV-7), although we observe that they are an important RLX class in marine Flavobacteriaceae and *SAR86* (Supplementary Figures S-IV-1 and 2). These differences highlight the underrepresentation of MAP annotation in public, curated databases.

The lower completeness of marine MAGs analyzed in this study (Supplementary Figure S-IV-3), their potential sampling bias, and the difficulties to identify plasmids in MAGs led us to study metagenomic assemblies from the *Tara* Oceans Expedition to understand RLX prevalence in the ocean (Supplementary Figure S-IV-4). By establishing a probabilistic measure of RLX abundance (P_{rlx}), we observed a significant depletion of marine and terrestrial RLXs relative to sewage and mammal GM by two orders of magnitude (Figure IV-4B). This deficiency is primarily attributed to lower counts within MOB classes MOB_P , MOB_V and MOB_Q (Supplementary Figure S-IV-5). Concordantly, the lower number of MOB_V+ MAGs in the ocean compared to the GM observed in Figure IV-3B correlates with the lower P_{MOBV} in this environment.

In contrast, USCG abundance remains remarkably consistent across biomes, with an average $P_{uscg}=2.4E-04$ (Figure IV-4A), corresponding to roughly one USCG per 4000 ORFs. For comparison, approximately 8 and 16 correspond to RLXs in sewage and human GM, respectively. Moreover, while RLX abundance varies by two orders of magnitude among environments, USCG abundance displays far less variation. Additionally, marine RLXs are phylogenetically diverse (Figure IV-7, Supplementary Figures S-IV-9 to 15). Together, these findings confirm that the marked depletion of RLXs in marine environments is a specific feature that does not extend to the overall ORF composition, thereby reinforcing the robustness of our analysis.

There are several notable features in the quantification analysis of the metagenomic samples. First, we observed a wide dispersion of RLX abundance in sewage metagenomes (Figure IV-4B). This is probably due to the high heterogeneity of bacterial content present in the human-derived effluents of the different cities, each subjected to different conditions and hygienic standards (Supplementary Figure S-IV-4). Second, the higher abundance of RLXs in the sewage and GM metagenomes compared to that of USCGs may reflect the presence of high copy-number plasmids in these environments. Additionally, the P_{rlx} and P_{uscg} values of all USCGs measured in rivers showed greater dispersion compared to other environments, likely due to the poorer assembly metrics obtained in this biome (Figure IV-4).

And third, we validated the robustness of the P_{rlx} measure used in this study against metagenomic contig fragmentation, as evidenced by the higher P_{rlx} found among longer ORFs that compose a lower fraction of the total metagenomic ORFs (Figure IV-5). This is summarized by the different average ORF and RLX sizes detected in the assemblies. Hence, the RLX domain structured in HMM profiles from MOBScan proves to be conserved enough to reliably quantify RLXs in metagenomic contigs assembled from short reads. Figure IV-5B also highlights discordant size distributions between RLXs from the GM and the rest of the environments. Specifically, pronounced peaks of the regression curves in whale and human GM at 360 residues contrast with the smoother curves for the other biomes. This distinction is most likely due to the different compositional and biological diversity of both biomes, coupled to their disparate microbial heterogeneity (Figure IV-6) and global P_{rlx} (Figure IV-4).

We also confirmed that pLA6_012²⁸⁶ and pP72_e²⁸⁷ plasmids, recently described as pervasive in the ocean, are identified within the marine metagenomes (Supplementary Figure S-IV-8). Specifically, two contigs from different samples contain the complete sequence of pLA6_012. This control finding indicates that the metagenomic samples from the *Tara* Oceans Expedition are a valid dataset for MAP prediction.

Notably, using the RLX as a proxy for bacterial conjugation has the caveat of excluding the fraction of plasmids that are non-mobilizable and non-conjugative. While this limitation may lead to an underestimation of the total plasmid pool, previous studies -primarily focused on human-associated microbiomes- suggest that the contribution of non-mobilizable plasmids to the propagation of ARGs is relatively minor^{331–333}. More broadly, recent evidence indicates that the vast majority of known ARGs across environments are typically encoded on mobile, broad-host range plasmids³³⁴.

To evaluate whether this trend holds across diverse ecosystems, we complemented MAP-based mobilization results based on RLX analysis with an assessment of antimicrobial resistance dispersion in marine environments. To do so, we interrogated the same metagenomic samples about the abundance of ARGs with an analogous probabilistic measure: P_{arg} . Our results show that ARGs are virtually depleted in natural aquatic environments, but they reach their maximum levels in sewage and the human GM (Figure IV-8A). ARG levels are equally depleted in rivers and agricultural soil, and the minimal ARG content observed in the ocean is not transmitted to whales. Notably, river and soil outlier samples

exhibited P_{arg} levels similar to those in human GM and sewage. These results confirm previous observations reporting that water and soil ARG abundance, expressed as ARGs per 16S rRNA, were 2 orders of magnitude lower compared to human GM, whereas ARG abundance in human GM and sludge (which is probably similar in its composition than sewage) were around 0.17 and 0.26 copies of ARG per 16S rRNA, respectively¹³⁵.

Analysis of ARG family composition uncovered distinct environment-specific patterns (Figure IV-8B). Notably, the sewage is markedly enriched in a diverse collection of ARG families compared to other biomes. Among these, only three but clinically relevant ARG families - *blaTEM*, *blaOXA* (beta lactam resistance-genes), and *ANT(3)* (macrolide resistance-genes)- are detected in four out of five biomes. Also, *tet* (tetracyclin resistance-genes), *erm*, *APH(3)* and *APH(6)* (macrolide resistance-genes) are found in several soil metagenomic samples. The presence of *tet* as the most abundant ARG family in the human GM has already been observed¹³⁵. In fact, most of these ARG families have been previously categorized as widespread in multiple environments¹³⁵. This pattern underscores the significant impact of anthropogenic activities on the propagation of antimicrobial resistance. Furthermore, among freshwater samples, only 4 are ARG-positive, with 3 of them containing 6 copies of *blaOXA*. ARG-positive soil samples ($P_{arg} > 2E-04$) are predominantly located in eastern China (n=4), Mexico (n=2), and the US (n=2), and ocean ARG-positive samples are heterogeneously distributed across the globe (Supplementary Figure S-IV-6).

To further validate our ORF quantification approach, applied through the P_{rlx} and P_{arg} metrics, we conducted a controlled experiment using synthetic communities with known genome compositions and relative abundances (Table IV-3, Supplementary Figure S-IV-16). The resulting distribution of $\Delta_{abundance}$ values, defined as the absolute difference between normalized genomic and metagenomic ORF abundances, was tightly centered around zero (Figure IV-9). This narrow distribution confirms that our method reliably recovers expected ORF abundances and supports the use of contig coverage as a valid proxy for true genomic ORF abundance in metagenomic samples.

We also extended the analysis of the RLX and ARGs in the metagenomic samples as predictors of MAP deficiency by classifying the metagenomic contigs through a set of recently published protein profiles characteristics of MGEs³¹⁹. These marker sets include genes commonly associated with plasmids, such as those encoding conjugation proteins, secretion

systems and proteins involved in quorum sensing and motility. We observed that the *rpc* is significantly lower in the oceans than in the human GM and the sewage by one order of magnitude (Figure IV-10B), confirming that viral content is higher in the ocean compared to the rest of the environments (Figure IV-10A)^{327,328}. This means that MAP depletion could be an evolutionary consequence of the high prevalence of viruses, as it is known that plasmids are targeted by very diverse phages³³⁵. In any case, this result indicates that oceans are relatively free of plasmids, and they behave as different HGT platforms from the rest of microbiomes. It is worth noting that geNomad is subject to the same limitations of plasmid-signature based identification methods discussed previously. Consequently, some of the contigs identified as plasmids may originate from ICEs.

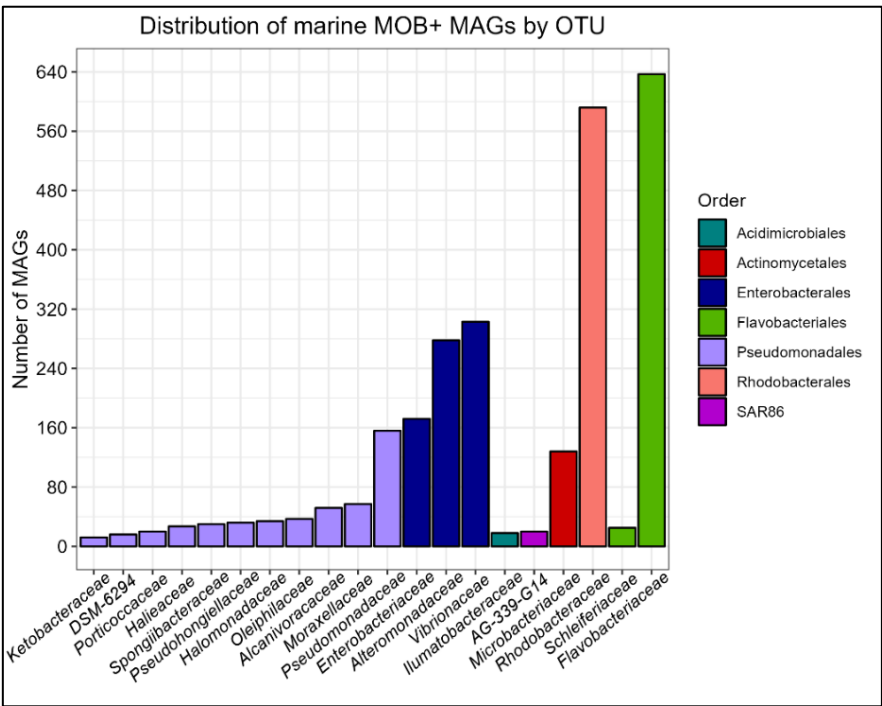
In summary, our results indicate that marine microbial ecosystems are, at present, largely free of plasmids and of the ARGs they disseminate. Thus, they will not be a significant source of antibiotic resistance threats in the coming years. On the contrary, sewage waters are enriched in RLXs (Figure IV-4), ARGs (Figure IV-8), plasmids (Figure IV-10) and multiple bacterial orders (Figure IV-6). Because PTUs circulate within bacterial orders, sewage is a key environment for investigating ARG transmission through bacterial conjugation and its potential interconnection with other aquatic, less anthropized environments. For instance, both rivers and sewage share two gammaproteobacterial orders: Burkholderiales and Pseudomonadales, with the latter also found in the ocean. This taxonomic overlap enables plasmids to potentially mediate ARG transfer across ecosystems via shared bacterial hosts. Importantly, some plasmids with broad host-range can transfer between bacterial species from different orders or even classes. For instance, PTU-P1 plasmids have high host-range and are predominantly associated with species from the order Burkholderiales¹⁶⁵, which are well represented in sewage (Figure IV-6). *Burkholderia spp.* participates in PTU exchange networks that link all relevant enterobacterial species with other Gammaproteobacteria from the order Pseudomonadales¹⁶⁵. This example suggests a potential conduit for ARG dissemination across taxonomic and environmental boundaries. For this reason, a One-Health approach should rather concentrate on studying these taxa on the aforementioned environments, including coastal waters, which are more heavily contaminated with ARGs, as potential sources of resistance threats for humans and domestic animals.

5. Conclusions

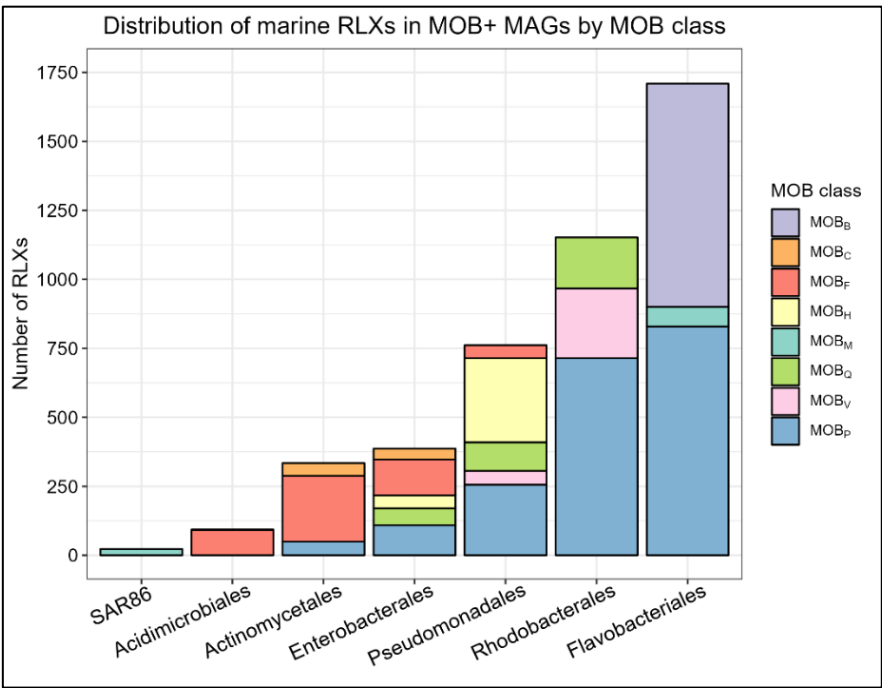
Our analysis reveals a fundamental difference in HGT dynamics between oceanic environments and more anthropogenically influenced biomes. By using the RLX gene as an effective proxy for plasmid distribution and abundance in both genomes and metagenomes, we demonstrate that marine ecosystems exhibit a notably reduced presence of conjugative plasmids and associated ARGs. This finding suggests that ARG contamination in the ocean remains in its early stages of spread via plasmid-mediated conjugation, potentially due to the relative scarcity of MGEs necessary for their dissemination. However, it is important to note that conjugative plasmids are not the sole vectors of HGT. Phages, which are highly abundant in marine ecosystems, may also contribute to ARG dissemination via transduction. Additional work is needed to evaluate the role of phage-mediated transfer and other MGEs, such as ICEs, in driving ARG mobility in the ocean.

In contrast, the abundance of RLX genes, ARGs, and plasmid sequences in human sewage represents a significant source of resistance gene contamination, posing a risk of altering environmental resistomes through punctual pollution events. Aquatic and terrestrial ecosystems may sporadically present ARG levels comparable to those observed in sewage, which underscores the importance of continuously monitoring these environments. In summary, the use of RLX genes as a proxy for quantifying plasmid abundance reinforces our understanding of differential HGT dynamics across biomes and highlights the use of targeted metagenomic approaches for advancing this understanding.

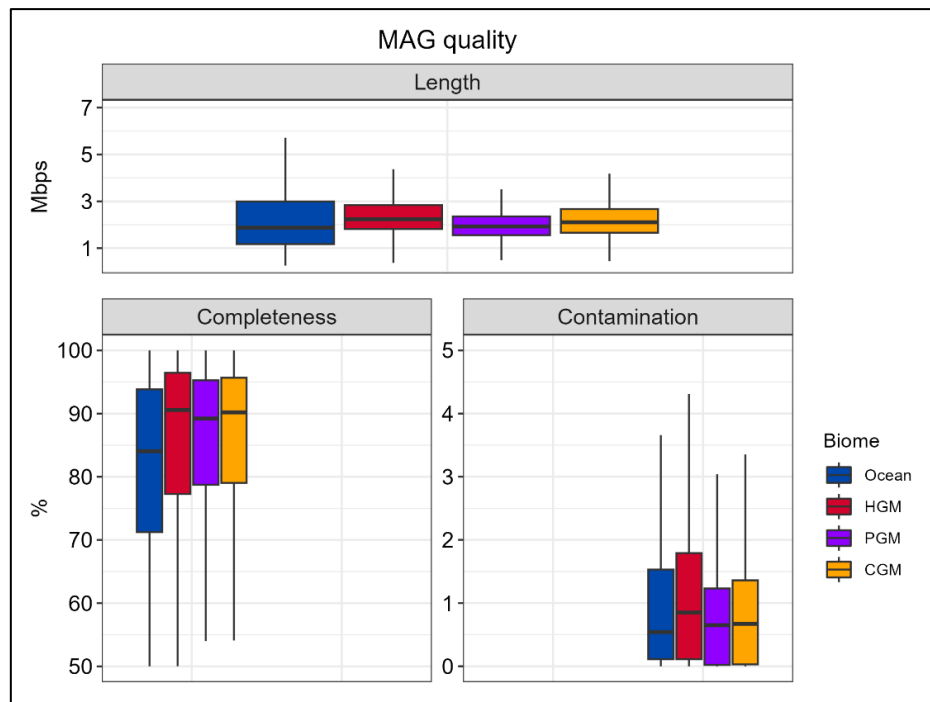
6. Supplementary material



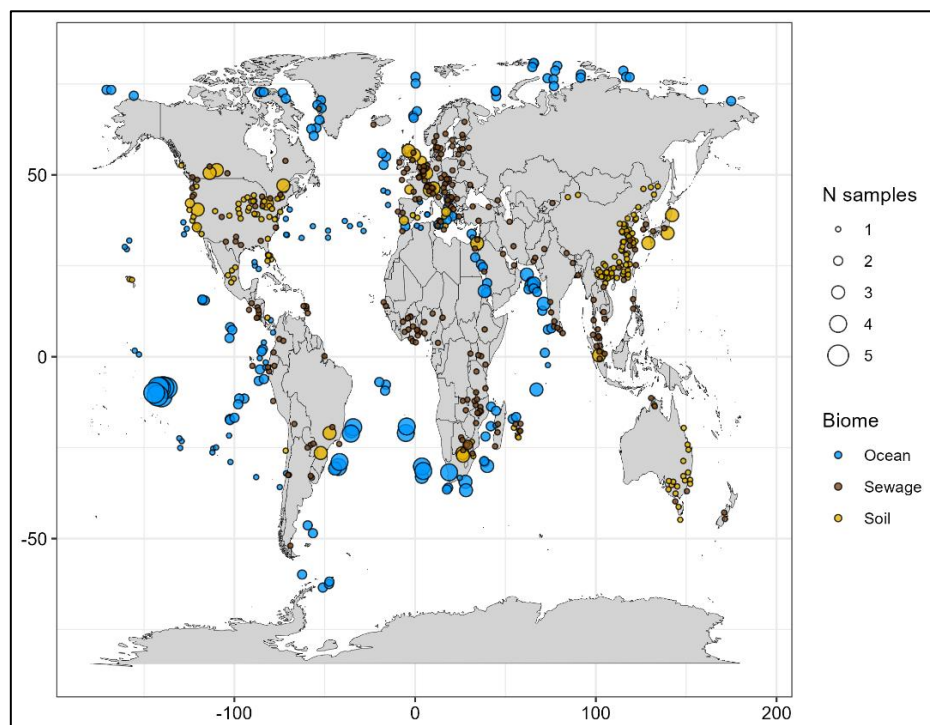
Supplementary Figure S-IV-1. Taxonomic distribution of marine, bacterial MOB+ MAGs. Bars represent the absolute abundance of MAGs from each bacterial family, colored by order. Orders encompassing <3% of the total MAGs were removed for visualization purposes.



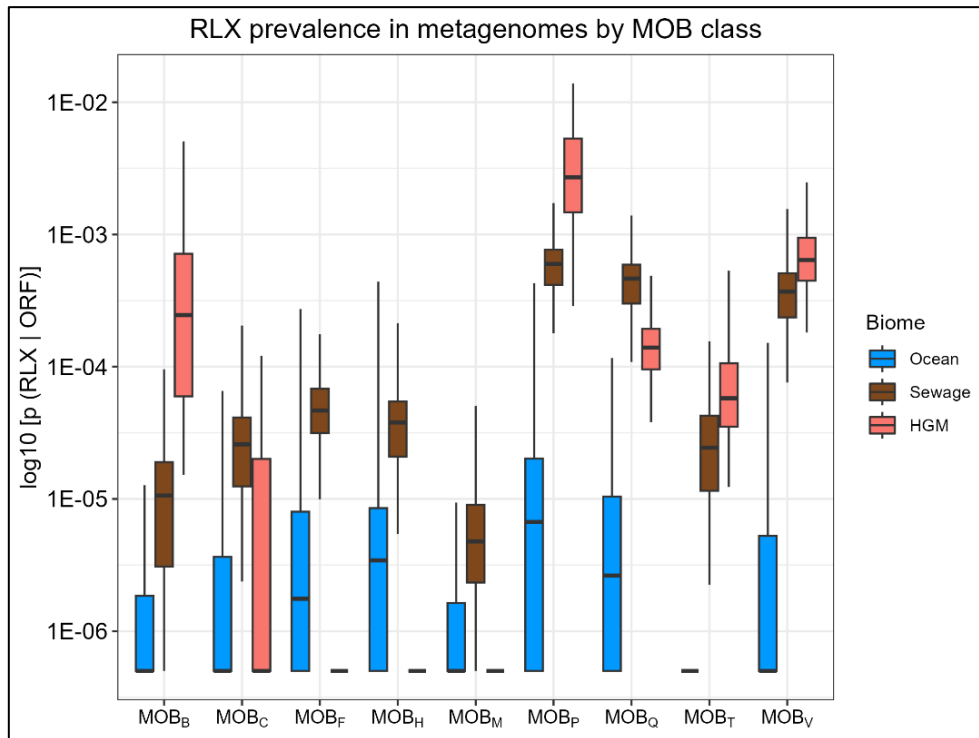
Supplementary Figure S-IV-2. Abundance of marine, MAG-encoded RLXs. Bars represent the absolute abundance of RLXs from each MOB class in the most abundant MOB+ marine bacterial orders, represented across x-axis. Minor MOB classes from the MAGs of each bacterial order were excluded for visualization purposes.



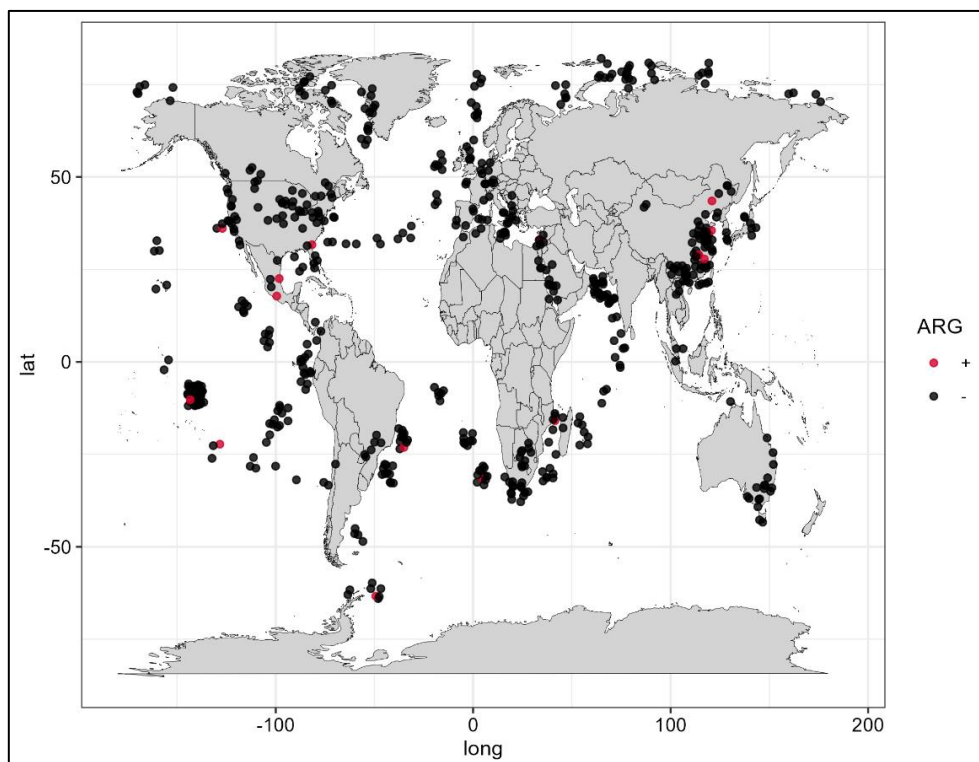
Supplementary Figure S-IV-3. Bacterial MAG quality. The measured parameters are genomic length, completeness and contamination. Horizontal black bars represent median levels, while boxes and whiskers represent the data from first to third quartiles and from the quartiles to the minimum and maximum, respectively.



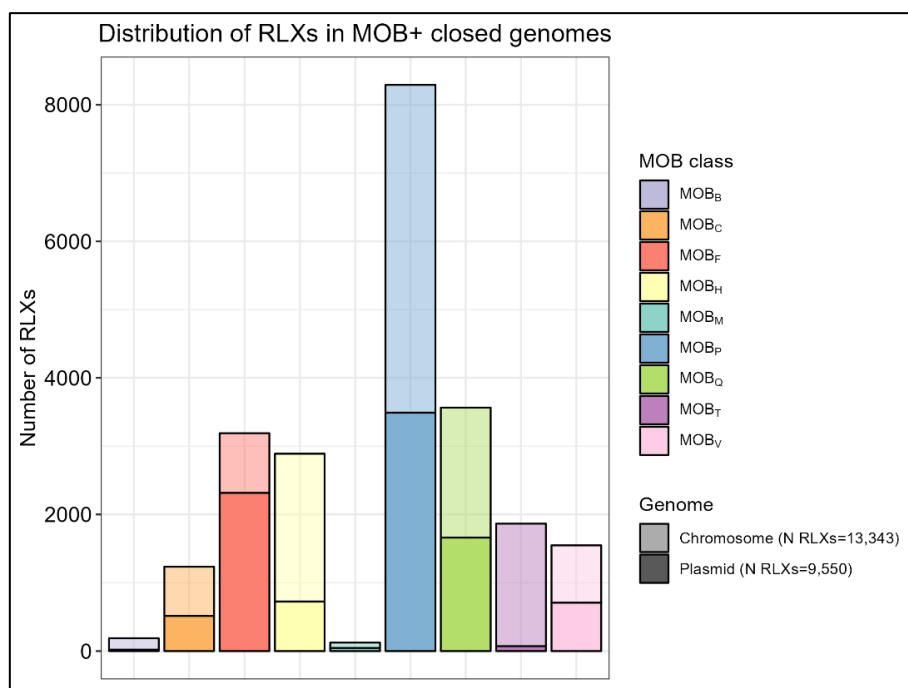
Supplementary Figure S-IV-4. Geographic distribution of metagenomic samples. Dots indicate samples from the *Tara* Oceans Expedition, the SMAG catalogue and the Global Sewage study. Color and size indicate biome and number of samples analyzed in each location, respectively.



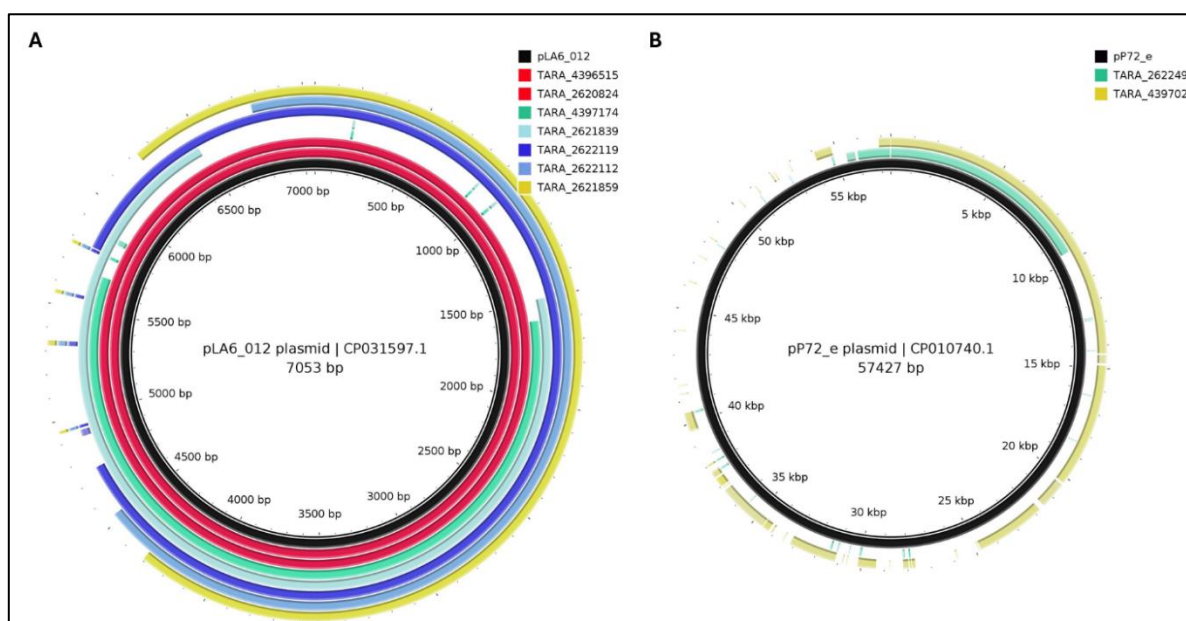
Supplementary Figure S-IV-5. RLX prevalence in metagenomes, by MOB class. Boxplots represent P_{rlx} in each biome. No RLXs from class MOB_T were predicted in the ocean contigs.



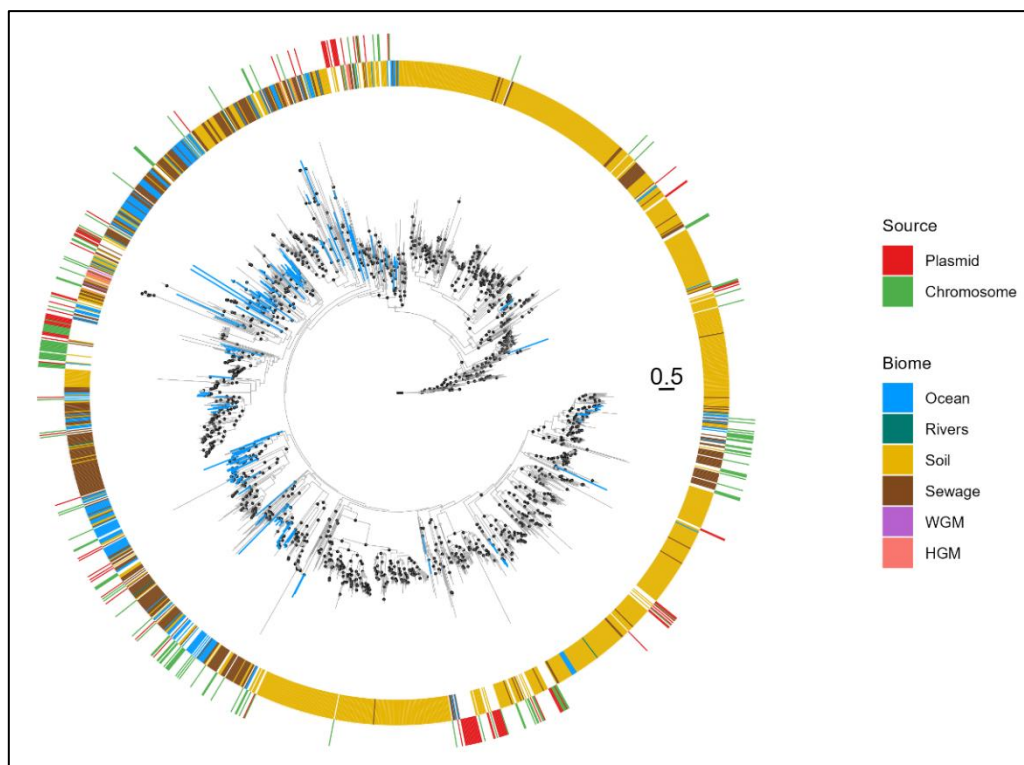
Supplementary Figure S-IV-6. Geographic distribution of the marine and soil metagenomic samples. Red dots represent ARG(+) samples, established as those with $P_{arg} > 2E-04$.



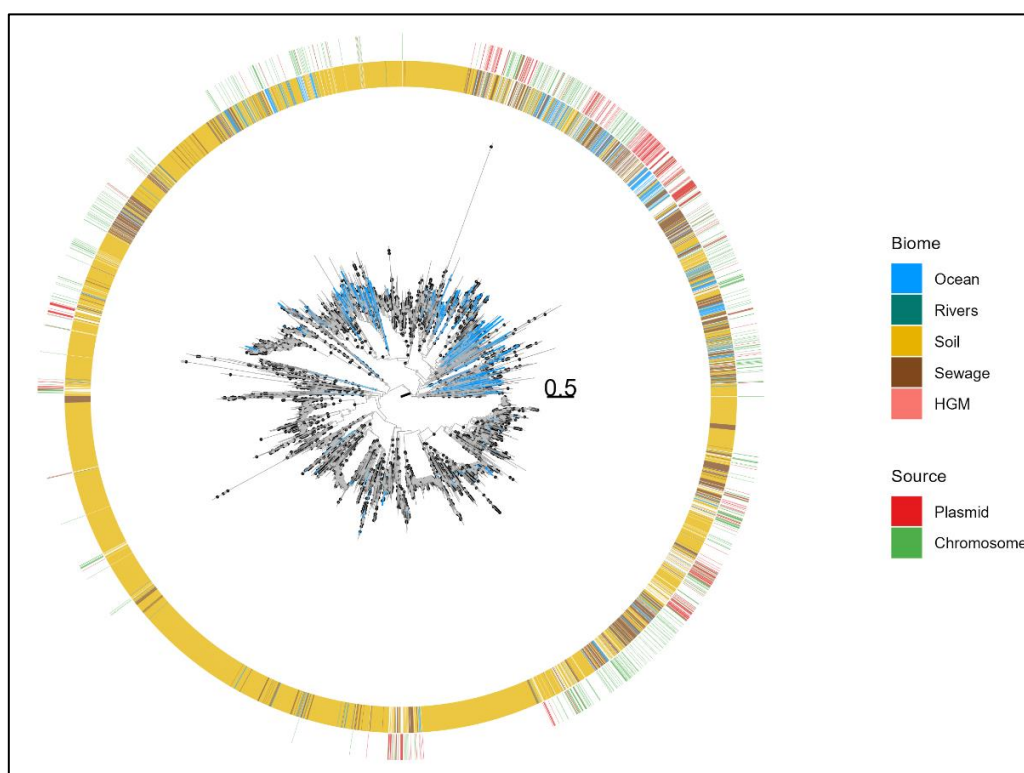
Supplementary Figure S-IV-7. Distribution of RLXs in closed genomes. Bars represent the absolute abundance of RLXs from the RefSeq200 genomic collection, split and colored by MOB class. The shaded part of the bars corresponds to the number of MOB+ MAGs, whereas the clearer part of the bar corresponds to the number of MOB- MAGs.



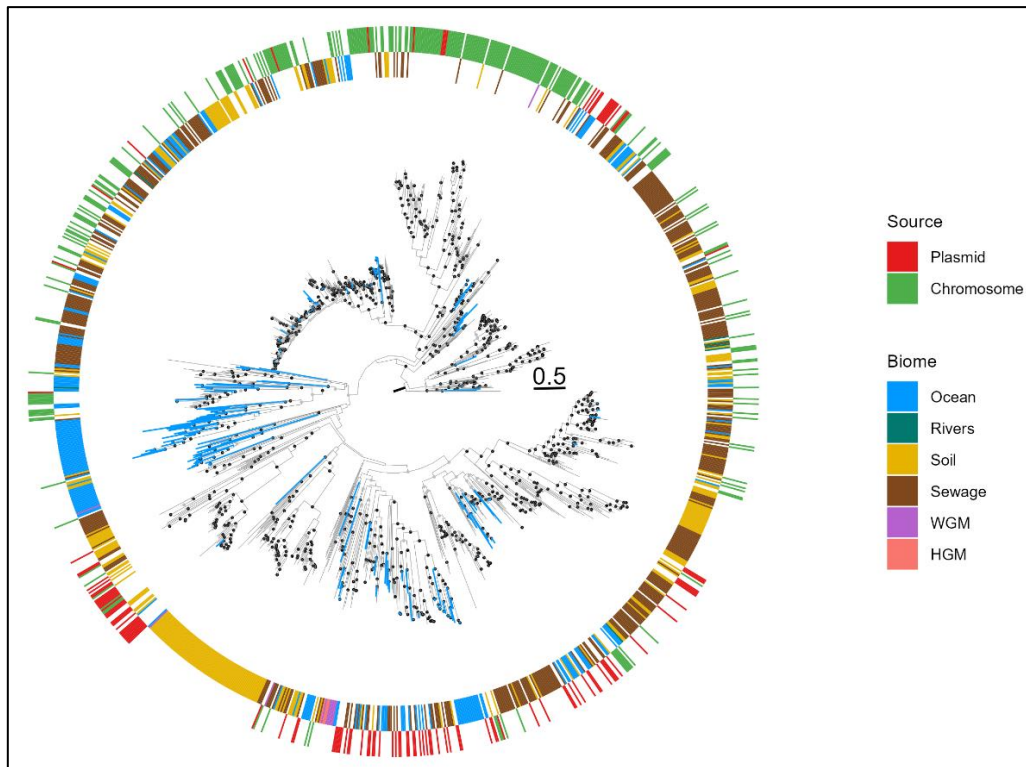
Supplementary Figure S-IV-8. Detection of MAPs (A) pLA6_012 and (B) the pP72_e. The inner black circles represent the genomic sequence of pLA6_012 and pP72_e, respectively. The colored circles represent marine metagenomic contigs from the *Tara* Oceans Expedition aligning with >99.9% identity, an E-value=0 and an alignment length >4,000 bp between plasmid and contig. Each matching contig depicted was found in a different metagenomic sample. Red circles represent perfect matches, i.e., metagenomic samples in which the complete sequence of pLA6_012 was found. The rest of the circles represent partial matches.



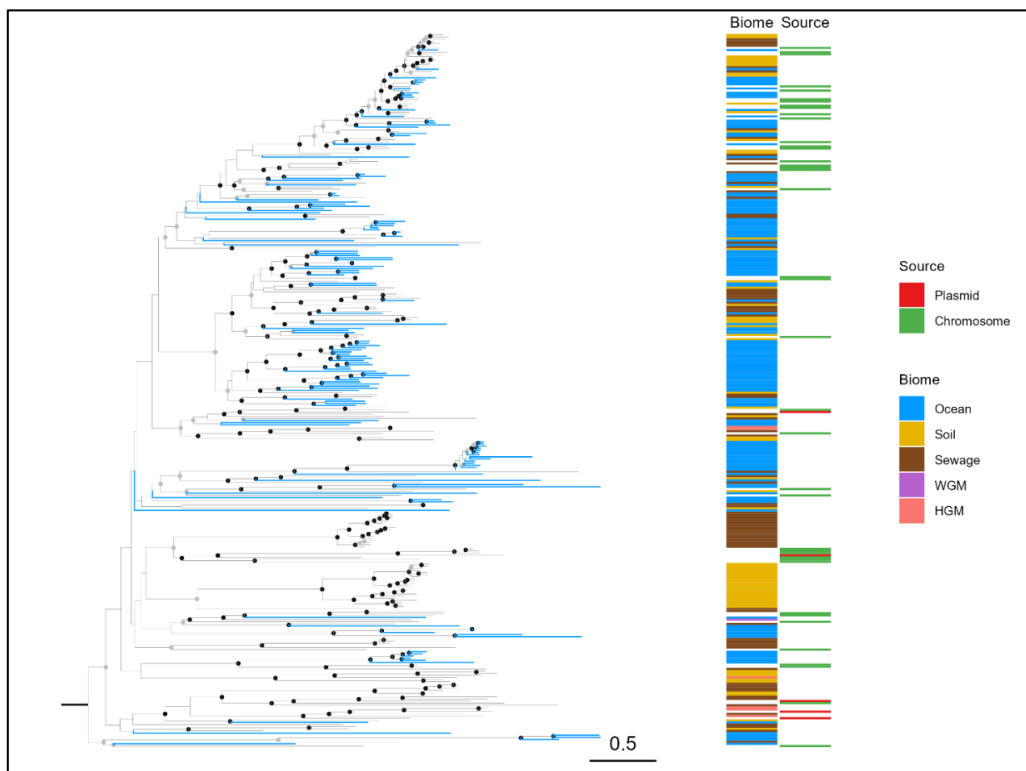
Supplementary Figure S-IV-9. Phylogenetic tree of MOB_c RLXs. The tree was built using 1,951 MOB_c RLXs with maximum likelihood with IQ-TREE¹⁷⁸ (model Q.pfam+F+R6 according to BIC and 1000 ultrafast bootstraps). The tree has been elaborated analogously as detailed in Figure IV-7.



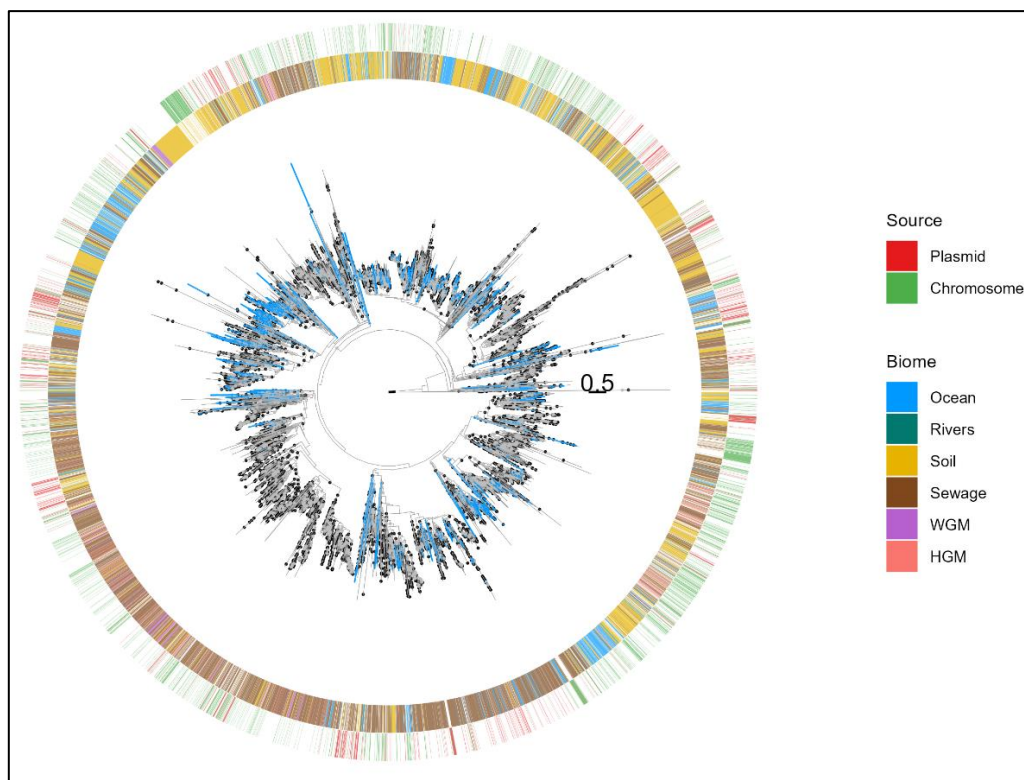
Supplementary Figure S-IV-10. Phylogenetic tree of MOB_f RLXs. The tree was built using 8,408 MOB_f RLXs with maximum likelihood with IQ-TREE¹⁷⁸ (model Q.pfam+F+R10 according to BIC and 1000 ultrafast bootstraps). The tree has been elaborated analogously as detailed in Figure IV-7.



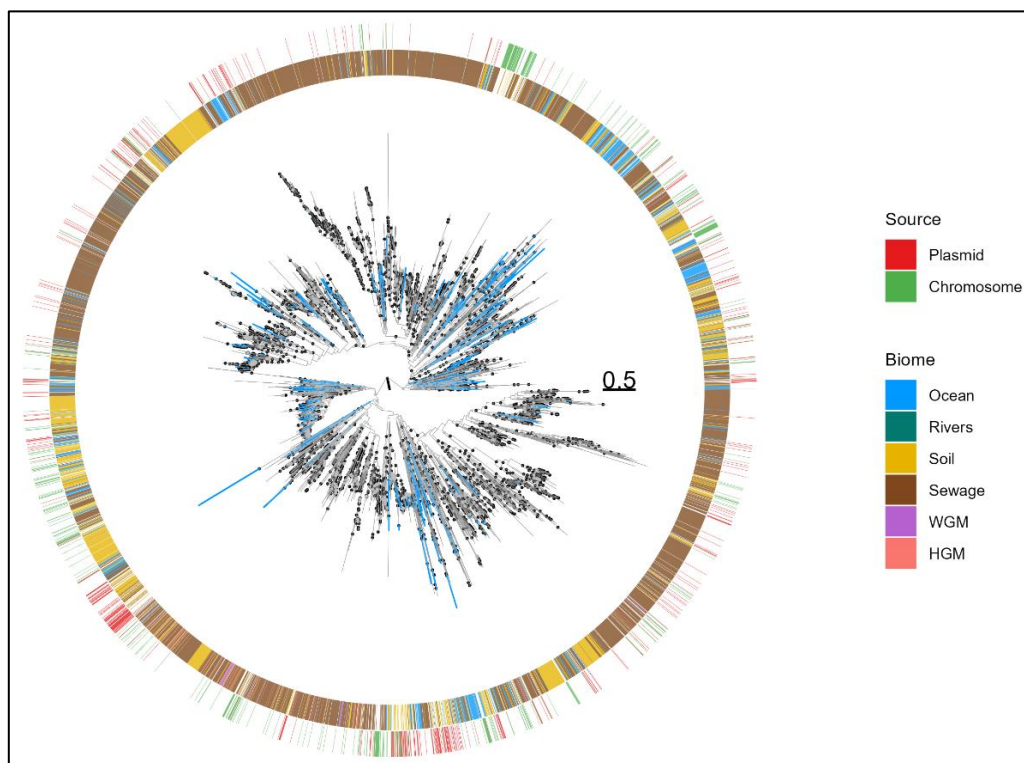
Supplementary Figure S-IV-11. Phylogenetic tree of MOB_H RLXs. The tree was built using 1,368 MOB_H RLXs with maximum likelihood with IQ-TREE¹⁷⁸ (model Q.pfam+R10 according to BIC and 1000 ultrafast bootstraps). The tree has been elaborated analogously as detailed in Figure IV-7.



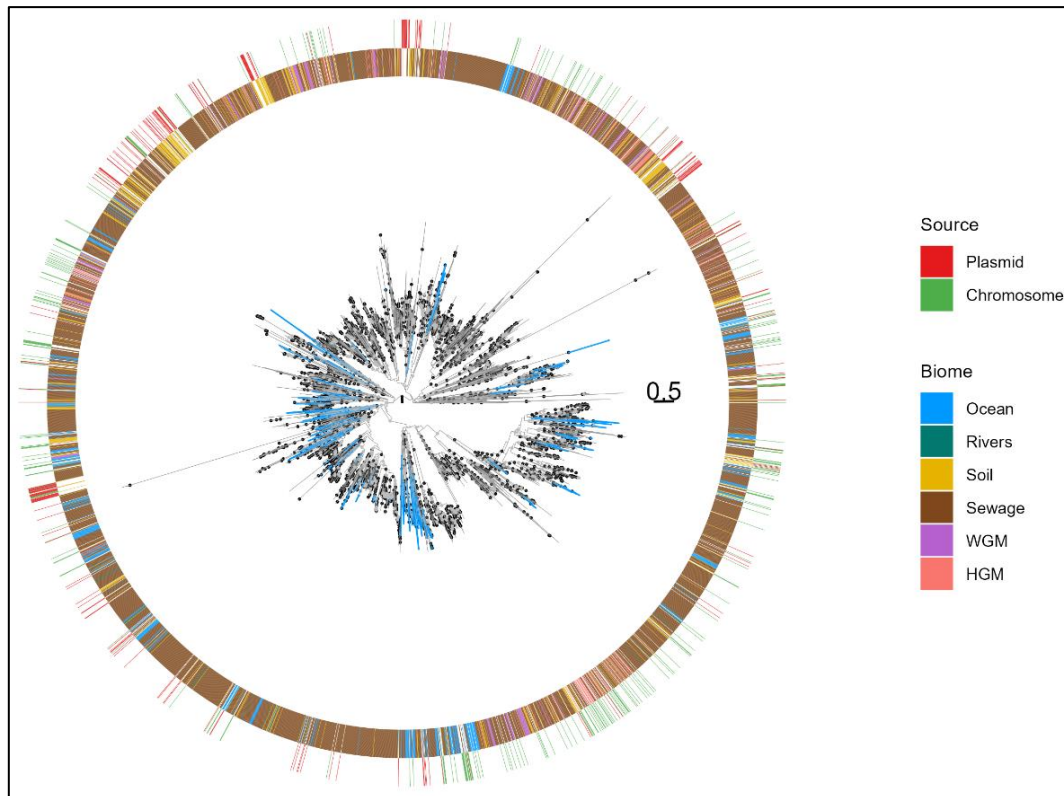
Supplementary Figure S-IV-12. Phylogenetic tree of MOB_M RLXs. The tree was built using 333 MOB_M RLXs with maximum likelihood with IQ-TREE¹⁷⁸ (model Q.pfam+F+R9 according to BIC and 1000 ultrafast bootstraps). The tree has been elaborated analogously as detailed in Figure IV-7.



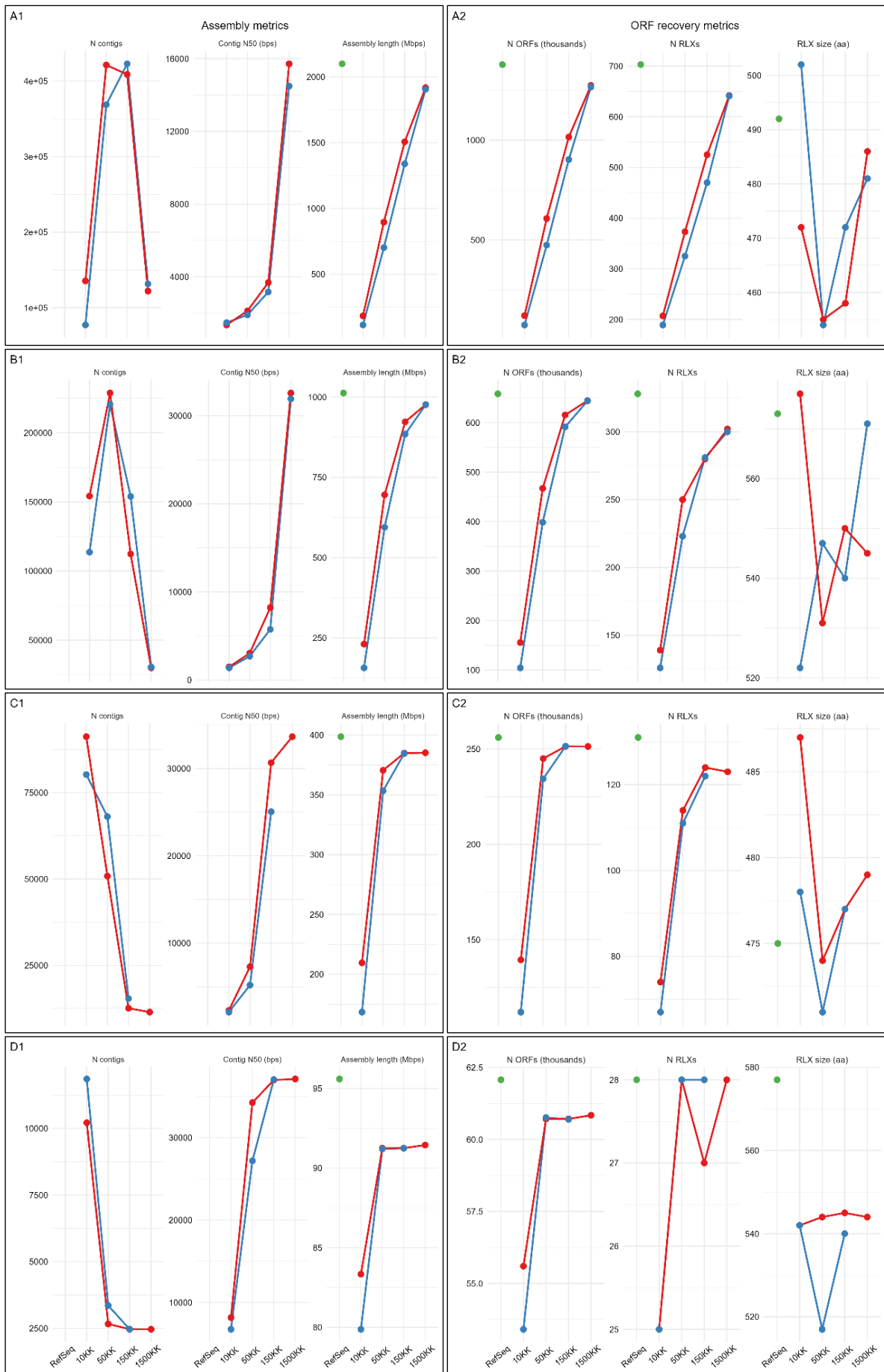
Supplementary Figure S-IV-13. Phylogenetic tree of MOB_p RLXs. The tree was built using 15,090 MOB_p RLXs with maximum likelihood with IQ-TREE¹⁷⁸ (model Q.pfam+R10 according to BIC and 1000 ultrafast bootstraps). The tree has been elaborated analogously as detailed in Figure IV-7.



Supplementary Figure S-IV-14. Phylogenetic tree of MOB_q RLXs. The tree was built using 7,189 MOB_q RLXs with maximum likelihood with IQ-TREE¹⁷⁸ (model Q.pfam+R10 according to BIC and 1000 ultrafast bootstraps). The tree has been elaborated analogously as detailed in Figure IV-7.



Supplementary Figure S-IV-15. Phylogenetic tree of MOB_v RLXs. The tree was built using 5,336 MOB_v RLXs with maximum likelihood with IQ-TREE¹⁷⁸ (model Q.pfam+F+R10 according to BIC and 1000 ultrafast bootstraps). The tree has been elaborated analogously as detailed in Figure IV-7.



Supplementary Figure S-IV-16. Contig and ORF metrics from synthetic communities. Red and blue lines represent results from metagenomic read simulations using NovaSeq and HiSeq error models, respectively. Green dots indicate the corresponding metric values calculated directly from the original synthetic community genomes, prior to read simulation. Metrics include the total number of contigs, contig length distributions, and the number of predicted ORFs recovered from each assembly. These comparisons illustrate the impact of sequencing platform and community complexity on assembly and gene prediction performance.

V. GLOBAL DISCUSSION

The technical foundation of this thesis encompasses three sequential steps: a) the characterization and isolation of one or a few gene families, b) the gene quantification in multiple metagenomes with different phenotypes and c) the correlation between gene abundance and phenotype. In this work, we present a robust methodology for accurate gene quantification in metagenomes, and illustrate its application through two comparative studies aimed at identifying differentially abundant genes in a disease-related context (Chapter I) and an environmental context (Chapter II) (Table V-1).

Feature	Chapter I	Chapter II
Gene	Metabolic gene	RLX (domain)
Phenotype	MASLD	Plasmid load

Table 0-1: Summary of gene-phenotype associations investigated.

In Chapter I, we employ an absolute abundance metric. Sequencing reads are aligned against several reference databases of metabolic genes, and the raw count of reads mapping to each gene is used as a proxy for its abundance in the metagenome (Figure III-3). To correct for sequencing-depth bias, read counts are normalized by the total library size. Gene-length bias is corrected by normalizing by the length of each target gene. To account for differences in reference database size, read counts are further normalized by the number of sequences in each database. These three corrections are combined using the RPKSM formula (Equation III-1). Although differences in average genome size between study groups are not explicitly adjusted, the consistent distribution of USCGs across groups and cohorts (Supplementary Figure S-III-4) suggests that any such bias is likely negligible.

In Chapter II, a relative abundance metric is adopted. First, sequencing reads are *de novo* assembled into contigs, from which all ORFs are predicted. RLX genes are then identified by scanning the ORF collections from each biome with HMM profiles, and their abundance is calculated as the number of predicted RLX hits divided by the total number of predicted ORFs (Figure IV-2). Sequencing-depth bias is implicitly addressed, since fewer reads yield fewer ORFs and, by extension, proportionally fewer RLX hits. Because the same protein domain is compared across samples, additional normalization by gene length is not necessary¹³⁶. While length correction is typically applied in gene-level comparisons, the rationale extends to domain-based analysis: when the same query, whether gene or domain, is used consistently

across samples, any potential length bias becomes negligible. As in Chapter I, average genome size is not explicitly accounted for. However, the distributions of USCGs (Figure IV-4A) and total ORFs (Figure IV-5A) across different biomes are nearly identical, indicating that this factor exerts minimal influence on this analysis.

Functional profiling of metagenomes can be conducted via two primary strategies: read-based methods and assembly-based approaches. Read-based profiling aligns sequencing reads directly to reference databases^{35,36}, bypassing the computationally intensive bottleneck of *de novo* assembly and achieving faster analyses. However, short reads may not be discriminative enough to distinguish between closely related gene families or to avoid ambiguous alignments to promiscuous protein domains. Consequently, functional profiles may be over- or under-assigned depending on how gene families are defined and the degree of sequence conservation among them³⁷. The key determinant of read-based accuracy is the completeness of reference databases, understood as the number of clades represented. Because most public catalogs are biased towards human-associated microbiomes (Figure I-1), read-based methods perform correctly on small, well-characterized datasets such as human GM samples where reference coverage is high⁹⁹, but become less reliable in complex, poorly characterized environments.

In contrast, assembly-based profiling reconstructs longer genomic sequences (contigs) before annotation (Figure I-4), capturing full genes and preserving synteny. Assembled contigs mitigate short-read multi-mapping problems and enable differentiation between functionally divergent paralogs that share high sequence similarity¹⁴⁰. Even when individual genes lack close database matches, their genomic context can improve functional assignments³⁷. The effectiveness of this approach highly depends on assembly completeness, understood as the percentage of reads that are successfully assembled. Achieving high completeness in complex communities requires substantial sequencing depth and incurs significant computational cost. Additionally, assemblies often demand manual curation to correct misassemblies or properly bin contigs⁸³. Moreover, annotation bias can arise because abundant taxa tend to produce more complete assemblies, leading to overrepresentation of genes from dominant community members, whereas contigs from low-abundance organisms may be fragmented or entirely missing⁹⁹. As a result, assembly-based methods offer more precise functional

resolution, particularly in novel or non-human microbiomes, but at the expense of increased resource requirements³⁷.

As a result, metagenomics enables accurate gene quantification only within a defined range of community complexity and sequencing depth, as demonstrated by our results using synthetic communities (Figure IV-9). Beyond that range, *de novo* assembly becomes the primary challenge. By collapsing reads into *k-mers*, *de Bruijn* graphs inevitably lose information: short *k-mers* lack the full context of the original reads and can merge disparate genomic regions (e.g. repeats) into a single path³³⁶. Increasing *k-mer* size preserves more sequence context and improves specificity, but larger *k-mers* further fragment the graph when gaps or sequencing errors occur, leading to misassemblies⁷¹. Despite their higher base-calling error rates³⁴, long-read technologies such as Oxford Nanopore or PacBio can often span entire repeats and bypass these assembly bottlenecks, but generating large-scale long-read datasets remains prohibitively expensive.

These considerations guided the methodological choices of this thesis, which focuses on short read, Illumina-sequenced metagenomes. We evaluated the reliability of gene quantification under controlled conditions by analyzing 554 and 974 metagenomic samples in Chapter I and II, respectively. In Chapter I, functional profiling was performed using read-based assignment methods for small human GM metagenomes, leveraging the high completeness of human metabolic gene reference databases derived from UHGG to obtain rapid and sufficiently accurate functional profiles. In Chapter II, an assembly-based approach was adopted for diverse environmental microbiomes (e.g., aquatic and terrestrial ecosystems), where high novelty and complexity required contig reconstruction despite its higher computational demands. To quantify gene abundance in assembled contigs, two equivalent approaches were tested: mapping reads back to contigs and counting alignments^{337,338}, and using assembler-provided coverage metrics⁴⁸. In our unpublished benchmarks, both methods yielded similar results.

Target gene comparative studies often require annotation protocols that differ substantially from those used in whole-metagenome workflows. A key factor influencing annotation accuracy is the phylogenetic diversity and quality of the reference database used for alignment¹⁰⁰. This limitation is especially relevant when targeting specific gene families, as poorly represented or misannotated sequences in public databases can lead to inaccurate or

incomplete functional assignments. To address this issue, custom and manually curated gene databases tailored to the specific targets of each analysis were developed in this thesis: metabolic gene families in Chapter I (this work) and RLX MOB families³¹⁴ in Chapter II. These curated resources improve annotation accuracy by ensuring that the sequence space is both phylogenetically diverse and functionally relevant to the research questions posed.

Focusing on a single (Chapter II) or a few (Chapter I) gene families instead of attempting to annotate the entire functional repertoire of a metagenome offers several advantages. First, it removes the need to normalize gene abundance by the total abundance of all predicted proteins in a sample, avoiding biases associated with their measure, as done previously¹⁴¹. Second, it mitigates the ambiguity from assigning short reads to multiple protein families with high sequence similarity, a common source of error in large-scale functional annotation⁷⁶. In such targeted approaches, gene classification becomes a binary task: either a gene is detected through an alignment or not. This minimizes multi-mapping artifacts, simplifies interpretation and improves the overall robustness of the results.

This strategy is particularly relevant for gene quantification because many gene families or orthologous groups in reference databases are defined based on sequence homology to only one or a few experimentally validated genes. Consequently, most database entries lack direct functional characterization. For example, in the carbohydrate-active enzymes database³³⁹, only a single member per family is required to be biochemically characterized, with the remaining members grouped solely by sequence similarity. This can lead to functional redundancy across between families and hidden functional diversity within a single family, complicating downstream interpretation⁹⁹. We acknowledge that this limitation affects the analyses in Chapter I, although the potential error is minimized when comparisons are limited to a few protein families.

Importantly, although filtering duplicate metagenomic reads is a common preprocessing step to minimize PCR amplification biases^{54,340}, we intentionally opted not to apply this filter. Biological duplicates can arise from highly abundant organisms, particularly in deeply sequenced libraries, and their removal may artificially reduce gene abundance estimates¹³³. Therefore, retaining potentially duplicate reads was preferred, as it allowed a more accurate representation of gene prevalence, particularly in high-abundance taxa.

For the assembly of metagenomic reads in Chapter II, we selected MEGAHIT⁴⁸ as the assembler of choice. MEGAHIT has been shown to recover a higher number of genes that can be functionally annotated from complex environmental samples, such as soil and ocean³⁴¹, and it offers greater computational efficiency⁴¹ over alternatives like metaSPAdes⁴⁹. While metaSPAdes generally produces longer contigs, these have been reported to be less accurate in highly complex metagenomes³⁴¹. This, combined with the large number of environmental samples analyzed (Table IV-1), further supports our choice of assembler.

Applying logarithmic transformation to metagenomic count data generates a large number of zeros, which poses two challenges. First, zero counts cannot be log-transformed. Second, zeros are inherently ambiguous, as they may indicate either true absence of a gene or a presence below the detection threshold due to limited sequencing depth. The first problem can be addressed by adding pseudocounts (typically one) to all observations in the dataset, allowing the transformation to proceed. However, pseudocounts can distort effect sizes -and thus statistical significances-, particularly when gene counts are low¹³⁶. Due to this limitation, we avoided log-transformation methods in Chapter II whenever possible (Figures IV-4 and IV-8A).

Finally, throughout this thesis, the Benjamini-Hochberg false discovery rate correction was applied to control for multiple hypothesis testing, a standard approach in large-scale metagenomic analyses to reduce the number of false-positive results¹³⁶. This ensures that statistical findings reported across both comparative studies maintain robust control of type I error rates.

VI. GLOBAL CONCLUSIONS

1. Gene-level metagenomic profiling reveals functional shifts in metabolic pathways and plasmid-marker genes that are overlooked by 16S rRNA-based taxonomic approaches, yet are essential for understanding microbial contributions to host physiology and ecosystem dynamics.
2. MASLD is marked by a coordinated depletion of butyrate- and methane-producing genes and enrichment of SCA- and TMA-producing genes, reflecting a microbial shift towards metabolic pathways that may contribute to pathogenesis.
3. Accessory metabolic genes, often carried on MGEs, vary across strains and environments. Integrating their analysis with plasmid mobilization data identifies horizontally transferable functions linked to disease and ecological adaptation.
4. The ocean is depleted in RLXs, ARGs and plasmid marker genes, indicating low conjugative HGT activity. In contrast, sewage harbors RLX- and ARG-rich plasmids, forming hotspots whose episodic discharges can raise coastal ARG load to terrestrial levels.
5. Combining metabolic pathway profiling, accessory genome analysis and MGE tracking provides a unified framework for identifying clinically relevant biomarkers in MASLD and detecting ARG pollution for environmental monitoring.
6. Gene-level surveillance of biomarkers, from butyrate/TMA shifts in the GM to RLX abundance in sewage, offers actionable tools for patient stratification, microbiome-targeted therapies, and early detection systems in microbial ecosystem management.

BIBLIOGRAPHY

1. Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* **5**, 245–249 (1998).
2. Marchesi, J. R. & Ravel, J. The vocabulary of microbiome research: a proposal. *Microbiome* **3**, 31 (2015).
3. Berg, G., Rybakova, D. & Fischer, D. Microbiome definition re-visited: old concepts and new challenges. *Microbiome* **8**, 103 (2020).
4. Pande, S. & Kost, C. Bacterial unculturability and the formation of intercellular metabolic networks. *Trends Microbiol.* **25**, 349–361 (2017).
5. Lewis, W. H., Tahon, G. & Geesink, P. Innovations to culturing the uncultured microbial majority. *Nat. Rev. Microbiol.* **19**, 225–240 (2020).
6. Hoehler, T. M. & Jørgensen, B. B. Microbial life under extreme energy limitation. *Nat. Rev. Microbiol.* **11**, 83–94 (2013).
7. Nagler, M., Insam, H., Pietramellara, G. & Ascher-Jenull, J. Extracellular DNA in natural environments: features, relevance and applications. *Appl. Microbiol. Biotechnol.* **102**, 6343–6356 (2018).
8. Bodor, A., Bounedjoum, N. & Vincze, G. E. Challenges of unculturable bacteria: environmental perspectives. *Rev. Environ. Sci. Biotechnol.* **19**, 1–22 (2020).
9. Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J. & Crosby, L. Phylogenetically novel uncultured microbial cells cominate earth microbiomes. *mSystems* **3**, 18 (2018).
10. Wu, D., Seshadri, R., Kyrpides, N. C. & Ivanova, N. N. A metagenomic perspective on the microbial prokaryotic genome census. *Sci. Adv.* **11**, eadq2166 (2025).
11. Cleary, B., Brito, I. L. & Huang, K. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nat. Biotechnol.* **33**, 1053–1060 (2015).
12. Vila, A. V. *et al.* Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Sci Transl Med* **10**, 1–11 (2018).
13. Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* **4**, 293–305 (2019).

14. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
15. Starr, E. P. *et al.* Stable-isotope-informed, genome-resolved metagenomics uncovers potential cross-kingdom interactions in rhizosphere soil. *mSphere* **6**, 1–18 (2021).
16. Hugerth, L. W. & Andersson, A. F. Analysing microbial community composition through amplicon sequencing: from sampling to hypothesis testing. *Front. Microbiol.* **8**, 1–22 (2017).
17. Dermastia, T. T., Vascotto, I., Francé, J., Stanković, D. & Mozetič, P. Evaluation of the *rbcL* marker for metabarcoding of marine diatoms and inference of population structure of selected genera. *Front. Microbiol.* **14**, 1–21 (2023).
18. Heller, P., Casaletto, J., Ruiz, G. & Geller, J. A database of metazoan cytochrome c oxidase subunit I gene sequences derived from GenBank with CO-ARBitrator. *Sci. Data* **5**, 180156 (2018).
19. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2012).
20. McDonald, D. *et al.* Greengenes2 unifies microbial data in a single reference tree. *Nat. Biotechnol.* **42**, 715–718 (2024).
21. Haft, D. H. *et al.* RefSeq and the prokaryotic genome annotation pipeline in the age of metagenomes. *Nucleic Acids Res.* **52**, D762–D769 (2024).
22. Abellan-Schneyder, I. *et al.* Primer, pipelines, parameters: issues in 16S rRNA gene sequencing. *mSphere* **6**, e01202-20 (2021).
23. O’Callaghan, J. L., Willner, D., Buttini, M., Huygens, F. & Pelzer, E. S. Limitations of 16S rRNA gene sequencing to characterize *Lactobacillus* species in the upper genital tract. *Front. Cell Dev. Biol.* **9**, 641921 (2021).
24. Větrovský, T. & Baldrian, P. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS ONE* **8**, e57923 (2013).
25. Thijs, S. *et al.* Comparative evaluation of four bacteria-specific primer pairs for 16S rRNA gene surveys. *Front. Microbiol.* **8**, 1–15 (2017).
26. Johnson, J. S. *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **10**, 5029 (2019).

- 27.** Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
- 28.** Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
- 29.** Odom, A. R., Faits, T., Castro-Nallar, E., Crandall, K. A. & Johnson, W. E. Metagenomic profiling pipelines improve taxonomic classification for 16S amplicon sequencing data. *Sci. Rep.* **13**, 13957 (2023).
- 30.** Douglas, G. M. *et al.* PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* **38**, 685–688 (2020).
- 31.** Sun, S., Jones, R. B. & Fodor, A. A. Inference-based accuracy of metagenome prediction tools varies across sample types and functional categories. *Microbiome* **8**, 46 (2020).
- 32.** Machado, M. S. *et al.* On the limits of 16S rRNA gene-based metagenome prediction and functional profiling. *Microb. Genomics* **10**, 1–14 (2024).
- 33.** Ortiz-Estrada, Á. M., Gollas-Galván, T., Martínez-Córdova, L. R. & Martínez-Porchas, M. Predictive functional profiles using metagenomic 16S rRNA data: a novel approach to understanding the microbial ecology of aquaculture systems. *Rev. Aquac.* **11**, 234–245 (2019).
- 34.** Rang, F. J., Kloosterman, W. P. & De Ridder, J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* **19**, 90 (2018).
- 35.** Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
- 36.** Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).
- 37.** Tamames, J., Cobo-Simón, M. & Puente-Sánchez, F. Assessing the performance of different approaches for functional and taxonomic annotation of metagenomes. *BMC Genomics* **20**, 960 (2019).
- 38.** Kolmogorov, M. *et al.* metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* **17**, 1103–1110 (2020).

- 39.** Sánchez-Navarro, R. *et al.* Long-read metagenome-assembled genomes improve identification of novel complete biosynthetic gene clusters in a complex microbial activated sludge ecosystem. *mSystems* **7**, e00632-22 (2022).
- 40.** Kang, X., Xu, J., Luo, X. & Schönhuth, A. Hybrid-hybrid correction of errors in long reads with HERO. *Genome Biol.* **24**, 275 (2023).
- 41.** Meyer, F. *et al.* Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nat. Methods* **19**, 429–440 (2022).
- 42.** El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).
- 43.** Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Res.* **53**, D672–D677 (2025).
- 44.** Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res.* **48**, D445–D453 (2020).
- 45.** Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
- 46.** Yang, C. *et al.* A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput. Struct. Biotechnol. J.* **19**, 6301–6314 (2021).
- 47.** Luan, T. *et al.* MetaCompass: reference-guided assembly of metagenomes. Preprint at <https://doi.org/10.1101/212506> (2024).
- 48.** Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. *Bioinformatics* **31**, 1674–1676 (2015).
- 49.** Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
- 50.** Hyatt, D., LoCascio, P. F., Hauser, L. J. & Uberbacher, E. C. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28**, 2223–2230 (2012).
- 51.** Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

52. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
53. Yue, Y. *et al.* Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. *BMC Bioinformatics* **21**, 334 (2020).
54. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
55. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
56. Wang, Z., Huang, P., You, R., Sun, F. & Zhu, S. MetaBinner: a high-performance and stand-alone ensemble binning method to recover individual genomes from complex microbial communities. *Genome Biol.* **24**, 1 (2023).
57. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
58. The Genome Standards Consortium *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
59. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
60. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
61. Arikawa, K. *et al.* Recovery of strain-resolved genomes from human microbiome through an integration framework of single-cell genomics and metagenomics. *Microbiome* **9**, 202 (2021).
62. Asnicar, F. *et al.* Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat. Commun.* **11**, 2500 (2020).
63. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).

- 64.** Alneberg, J. *et al.* Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. *Microbiome* **6**, 173 (2018).
- 65.** Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).
- 66.** Breitwieser, F. P., Lu, J. & Salzberg, S. L. A review of methods and databases for metagenomic classification and assembly. *Brief. Bioinform.* **20**, 1125–1136 (2019).
- 67.** Menzel, P., Ng, K. L. & Krogh, A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* **7**, 11257 (2016).
- 68.** Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).
- 69.** Olson, N. D. *et al.* Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes.
- 70.** Georganas, E. *et al.* Extreme scale *de novo* metagenome assembly. *ACM* **1**, (2018).
- 71.** Vollmers, J., Wiegand, S. & Kaster, A.-K. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - not only size matters! *PLOS ONE* **12**, e0169662 (2017).
- 72.** Anthony, W. E. *et al.* From soil to sequence: filling the critical gap in genome-resolved metagenomics is essential to the future of soil microbial ecology. *Environ. Microbiome* **19**, 56 (2024).
- 73.** Mise, K. & Iwasaki, W. Unexpected absence of ribosomal protein genes from metagenome-assembled genomes. *ISME Commun.* **2**, 118 (2022).
- 74.** Gemayel, K., Lomsadze, A. & Borodovsky, M. MetaGeneMark-2: improved gene prediction in metagenomes. Preprint at <https://doi.org/10.1101/2022.07.25.500264> (2022).
- 75.** Estrada, K., Garcarrubio, A. & Merino, E. Unraveling the plasticity of translation initiation in prokaryotes: beyond the invariant Shine-Dalgarno sequence. *PLOS ONE* **19**, e0289914 (2024).
- 76.** Tamames, J. & Puente-Sánchez, F. SqueezeMeta, a highly portable, fully automatic metagenomic analysis pipeline. *Front. Microbiol.* **9**, 3349 (2019).

77. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
78. The UniProt Consortium *et al.* UniProt: the Universal Protein knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617 (2025).
79. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
80. Hernández-Plaza, A. *et al.* eggNOG 6.0: enabling comparative genomics across 12 535 organisms. *Nucleic Acids Res.* **51**, D389–D394 (2023).
81. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
82. Puente-Sánchez, F., García-García, N. & Tamames, J. SQMtools: automated processing and visual analysis of 'omics data with R and anvi'o. *BMC Bioinformatics* **21**, 358 (2020).
83. Eren, A. M. *et al.* Community-led, integrated, reproducible multi-omics with anvi'o. *Nat. Microbiol.* **6**, 3–6 (2020).
84. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **39**, 105–114 (2021).
85. Chibani, C. M. *et al.* A catalogue of 1,167 genomes from the human gut archaeome. *Nat. Microbiol.* **7**, 48–61 (2022).
86. Lesker, T. R. *et al.* An integrated metagenome catalog reveals new insights into the murine gut microbiome. *Cell Rep.* **30**, 2909–2922.e6 (2020).
87. Xie, F. *et al.* An integrated gene catalog and over 10,000 metagenome-assembled genomes from the gastrointestinal microbiome of ruminants. *Microbiome* **9**, 137 (2021).
88. Chen, C. *et al.* Expanded catalog of microbial genes and metagenome-assembled genomes from the pig gut microbiome. *Nat. Commun.* **12**, 1106 (2021).
89. Feng, Y. *et al.* Metagenome-assembled genomes and gene catalog from the chicken gut microbiome aid in deciphering antibiotic resistomes. *Commun. Biol.* **4**, 1305 (2021).
90. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).

- 91.** Paoli, L. *et al.* Biosynthetic potential of the global ocean microbiome. *Nature* **607**, 111–118 (2022).
- 92.** Han, Y. *et al.* A comprehensive genomic catalog from global cold seeps. *Sci. Data* **10**, 596 (2023).
- 93.** Ma, B. *et al.* A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. *Nat. Commun.* **14**, 7318 (2023).
- 94.** Nayfach, S. *et al.* A genomic catalog of Earth’s microbiomes. *Nat. Biotechnol.* **39**, 499–509 (2021).
- 95.** Poyet, M. *et al.* A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research. *Nat. Med.* **25**, 1442–1452 (2019).
- 96.** Olsson, L. M. *et al.* Dynamics of the normal gut microbiota: A longitudinal one-year population study in Sweden. *Cell Host Microbe* **30**, 726–739.e3 (2022).
- 97.** Lloyd-Price, J. *et al.* Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* **569**, 655–662 (2019).
- 98.** Rosenboom, I. *et al.* Longitudinal development of the airway metagenome of preterm very low birth weight infants during the first two years of life. *ISME Commun.* **3**, 75 (2023).
- 99.** Treiber, M. L., Taft, D. H., Korf, I., Mills, D. A. & Lemay, D. G. Pre- and post-sequencing recommendations for functional annotation of human fecal metagenomes. *BMC Bioinformatics* **21**, 74 (2020).
- 100.** Carr, R. & Borenstein, E. Comparative analysis of functional metagenomic annotation and the mappability of short reads. *PLoS ONE* **9**, e105776 (2014).
- 101.** Maghini, D. G. *et al.* Quantifying bias introduced by sample collection in relative and absolute microbiome measurements. *Nat. Biotechnol.* **42**, 328–338 (2024).
- 102.** McCarthy, A., Chiang, E., Schmidt, M. L. & Denef, V. J. RNA preservation agents and nucleic acid extraction method bias perceived bacterial community composition. *PLOS ONE* **10**, e0121659 (2015).
- 103.** Choo, J. M., Leong, L. E. & Rogers, G. B. Sample storage conditions significantly influence faecal microbiome profiles. *Sci. Rep.* **5**, 16350 (2015).

- 104.** Jones, M. B. *et al.* Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proc. Natl. Acad. Sci.* **112**, 14024–14029 (2015).
- 105.** Poulsen, C. S., Ekstrøm, C. T., Aarestrup, F. M. & Pamp, S. J. Library preparation and sequencing platform introduce bias in metagenomic-based characterizations of microbiomes. *Microbiol. Spectr.* **10**, e00090-22 (2022).
- 106.** Browne, P. D. *et al.* GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. *GigaScience* **9**, giaa008 (2020).
- 107.** Duvallet, C., Gibbons, S. M., Gurry, T., Irizarry, R. A. & Alm, E. J. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat. Commun.* **8**, 1784 (2017).
- 108.** Ospino, M. C. *et al.* Evaluation of multiple displacement amplification for metagenomic analysis of low biomass samples. *ISME Commun.* **4**, ycae024 (2024).
- 109.** Parras-Moltó, M., Rodríguez-Galet, A., Suárez-Rodríguez, P. & López-Bueno, A. Evaluation of bias induced by viral enrichment and random amplification protocols in metagenomic surveys of saliva DNA viruses. *Microbiome* **6**, 119 (2018).
- 110.** Salipante, S. J. *et al.* Rapid 16S rRNA next generation sequencing of polymicrobial clinical samples for diagnosis of complex bacterial infections. *PLoS ONE* **8**, e65226 (2013).
- 111.** Brinkmann, A. *et al.* Development and preliminary evaluation of a multiplexed amplification and next generation sequencing method for viral hemorrhagic fever diagnostics. *PLoS Negl. Trop. Dis.* **11**, e0006075 (2017).
- 112.** Morsli, M. *et al.* Real-time metagenomics-based diagnosis of community-acquired meningitis: a prospective series, southern France. *eBioMedicine* **84**, 104247 (2022).
- 113.** Child, H. T. *et al.* Comparison of metagenomic and targeted methods for sequencing human pathogenic viruses from wastewater. *mBio* **14**, e01468-23 (2023).
- 114.** Thézé, J. *et al.* Genomic epidemiology reconstructs the introduction and spread of Zika virus in Central America and Mexico. *Cell Host Microbe* **23**, 855-864.e7 (2018).
- 115.** Kent, B. N. *et al.* Complete bacteriophage transfer in a bacterial endosymbiont (*Wolbachia*) determined by targeted genome capture. *Genome Biol. Evol.* **3**, 209–218 (2011).

- 116.** Noyes, N. R. *et al.* Enrichment allows identification of diverse, rare elements in metagenomic resistome-virulome sequencing. *Microbiome* **5**, 142 (2017).
- 117.** Lanza, V. F. *et al.* In-depth resistome analysis by targeted metagenomics. *Microbiome* **6**, 11 (2018).
- 118.** Macedo, G. *et al.* Targeted metagenomics reveals inferior resilience of farm soil resistome compared to soil microbiome after manure application. *Sci. Total Environ.* **770**, 145399 (2021).
- 119.** Siljanen, H. M. P. *et al.* Targeted metagenomics using probe capture detect a larger diversity of nitrogen and methane cycling genes in complex microbial communities than traditional metagenomics. Preprint at <https://doi.org/10.1101/2022.11.04.515048> (2025).
- 120.** Humily, F. *et al.* Development of a targeted metagenomic approach to study a genomic region involved in light harvesting in marine *Synechococcus*. *FEMS Microbiol. Ecol.* **88**, 231–249 (2014).
- 121.** Grieb, A. *et al.* A pipeline for targeted metagenomics of environmental bacteria. *Microbiome* **8**, 21 (2020).
- 122.** Li, H.-Z. *et al.* Active antibiotic resistome in soils unraveled by single-cell isotope probing and targeted metagenomics. *Proc. Natl. Acad. Sci.* **119**, e2201473119 (2022).
- 123.** Vollmers, J., Correa Cassal, M. & Kaster, A.-K. Midi-metagenomics: A novel approach for cultivation independent microbial genome reconstruction from environmental samples. Preprint at <https://doi.org/10.1101/2023.01.26.525644> (2023).
- 124.** Langenfeld, K. *et al.* Development of a quantitative metagenomic approach to establish quantitative limits and its application to viruses. *Nucleic Acids Res.* **53**, gkaf118 (2025).
- 125.** Zaramela, L. S., Tjuanta, M., Moyne, O., Neal, M. & Zengler, K. synDNA: a synthetic DNA spike-in method for absolute quantification of shotgun metagenomic sequencing. *mSystems* **7**, e00447-22 (2022).
- 126.** Crossette, E. *et al.* Metagenomic quantification of genes with internal standards. *mBio* **12**, e03173-20 (2021).
- 127.** Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* **8**, 1–6 (2017).

- 128.** Manor, O. & Borenstein, E. MUSiCC: a marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome. *Genome Biol.* **16**, 1–20 (2015).
- 129.** Lin, H. & Peddada, S. D. Analysis of compositions of microbiomes with bias correction. *Nat. Commun.* **11**, 3514 (2020).
- 130.** Lin, H. & Peddada, S. D. Multigroup analysis of compositions of microbiomes with covariate adjustments and repeated measures. *Nat. Methods* **21**, 83–91 (2024).
- 131.** Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10**, 1196–1199 (2013).
- 132.** Wang, S., Ventolero, M., Hu, H. & Li, X. A revisit to universal single-copy genes in bacterial genomes. *Sci. Rep.* **12**, 14550 (2022).
- 133.** Nayfach, S. & Pollard, K. S. Toward accurate and quantitative comparative metagenomics. *Cell* **166**, 1103–1116 (2016).
- 134.** Beszteri, B., Temperton, B., Frickenhaus, S. & Giovannoni, S. J. Average genome size: a potential source of bias in comparative metagenomics. *ISME J.* **4**, 1075–1077 (2010).
- 135.** Pal, C., Bengtsson-Palme, J., Kristiansson, E. & Larsson, D. G. J. The structure and diversity of human, animal and environmental resistomes. *Microbiome* **4**, 1–15 (2016).
- 136.** Bengtsson-Palme, J., Larsson, D. G. J. & Kristiansson, E. Using metagenomics to investigate human and environmental resistomes. *J. Antimicrob. Chemother.* **72**, 2690–2703 (2017).
- 137.** Nayfach, S. & Pollard, K. S. Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol.* **16**, 1103–1116 (2015).
- 138.** Chouvarine, P., Wiehlmann, L., Moran Losada, P., DeLuca, D. S. & Tümmeler, B. Filtration and normalization of sequencing read data in whole-metagenome shotgun samples. *PLOS ONE* **11**, e0165015 (2016).
- 139.** Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
- 140.** Zhao, C., Shi, Z. J. & Pollard, K. S. Pitfalls of genotyping microbial communities with rapidly growing genome collections. *Cell Syst.* **14**, 160-176.e3 (2023).

- 141.** The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
- 142.** Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
- 143.** Jonsson, V., Österlund, T., Nerman, O. & Kristiansson, E. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. *BMC Genomics* **17**, 78 (2016).
- 144.** Pereira, M. B., Wallroth, M., Jonsson, V. & Kristiansson, E. Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics* **19**, 274 (2018).
- 145.** Riazi, K. *et al.* The prevalence and incidence of NAFLD worldwide: a systematic review and meta-analysis. *Lancet Gastroenterol. Hepatol.* **7**, 851–861 (2022).
- 146.** EASL. EASL–EASD–EASO Clinical practice guidelines for the management of non-alcoholic fatty liver disease. *J. Hepatol.* **64**, 1388–1402 (2016).
- 147.** Allen, A. M. *et al.* Clinical course of non-alcoholic fatty liver disease and the implications for clinical trial design. *J. Hepatol.* **77**, 1237–1245 (2022).
- 148.** Lekakis, V. & Papatheodoridis, G. V. Natural history of metabolic dysfunction-associated steatotic liver disease. *Eur. J. Intern. Med.* **122**, 3–10 (2024).
- 149.** Powell, E. E., Wong, V. W.-S. & Rinella, M. Non-alcoholic fatty liver disease. *The Lancet* **397**, 2212–2224 (2021).
- 150.** Yki-Järvinen, H. Non-alcoholic fatty liver disease as a cause and a consequence of metabolic syndrome. *Lancet Diabetes Endocrinol.* **2**, 901–910 (2014).
- 151.** Di Vincenzo, F., Del Gaudio, A., Petito, V., Lopetuso, L. R. & Scaldaferri, F. Gut microbiota, intestinal permeability, and systemic inflammation: a narrative review. *Intern. Emerg. Med.* **19**, 275–293 (2024).
- 152.** Albillos, A., De Gottardi, A. & Rescigno, M. The gut-liver axis in liver disease: pathophysiological basis for therapy. *J. Hepatol.* **72**, 558–577 (2020).
- 153.** Martín-Mateos, R. & Albillos, A. The role of the gut-liver axis in metabolic dysfunction-associated fatty liver disease. *Front. Immunol.* **12**, 660179 (2021).

- 154.** Carpino, G. *et al.* Increased liver localization of lipopolysaccharides in human and experimental NAFLD. *Hepatology* **72**, 470–485 (2020).
- 155.** Li, Q., Rempel, J. D., Yang, J. & Minuk, G. Y. The effects of pathogen-associated molecular patterns on peripheral blood monocytes in patients with non-alcoholic fatty liver disease. *J. Clin. Exp. Hepatol.* **12**, 808–817 (2022).
- 156.** Rivera, C. A. *et al.* Toll-like receptor-4 signaling and Kupffer cells play pivotal roles in the pathogenesis of non-alcoholic steatohepatitis. *J. Hepatol.* **47**, 571–579 (2007).
- 157.** Miele, L. *et al.* Increased intestinal permeability and tight junction alterations in non-alcoholic fatty liver disease. *Hepatology* **49**, 1877–1887 (2009).
- 158.** Wang, W. *et al.* Tauroursodeoxycholic acid inhibits intestinal inflammation and barrier disruption in mice with non-alcoholic fatty liver disease. *Br. J. Pharmacol.* **175**, 469–484 (2018).
- 159.** Kuang, J. *et al.* Hyodeoxycholic acid alleviates non-alcoholic fatty liver disease through modulating the gut-liver axis. *Cell Metab.* **35**, 1752-1766.e8 (2023).
- 160.** Yang, C. *et al.* Altered gut microbial profile accompanied by abnormal short chain fatty acid metabolism exacerbates nonalcoholic fatty liver disease progression. *Sci. Rep.* **14**, 22385 (2024).
- 161.** Zhang, R. *et al.* Gut microbial metabolites in MASLD: implications of mitochondrial dysfunction in the pathogenesis and treatment. *Hepatol. Commun.* **8**, 1–21 (2024).
- 162.** Ponziani, F. R., Gasbarrini, A. & Pompili, M. NAFLD or comorbidities, that is the question. *J. Hepatol.* **73**, 723 (2020).
- 163.** Del Chierico, F. *et al.* Gut microbiota profiling of pediatric non-alcoholic fatty liver disease and obese patients unveiled by an integrated meta-omics-based approach. *Hepatology* **65**, 451–464 (2017).
- 164.** Lee, G. *et al.* Distinct signatures of gut microbiome and metabolites associated with significant fibrosis in non-obese NAFLD. *Nat. Commun.* **11**, 4982 (2020).
- 165.** Redondo-Salvo, S. *et al.* Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nat. Commun.* **11**, 3602 (2020).

- 166.** Leinonen, R., Sugawara, H., Shumway, M., & on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).
- 167.** Hoyles, L. *et al.* Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese women. *Nat. Med.* **24**, 1070–1080 (2018).
- 168.** Mardinoglu, A. *et al.* An integrated understanding of the rapid metabolic benefits of a carbohydrate-restricted diet on hepatic steatosis in humans. *Cell Metab.* **27**, 559-571.e5 (2018).
- 169.** Loomba, R. *et al.* Gut microbiome-based metagenomic signature for non-invasive detection of advanced fibrosis in human non-alcoholic fatty liver disease. *Cell Metab.* **25**, 1054-1062.e5 (2017).
- 170.** Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59–64 (2014).
- 171.** Bushnell, B. BBMap (<https://sourceforge.net/projects/bbmap/>). (2014).
- 172.** Andrews, S. FastQC: A quality control tool for high throughput sequence data (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>). (2010).
- 173.** Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
- 174.** Rosero, J. A. *et al.* Reclassification of *Eubacterium rectale* (Hauduroy *et al.* 1937) in a new genus *Agathobacter* gen. nov. as *Agathobacter rectalis* comb. nov., and description of *Agathobacter ruminis* sp. nov., isolated from the rumen contents of sheep and cows. *Int. J. Syst. Evol. Microbiol.* **66**, 768–773 (2016).
- 175.** Togo, A. H. *et al.* Description of *Mediterraneibacter massiliensis*, gen. nov., sp. nov., a new genus isolated from the gut microbiota of an obese patient and reclassification of *Ruminococcus faecis*, *Ruminococcus lactaris*, *Ruminococcus torques*, *Ruminococcus gnavus* and *Clostridium glycyrrhizinilyticum* as *Mediterraneibacter faecis* comb. nov., *Mediterraneibacter lactaris* comb. nov., *Mediterraneibacter torques* comb. nov., *Mediterraneibacter gnavus* comb. nov. and *Mediterraneibacter glycyrrhizinilyticus* comb. nov. *Antonie Van Leeuwenhoek* **111**, 2107–2128 (2018).
- 176.** Murotomi, K., Tourlousse, D. M., Hamajima, M. & Sekiguchi, Y. Complete genome sequence of *Roseburia faecis* M72/1^T. *Microbiol. Resour. Announc.* **14**, e01253-24 (2025).

- 177.** Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
- 178.** Minh, B. Q. *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
- 179.** Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
- 180.** Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
- 181.** Eddy, S. HMMER (<http://hmmer.org/>). (2019).
- 182.** Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
- 183.** Wickham, H. *et al.* Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
- 184.** Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).
- 185.** Kassambara, A. *ggpubr: ggplot2 based publication ready plots*. (2023).
- 186.** Kolde, R. *pheatmap: pretty heatmaps*. (2019).
- 187.** Kassambara, A. *rstatix: pipe-friendly framework for basic statistical tests*. (2023).
- 188.** Lanthier, N. *et al.* Microbiota analysis and transient elastography reveal new extra-hepatic components of liver steatosis and fibrosis in obese patients. *Sci. Rep.* **11**, 659 (2021).
- 189.** Boursier, J. *et al.* The severity of non-alcoholic fatty liver disease is associated with gut dysbiosis and shift in the metabolic function of the gut microbiota. *Hepatology* **63**, 764–775 (2016).
- 190.** Caussy, C. *et al.* A gut microbiome signature for cirrhosis due to non-alcoholic fatty liver disease. *Nat. Commun.* **10**, 1406 (2019).
- 191.** Oh, T. G. *et al.* A universal gut-microbiome-derived signature predicts cirrhosis. *Cell Metab.* **32**, 878–888.e6 (2020).

- 192.** Shen, F. *et al.* Gut microbiota dysbiosis in patients with non-alcoholic fatty liver disease. *Hepatobiliary Pancreat. Dis. Int.* **16**, 375–381 (2017).
- 193.** Aron-Wisnewsky, J. *et al.* Gut microbiota and human NAFLD: disentangling microbial signatures from metabolic disorders. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 279–297 (2020).
- 194.** McPherson, S. *et al.* Age as a confounding factor for the accurate non-invasive diagnosis of advanced NAFLD fibrosis. *Am. J. Gastroenterol.* **112**, 740–751 (2017).
- 195.** Chu, H., Duan, Y., Yang, L. & Schnabl, B. Small metabolites, possible big changes: a microbiota-centered view of non-alcoholic fatty liver disease. *Gut* **68**, 359–370 (2019).
- 196.** Dai, X. *et al.* Microbial metabolites: critical regulators in NAFLD. *Front. Microbiol.* **11**, 567654 (2020).
- 197.** Martin-Grau, M. & Monleón, D. The role of microbiota-related co-metabolites in MASLD progression: a narrative review. *Curr. Issues Mol. Biol.* **46**, 6377–6389 (2024).
- 198.** Oliphant, K. & Allen-Vercoe, E. Macronutrient metabolism by the human gut microbiome: major fermentation by-products and their impact on host health. *Microbiome* **7**, 91 (2019).
- 199.** Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.* **14**, 20–32 (2016).
- 200.** Iruzubieta, P., Medina, J. M., Fernández-López, R., Crespo, J. & De La Cruz, F. A role for gut microbiome fermentative pathways in fatty liver disease progression. *J. Clin. Med.* **9**, 1369 (2020).
- 201.** Miller, T. L. & Wolin, M. J. Pathways of acetate, propionate, and butyrate formation by the human fecal microbial flora. *Appl. Environ. Microbiol.* **62**, 1589–1592 (1996).
- 202.** Koh, A., De Vadder, F., Kovatcheva-Datchary, P. & Bäckhed, F. From dietary fiber to host physiology: short-chain fatty acids as key bacterial metabolites. *Cell* **165**, 1332–1345 (2016).
- 203.** Morrison, D. J. & Preston, T. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut Microbes* **7**, 189–200 (2016).
- 204.** Blaak, E. E. *et al.* Short chain fatty acids in human gut and metabolic health. *Benef. Microbes* **11**, 411–455 (2020).

205. Gaudier, E. *et al.* Butyrate specifically modulates *MUC* gene expression in intestinal epithelial goblet cells deprived of glucose. *Am. J. Physiol.-Gastrointest. Liver Physiol.* **287**, G1168–G1174 (2004).
206. Wang, H.-B., Wang, P.-Y., Wang, X., Wan, Y.-L. & Liu, Y.-C. Butyrate enhances Intestinal epithelial barrier function via up-regulation of tight junction protein claudin-1 transcription. *Dig. Dis. Sci.* **57**, 3126–3135 (2012).
207. Donohoe, D. R. *et al.* The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon. *Cell Metab.* **13**, 517–526 (2011).
208. Furusawa, Y. *et al.* Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. *Nature* **504**, 446–450 (2013).
209. Singh, N. *et al.* Activation of Gpr109a, receptor for niacin and the commensal metabolite butyrate, suppresses colonic inflammation and carcinogenesis. *Immunity* **40**, 128–139 (2014).
210. Schulthess, J. *et al.* The short-chain fatty acid butyrate imprints an antimicrobial program in macrophages. *Immunity* **50**, 432–445.e7 (2019).
211. Chang, P. V., Hao, L., Offermanns, S. & Medzhitov, R. The microbial metabolite butyrate regulates intestinal macrophage function via histone deacetylase inhibition. *Proc. Natl. Acad. Sci.* **111**, 2247–2252 (2014).
212. Fernando, M. R., Saxena, A., Reyes, J.-L. & McKay, D. M. Butyrate enhances antibacterial effects while suppressing other features of alternative activation in IL-4-induced macrophages. *Am. J. Physiol.-Gastrointest. Liver Physiol.* **310**, G822–G831 (2016).
213. Gao, Z. *et al.* Butyrate improves insulin sensitivity and increases energy expenditure in mice. *Diabetes* **58**, 1509–1517 (2009).
214. Li, Z. *et al.* Butyrate reduces appetite and activates brown adipose tissue via the gut-brain neural circuit. *Gut* **67**, 1269–1279 (2018).
215. Den Besten, G. *et al.* Short-chain fatty acids protect against high-fat diet–induced obesity via a PPAR γ -dependent switch from lipogenesis to fat oxidation. *Diabetes* **64**, 2398–2408 (2015).
216. De Vadder, F. *et al.* Microbiota-generated metabolites promote metabolic benefits via gut-brain neural circuits. *Cell* **156**, 84–96 (2014).

217. Sanna, S. *et al.* Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat. Genet.* **51**, 600–605 (2019).
218. Endo, H., Niioka, M., Kobayashi, N., Tanaka, M. & Watanabe, T. Butyrate-producing probiotics reduce non-alcoholic fatty liver disease progression in rats: new insight into the probiotics for the gut-liver axis. *PLoS ONE* **8**, e63388 (2013).
219. Honma, K., Oshima, K., Takami, S. & Goda, T. Regulation of hepatic genes related to lipid metabolism and antioxidant enzymes by sodium butyrate supplementation. *Metab. Open* **7**, 100043 (2020).
220. Jin, C. J., Sellmann, C., Engstler, A. J., Ziegenhardt, D. & Bergheim, I. Supplementation of sodium butyrate protects mice from the development of non-alcoholic steatohepatitis (NASH). *Br. J. Nutr.* **114**, 1745–1755 (2015).
221. Zhou, D. *et al.* Sodium butyrate attenuates high-fat diet-induced steatohepatitis in mice by improving gut microbiota and gastrointestinal barrier. *World J. Gastroenterol.* **23**, 60 (2017).
222. Zhou, D. *et al.* Sodium butyrate reduces high-fat diet-induced non-alcoholic steatohepatitis through upregulation of hepatic GLP-1R expression. *Exp. Mol. Med.* **50**, 1–12 (2018).
223. Zhou, D. *et al.* Total fecal microbiota transplantation alleviates high-fat diet-induced steatohepatitis in mice via beneficial regulation of gut microbiota. *Sci. Rep.* **7**, 1529 (2017).
224. Vrieze, A. *et al.* Transfer of intestinal microbiota from lean donors increases insulin sensitivity in individuals with metabolic syndrome. *Gastroenterology* **143**, 913-916.e7 (2012).
225. Kootte, R. S. *et al.* Improvement of insulin sensitivity after lean donor feces in metabolic syndrome is driven by baseline intestinal microbiota composition. *Cell Metab.* **26**, 611-619.e6 (2017).
226. Louis, P. & Flint, H. J. Formation of propionate and butyrate by the human colonic microbiota. *Environ. Microbiol.* **19**, 29–41 (2017).
227. Zhu, L. *et al.* Characterization of gut microbiomes in non-alcoholic steatohepatitis (NASH) patients: a connection between endogenous alcohol and NASH. *Hepatology* **57**, 601–609 (2013).
228. Michail, S. *et al.* Altered gut microbial energy and metabolism in children with non-alcoholic fatty liver disease. *FEMS Microbiol. Ecol.* **91**, 1–9 (2015).

- 229.** Meijnikman, A. S. *et al.* Microbiome-derived ethanol in non-alcoholic fatty liver disease. *Nat. Med.* **28**, 2100–2106 (2022).
- 230.** Yuan, J. *et al.* Fatty liver disease caused by high-alcohol-producing *Klebsiella pneumoniae*. *Cell Metab.* **30**, 675-688.e7 (2019).
- 231.** Chen, X. *et al.* Endogenous ethanol produced by intestinal bacteria induces mitochondrial dysfunction in non-alcoholic fatty liver disease. *J. Gastroenterol. Hepatol.* **35**, 2009–2019 (2020).
- 232.** Fromenty, B. & Roden, M. Mitochondrial alterations in fatty liver diseases. *J. Hepatol.* **78**, 415–429 (2023).
- 233.** Ruuskanen, M. O. *et al.* Links between gut microbiome composition and fatty liver disease in a large population sample. *Gut Microbes* **13**, 1888673 (2021).
- 234.** Malaguarnera, G. Gut microbiota in alcoholic liver disease: pathogenetic role and therapeutic perspectives. *World J. Gastroenterol.* **20**, 16639 (2014).
- 235.** Schut, G. J. *et al.* Tungsten enzymes play a role in detoxifying food and antimicrobial aldehydes in the human gut microbiome. *Proc. Natl. Acad. Sci.* **118**, e2109008118 (2021).
- 236.** Pony, P., Rapisarda, C., Terradot, L., Marza, E. & Fronzes, R. Filamentation of the bacterial bi-functional alcohol/aldehyde dehydrogenase AdhE is essential for substrate channeling and enzymatic regulation. *Nat. Commun.* **11**, 1426 (2020).
- 237.** Hitschler, L., Nissen, L. S., Kuntz, M. & Basen, M. Alcohol dehydrogenases AdhE and AdhB with broad substrate ranges are important enzymes for organic acid reduction in *Thermoanaerobacter* sp. strain X514. *Biotechnol. Biofuels* **14**, 187 (2021).
- 238.** Pick, A., Rühmann, B., Schmid, J. & Sieber, V. Novel CAD-like enzymes from *Escherichia coli* K-12 as additional tools in chemical production. *Appl. Microbiol. Biotechnol.* **97**, 5815–5824 (2013).
- 239.** Merchel Piovesan Pereira, B., Adil Salim, M., Rai, N. & Tagkopoulos, I. Tolerance to glutaraldehyde in *Escherichia coli* mediated by overexpression of the aldehyde reductase YqhD by YqhC. *Front. Microbiol.* **12**, 680553 (2021).
- 240.** Moreira De Gouveia, M. I., Daniel, J., Garrivier, A., Bernalier-Donadille, A. & Jubelin, G. Diversity of ethanolamine utilization by human commensal *Escherichia coli*. *Res. Microbiol.* **174**, 103989 (2023).

- 241.** Dellomonaco, C., Clomburg, J. M., Miller, E. N. & Gonzalez, R. Engineered reversal of the β -oxidation cycle for the synthesis of fuels and chemicals. *Nature* **476**, 355–359 (2011).
- 242.** Eram, M. & Ma, K. Decarboxylation of pyruvate to acetaldehyde for ethanol production by hyperthermophiles. *Biomolecules* **3**, 578–596 (2013).
- 243.** Ho, K. K. & Weiner, H. Isolation and characterization of an aldehyde dehydrogenase encoded by the *aldB* gene of *Escherichia coli*. *J. Bacteriol.* **187**, 1067–1073 (2005).
- 244.** Ravcheev, D. A., Moussu, L., Smajic, S. & Thiele, I. Comparative genomic analysis reveals novel microcompartment-associated metabolic pathways in the human gut microbiome. *Front. Genet.* **10**, 636 (2019).
- 245.** Dank, A. *et al.* Bacterial microcompartment-dependent 1,2-propanediol utilization of *Propionibacterium freudenreichii*. *Front. Microbiol.* **12**, 679827 (2021).
- 246.** Chen, Y. *et al.* Associations of gut-flora-dependent metabolite trimethylamine-N-oxide, betaine and choline with non-alcoholic fatty liver disease in adults. *Sci. Rep.* **6**, 19076 (2016).
- 247.** Flores-Guerrero, J. L. *et al.* Circulating trimethylamine-N-oxide is associated with all-cause mortality in subjects with non-alcoholic fatty liver disease. *Liver Int.* **41**, 2371–2382 (2021).
- 248.** Wang, Z. *et al.* Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* **472**, 57–63 (2011).
- 249.** Roncal, C. *et al.* Trimethylamine-N-oxide (TMAO) predicts cardiovascular mortality in peripheral artery disease. *Sci. Rep.* **9**, 15580 (2019).
- 250.** Martínez-del Campo, A. *et al.* Characterization and detection of a widely distributed gene cluster that predicts anaerobic choline utilization by human gut bacteria. *mBio* **6**, e00042-15 (2015).
- 251.** Goris, M. *et al.* Increased thermostability of an engineered flavin-containing monooxygenase to remediate trimethylamine in fish protein hydrolysates. *Appl. Environ. Microbiol.* **89**, e00390-23 (2023).
- 252.** Leimkühler, S. The biosynthesis of the molybdenum cofactors in *Escherichia coli*. *Environ. Microbiol.* **22**, 2007–2026 (2020).

253. Ferguson, D. J. & Krzycki, J. A. Reconstitution of trimethylamine-dependent coenzyme M methylation with the trimethylamine corrinoid protein and the isozymes of methyltransferase II from *Methanosarcina barkeri*. *J. Bacteriol.* **179**, 846–852 (1997).
254. Ferguson, D. J., Krzycki, J. A. & Grahame, D. A. Specific roles of methylcobamide:coenzyme M methyltransferase isozymes in metabolism of methanol and methylamines in *Methanosarcina barkeri*. *J. Biol. Chem.* **271**, 5189–5194 (1996).
255. Kurth, J. M. *et al.* Methanogenic archaea use a bacteria-like methyltransferase system to demethoxylate aromatic compounds. *ISME J.* **15**, 3549–3565 (2021).
256. Gendron, A. & Allen, K. D. Overview of diverse methyl/alkyl-coenzyme M reductases and considerations for their potential heterologous expression. *Front. Microbiol.* **13**, 867342 (2022).
257. Thauer, R. K. Methyl (alkyl)-coenzyme M reductases: nickel F-430-containing enzymes involved in anaerobic methane formation and in anaerobic oxidation of methane or of short chain alkanes. *Biochemistry* **58**, 5198–5220 (2019).
258. Wagner, T., Koch, J., Ermler, U. & Shima, S. Methanogenic heterodisulfide reductase (HdrABC-MvhAGD) uses two noncubane [4Fe-4S] clusters for reduction. *Science* **357**, 699–703 (2017).
259. Aziz, I. *et al.* Structural and mechanistic basis of the central energy-converting methyltransferase complex of methanogenesis. *Proc. Natl. Acad. Sci.* **121**, e2315568121 (2024).
260. Rosconi, F. *et al.* A bacterial pan-genome makes gene essentiality strain-dependent and evolvable. *Nat. Microbiol.* **7**, 1580–1592 (2022).
261. Yu, M. K., Fogarty, E. C. & Eren, A. M. Diverse plasmid systems and their ecology across human gut metagenomes revealed by PlasX and MobMess. *Nat. Microbiol.* **9**, 830–847 (2024).
262. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).
263. Tomita, S., Saito, K., Nakamura, T., Sekiyama, Y. & Kikuchi, J. Rapid discrimination of strain-dependent fermentation characteristics among *Lactobacillus* strains by NMR-based metabolomics of fermented vegetable juice. *PLOS ONE* **12**, e0182229 (2017).
264. Forster, A. H. & Gescher, J. Metabolic engineering of *Escherichia coli* for production of mixed-acid fermentation end products. *Front. Bioeng. Biotechnol.* **2**, (2014).

265. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
266. Medina, J. M., Fernández-López, R., Crespo, J. & Cruz, F. D. L. Propionate fermentative genes of the gut microbiome decrease in inflammatory bowel disease. *J. Clin. Med.* **10**, 2176 (2021).
267. Louis, P., McCrae, S. I., Charrier, C. & Flint, H. J. Organization of butyrate synthetic genes in human colonic bacteria: phylogenetic conservation and horizontal gene transfer. *FEMS Microbiol. Lett.* **269**, 240–247 (2007).
268. Cui, Y. *et al.* Production of butyl butyrate from lignocellulosic biomass through *Escherichia coli*-*Clostridium beijerinckii* G117 co-culture. *Process Biochem.* **128**, 58–67 (2023).
269. Gonzalez-Alba, J. M., Baquero, F., Cantón, R. & Galán, J. C. Stratified reconstruction of ancestral *Escherichia coli* diversification. *BMC Genomics* **20**, 936 (2019).
270. Palomino, A. *et al.* Metabolic genes on conjugative plasmids are highly prevalent in *Escherichia coli* and can protect against antibiotic treatment. *ISME J.* **17**, 151–162 (2023).
271. Abriouel, H. *et al.* New insights into the role of plasmids from probiotic *Lactobacillus pentosus* MP-10 in Aloreña table olive brine fermentation. *Sci. Rep.* **9**, 10938 (2019).
272. Pudlo, N. A. *et al.* Phenotypic and genomic diversification in complex carbohydrate-degrading human gut bacteria. *mSystems* **7**, e00947-21 (2022).
273. Murray, C. J. L. *et al.* Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet* **399**, 629–655 (2022).
274. Hassoun-Kheir, N. *et al.* Comparison of antibiotic-resistant bacteria and antibiotic resistance genes abundance in hospital and community wastewater: a systematic review. *Sci. Total Environ.* **743**, 140804 (2020).
275. Karkman, A., Pärnänen, K. & Larsson, D. G. J. Fecal pollution can explain antibiotic resistance gene abundances in anthropogenically impacted environments. *Nat. Commun.* **10**, 80 (2019).
276. Larsson, D. G. J. & Flach, C.-F. Antibiotic resistance in the environment. *Nat. Rev. Microbiol.* **20**, 257–269 (2022).
277. Zhu, Y.-G. *et al.* Continental-scale pollution of estuaries with antibiotic resistance genes. *Nat. Microbiol.* **2**, 16270 (2017).

- 278.** Zhang, Z. *et al.* Assessment of global health risk of antibiotic resistance genes. *Nat. Commun.* **13**, 1553 (2022).
- 279.** Zheng, D. *et al.* A systematic review of antibiotics and antibiotic resistance genes in estuarine and coastal environments. *Sci. Total Environ.* **777**, 146009 (2021).
- 280.** Jara, D. *et al.* Antibiotic resistance in bacterial isolates from freshwater samples in Fildes Peninsula, King George Island, Antarctica. *Sci. Rep.* **10**, 3145 (2020).
- 281.** Klümper, U. *et al.* Broad host range plasmids can invade an unexpectedly diverse fraction of a soil bacterial community. *ISME J.* **9**, 934–945 (2015).
- 282.** Sun, J. *et al.* Antibiotic resistance genes in agricultural soils from the Yangtze River Delta, China. *Sci. Total Environ.* **740**, 140001 (2020).
- 283.** Risely, A. *et al.* Host- plasmid network structure in wastewater is linked to antimicrobial resistance genes. *Nat. Commun.* **15**, 555 (2024).
- 284.** Zhu, Y.-G. *et al.* Diverse and abundant antibiotic resistance genes in Chinese swine farms. *Proc. Natl. Acad. Sci.* **110**, 3435–3440 (2013).
- 285.** Forster, S. C. *et al.* Strain-level characterization of broad host range mobile genetic elements transferring antibiotic resistance from the human microbiome. *Nat. Commun.* **13**, 1445 (2022).
- 286.** Petersen, J. *et al.* A marine plasmid hitchhiking vast phylogenetic and geographic distances. *Proc. Natl. Acad. Sci.* **116**, 20568–20573 (2019).
- 287.** Birmes, L., Freese, H. M. & Petersen, J. RepC_soli: a novel promiscuous plasmid type of *Rhodobacteraceae* mediates horizontal transfer of antibiotic resistances in the ocean. *Environ. Microbiol.* **23**, 5395–5411 (2021).
- 288.** Hehemann, J.-H. *et al.* Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* **464**, 908–912 (2010).
- 289.** Pudlo, N. A. *et al.* Diverse events have transferred genes for edible seaweed digestion from marine to human gut bacteria. *Cell Host Microbe* **30**, 314–328.e11 (2022).
- 290.** Hatosy, S. M. & Martiny, A. C. The ocean as a global reservoir of antibiotic resistance genes. *Appl. Environ. Microbiol.* **81**, 7593–7599 (2015).

- 291.** Berendonk, T. U. *et al.* Tackling antibiotic resistance: the environmental framework. *Nat. Rev. Microbiol.* **13**, 310–317 (2015).
- 292.** Zhuang, M. *et al.* Distribution of antibiotic resistance genes in the environment. *Environ. Pollut.* **285**, 117402 (2021).
- 293.** Bonanno Ferraro, G. *et al.* Global quantification and distribution of antibiotic resistance genes in oceans and seas: Anthropogenic impacts and regional variability. *Sci. Total Environ.* **955**, 176765 (2024).
- 294.** Jørgensen, T. S., Xu, Z., Hansen, M. A., Sørensen, S. J. & Hansen, L. H. Hundreds of circular novel plasmids and DNA elements identified in a rat cecum metatranscriptome. *PLoS ONE* **9**, e87924 (2014).
- 295.** Kothari, A. *et al.* Large circular plasmids from groundwater plasmidomes span multiple incompatibility groups and are enriched in multimetal resistance genes. *mBio* **10**, e02899-18 (2019).
- 296.** Androsiuk, L., Shay, T. & Tal, S. Characterization of the environmental plasmidome of the red sea. *Microbiol. Spectr.* **11**, e00400-23 (2023).
- 297.** Rozov, R. *et al.* Recycler: an algorithm for detecting plasmids from *de novo* assembly graphs. *Bioinformatics* **33**, 475–482 (2017).
- 298.** Antipov, D., Raiko, M., Lapidus, A. & Pevzner, P. A. Plasmid detection and assembly in genomic and metagenomic data sets. *Genome Res.* **29**, 961–968 (2019).
- 299.** Pellow, D. *et al.* SCAPP: an algorithm for improved plasmid assembly in metagenomes. *Microbiome* **9**, 144 (2021).
- 300.** Arredondo-Alonso, S., Willems, R. J., Van Schaik, W. & Schürch, A. C. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb. Genomics* **3**, 1–8 (2017).
- 301.** Brown Kav, A., Benhar, I. & Mizrahi, I. A method for purifying high quality and high yield plasmid DNA for metagenomic and deep sequencing approaches. *J. Microbiol. Methods* **95**, 272–279 (2013).
- 302.** Krawczyk, P. S., Lipinski, L. & Dziembowski, A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.* **46**, e35–e35 (2018).

- 303.** Pellow, D., Mizrahi, I. & Shamir, R. PlasClass improves plasmid sequence classification. *PLOS Comput. Biol.* **16**, e1007781 (2020).
- 304.** Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. C. & De La Cruz, F. Mobility of plasmids. *Microbiol. Mol. Biol. Rev.* **74**, 434–452 (2010).
- 305.** Francia, M. V. *et al.* A classification scheme for mobilization regions of bacterial plasmids. *FEMS Microbiol. Rev.* **28**, 79–100 (2004).
- 306.** Garcillán-Barcia, M. P., Francia, M. V. & De La Cruz, F. The diversity of conjugative relaxases and its application in plasmid classification. *FEMS Microbiol. Rev.* **33**, 657–687 (2009).
- 307.** Garcillán-Barcia, M. P., Redondo-Salvo, S. & De La Cruz, F. Plasmid classifications. *Plasmid* **126**, 102684 (2023).
- 308.** Richardson, L. *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).
- 309.** Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* **2**, 150023 (2015).
- 310.** Coelho, L. P. *et al.* Towards the biogeography of prokaryotic genes. *Nature* **601**, 252–256 (2022).
- 311.** Munk, P. *et al.* Genomic analysis of sewage from 101 countries reveals global landscape of antimicrobial resistance. *Nat. Commun.* **13**, 7251 (2022).
- 312.** Youngblut, N. D. *et al.* Large-scale metagenome assembly reveals novel animal-associated microbial genomes, biosynthetic gene clusters, and other genetic diversity. *mSystems* **5**, e01045-20 (2020).
- 313.** Pu, L. & Shamir, R. 4CAC: 4-class classifier of metagenome contigs using machine learning and assembly graphs. *Nucleic Acids Res.* **52**, e94–e94 (2024).
- 314.** Garcillán-Barcia, M., Redondo-Salvo, S., Vielva, L. & Cruz, F. D. L. MOBscan: automated annotation of MOB relaxases. in *Horizontal gene transfer: methods and protocols* 295–308 (New York: Springer, 2020).
- 315.** Haft, D. H. *et al.* TIGRFAMs and genome properties in 2013. *Nucleic Acids Res.* **41**, D387–D395 (2012).

- 316.** Gourelé, H., Karlsson-Lindsjö, O., Hayer, J. & Bongcam-Rudloff, E. Simulating Illumina metagenomic data with InSilicoSeq. *Bioinformatics* **35**, 521–522 (2019).
- 317.** Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
- 318.** Feldgarden, M. *et al.* Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob. Agents Chemother.* **63**, e00483-19 (2019).
- 319.** Camargo, A. P. *et al.* Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.* **42**, 1303–1312 (2024).
- 320.** Alikhan, N.-F., Petty, N. K., Ben Zakour, N. L. & Beatson, S. A. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* **12**, 402 (2011).
- 321.** Revell, L. J. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). *PeerJ* **12**, e16505 (2024).
- 322.** Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. GGTREE: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
- 323.** Eisenhofer, R., Odriozola, I. & Alberdi, A. Impact of microbial genome completeness on metagenomic functional inference. *ISME Commun.* **3**, 12 (2023).
- 324.** Maguire, F. *et al.* Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic islands. *Microb. Genomics* **6**, 1–12 (2020).
- 325.** Arumugam, K. *et al.* Recovery of complete genomes and non-chromosomal replicons from activated sludge enrichment microbial communities with long read metagenome sequencing. *Npj Biofilms Microbiomes* **7**, 23 (2021).
- 326.** Johnson, J., Soehnlén, M. & Blankenship, H. M. Long read genome assemblers struggle with small plasmids. *Microb. Genomics* **9**, 1–8 (2023).
- 327.** Marx, V. Why the ocean virome matters. *Nat. Methods* **19**, 924–927 (2022).
- 328.** Yi, Y. *et al.* A systematic analysis of marine lysogens and proviruses. *Nat. Commun.* **14**, 6013 (2023).

- 329.** Dupont, C. L. *et al.* Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J.* **6**, 1186–1199 (2012).
- 330.** Ramfelt, O., Freel, K. C., Tucker, S. J., Nigro, O. D. & Rappé, M. S. Isolate-anchored comparisons reveal evolutionary and functional differentiation across SAR86 marine bacteria. *ISME J.* **18**, wræ227 (2024).
- 331.** Von Wintersdorff, C. J. H. *et al.* Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front. Microbiol.* **7**, 1–10 (2016).
- 332.** Che, Y. *et al.* Conjugative plasmids interact with insertion sequences to shape the horizontal transfer of antimicrobial resistance genes. *Proc. Natl. Acad. Sci.* **118**, e2008731118 (2021).
- 333.** Nielsen, T. K., Browne, P. D. & Hansen, L. H. Antibiotic resistance genes are differentially mobilized according to resistance mechanism. *GigaScience* **11**, giac072 (2022).
- 334.** Coluzzi, C. & Rocha, E. P. C. The spread of antibiotic resistance is driven by plasmids among the fastest evolving and of broadest host range. *Mol. Biol. Evol.* **42**, 1–12 (2025).
- 335.** Quinones-Olvera, N. *et al.* Diverse and abundant phages exploit conjugative plasmids. *Nat. Commun.* **15**, 3197 (2024).
- 336.** Ndiaye, M. *et al.* When less is more: sketching with minimizers in genomics. *Genome Biol.* **25**, 270 (2024).
- 337.** Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- 338.** Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
- 339.** Drula, E. *et al.* The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res.* **50**, D571–D577 (2022).
- 340.** Andreu-Sánchez, S. *et al.* A benchmark of genetic variant calling pipelines using metagenomic short-read sequencing. *Front. Genet.* **12**, 648229 (2021).
- 341.** Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).

List of abbreviations

ARG	Antibiotic resistance gene
bp	Base pairs
Contig	Contiguous sequence
EC	Enzyme Commission
FMT	Fecal microbiota transplantation
GLA	Gut-liver axis
GM	Gut microbiome
HGT	Horizontal gene transfer
HMM	Hidden Markov model
ICE	Integrative and conjugative element
KO	Kegg Orthology
MAG	Metagenome-assembled genome
MAP	Marine plasmid
MASLD	Metabolic dysfunction-associated steatotic liver disease
MASH	Metabolic dysfunction-associated steatohepatitis
MCR	Methyl-CoM reductase
MGE	Mobile genetic element
MOB	Mobilization (typing)
ORF	Open reading frame
PCR	Polymerase chain reaction
PTU	Plasmid taxonomic unit
RPKM	Reads per kilobase per million reads
RPKSM	Reads per kilobase per size per million reads
RLX	Relaxase
rRNA	Ribosomal RNA
SCA	Short chain alcohol
SRA	Sequence Read Archive
TMA	Trimethylamine

TMAO	Trimethylamine N-oxide
UHGG	Unified Human Gastrointestinal Genome
USCG	Universal single-copy gene

List of figures

Figure I-1: Distribution of cultured bacterial genera across different biomes.....	24
Figure I-2: Metataxonomics protocol for microbiome analysis using the 16S rRNA gene.	27
Figure I-3: Metagenomics protocol for microbiome analysis.....	29
Figure I-4: Overview of <i>de novo</i> assembly algorithm.	32
Figure I-5: HUMAnN workflow overview.	35
Figure I-6: Genomic catalogues from metagenomic studies.....	37
Figure I-7: Microbial genomic enrichment strategies.....	42
Figure I-8: Compositional problem in metagenomics.....	44
Figure I-9: Gene abundance metrics used in metagenomics.	45
Table I-1: Bias sources in metagenomic gene quantification and normalization strategies. ..	48
Figure III-1: Overview of MASLD progression.	57
Figure III-2: Gut-liver axis structure.....	58
Figure III-3: Overview of the methodology used in Chapter I.	60
Table III-1: MASLD patient cohorts and fecal metagenomic datasets analyzed in Chapter I..	61
Equation III-1: RPKSM formula.	64
Figure III-4: Taxonomic profiling of the GM in MASLD.	69
Figure III-5: Abundance of butyrate-producing genes in MASLD.....	73
Figure III-6: Abundance of SCA-producing genes in MASLD.....	77
Figure III-7: Abundance of methane and TMA-producing genes in MASLD.	81
Figure III-8: Presence of candidate metabolic genes in the human GM.....	85
Figure III-9: Geno-metabolic alterations associated with MASLD.....	89
Figure IV-1: RLX organization and function in transmissible plasmids.....	99
Figure IV-2: Overview of the methodology used in Chapter II.	100
Table IV-1: Environmental metagenomic datasets analyzed in Chapter II.....	102
Table IV-2: USCGs analyzed in Chapter II.	103
Table IV-3: Genomic composition of the synthetic communities.....	104
Figure IV-3: Distribution of RLXs in marine and vertebrate GM MAGs.	109
Figure IV-4: RLX and USCG prevalence in the metagenomes.	111
Figure IV-5: RLX prevalence in the metagenomes, by ORF size.....	111

Figure IV-6: Taxonomic composition of bacterial contigs.....	112
Figure IV-7: Phylogenetic tree of MOB _B RLXs.....	113
Figure IV-8: ARG prevalence and composition in the metagenomes.....	116
Equation IV-1: Min-max normalization and abundance difference calculation.....	117
Figure IV-9: Comparison between genomic and metagenomic ORF abundances.....	118
Figure IV-10: Genomic content classified by molecular origin in the metagenomes.....	119
Table V-1: Summary of gene-phenotype associations investigated.....	138

Supplementary figures

- S-III-1** Top abundant GM species.
- S-III-2** *Agathobacter rectalis* abundance across MASLD stages.
- S-III-3** Abundance of Mtb and Mta-complex genes in MASLD.
- S-III-4** USCG abundance in MASLD.
- S-III-5** Abundance of candidate KO groups in MASLD.
- S-IV-1** Taxonomic distribution of marine, bacterial MOB+ MAGs.
- S-IV-2** Abundance of marine, MAG-encoded RLXs.
- S-IV-3** Bacterial MAG quality.
- S-IV-4** Geographic distribution of metagenomic samples.
- S-IV-5** RLX prevalence in metagenomes, by MOB class.
- S-IV-6** Geographic distribution of the marine and soil metagenomic samples.
- S-IV-7** Distribution of RLXs in closed genomes.
- S-IV-8** Detection of MAPs (A) pLA6_012 and (B) the pP72_e.
- S-IV-9** Phylogenetic tree of MOB_C RLXs.
- S-IV-10** Phylogenetic tree of MOB_F RLXs.
- S-IV-11** Phylogenetic tree of MOB_H RLXs.
- S-IV-12** Phylogenetic tree of MOB_M RLXs.
- S-IV-13** Phylogenetic tree of MOB_P RLXs.
- S-IV-14** Phylogenetic tree of MOB_Q RLXs.
- S-IV-15** Phylogenetic tree of MOB_V RLXs.
- S-IV-16** Contig and ORF metrics from synthetic communities.

APPENDIX

PROPIONATE GENES IN IBD


This study investigates the role of microbial propionate in inflammatory bowel disease (IBD) by quantifying the abundance of genes involved in its production across a large cohort of metagenomic samples from both healthy individuals and IBD patients. Propionate, a short-chain fatty acid, is synthesized by the GM through multiple metabolic pathways. This analysis focused on the terminal reactions of bacterial propionate synthesis, catalyzed by propionate kinase, propionate CoA transferase, and propionate CoA ligase. These enzymes are encoded by the genes *tdcD*, *pduW*, *pct* and *prpE*. Gene families were defined by linking the corresponding EC numbers to Pfam domains, and HMMs associated with these domains were queried against the UHGG database.

Gene abundance was estimated by aligning metagenomic reads against the reference gene families, with the total number of aligned reads per family used to quantify their presence in each sample. This functional profiling was integrated with metataxonomic and metabolomic data to contextualize the results. Data analysis revealed that propionate is selectively depleted in specific manifestations of IBD. Among the terminal enzymes involved, propionate kinase was the most abundant in the GM. In IBD patients, reductions in fecal propionate levels coincided with significantly lower abundances of genes encoding this enzyme.

Notably, the genes involved in the terminal steps of propionate production exhibited taxonomic shifts in IBD that were not reflected by 16S rRNA-based profiles. This decoupling suggests that changes in gene abundance are not solely driven by shifts in species-level composition, but may instead reflect differences in the accessory genome. These findings underscore the value of gene-centric approaches for detecting functionally relevant microbial alterations associated with gut disease. This work lays the foundation for Chapter I, where the focus shifts from metabolic gene families associated with MASLD to short-chain fatty acid metabolism in the context of IBD.

Article

Propionate Fermentative Genes of the Gut Microbiome Decrease in Inflammatory Bowel Disease

Juan Manuel Medina ^{1,2}, Raúl Fernández-López ¹, Javier Crespo ²  and Fernando de la Cruz ^{1,*}¹ Instituto de Biomedicina y Biotecnología de Cantabria (IBBTec), 39011 Santander, Spain; jmedina@idival.org (J.M.M.); raul.fernandez@unican.es (R.F.-L.)² Clinical and Translational Digestive Research Group, Gastroenterology and Hepatology Department, IDIVAL, Marqués de Valdecilla University Hospital, 39008 Santander, Spain; javier.crespo@scsalud.es

* Correspondence: delacruz@unican.es; Tel.: +34-942-201942

Abstract: Changes in the gut microbiome have been associated with inflammatory bowel disease. A protective role of short chain fatty acids produced by the gut microbiota has been suggested as a causal mechanism. Nevertheless, multi-omic analyses have failed to identify a clear link between changes in specific taxa and disease states. Recently, metagenomic analyses unveiled that gut bacterial species have a previously unappreciated genomic diversity, implying that a geno-centric approach may be better suited to identifying the mechanisms involved. Here, we quantify the abundance of terminal genes in propionate-producing fermentative pathways in the microbiome of a large cohort of healthy subjects and patients with inflammatory bowel disease. The results show that propionate kinases responsible for propionate production in the gut are depleted in patients with Crohn's disease. Our results also indicate that changes in overall species abundances do not necessarily correlate with changes in the abundances of metabolic genes, suggesting that these genes are not part of the core genome. This, in turn, suggests that changes in strain composition may be as important as changes in species abundance in alterations of the gut microbiome associated with pathological conditions.

Keywords: short chain fatty acids; microbiota; Crohn's disease; metagenomics



Citation: Medina, J.M.; Fernández-López, R.; Crespo, J.; Cruz, F.d.l. Propionate Fermentative Genes of the Gut Microbiome Decrease in Inflammatory Bowel Disease. *J. Clin. Med.* **2021**, *10*, 2176. <https://doi.org/10.3390/jcm10102176>

Academic Editor: Gian Paolo Cavaglia

Received: 5 April 2021

Accepted: 7 May 2021

Published: 18 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Many pathological conditions of the gut are linked to changes in the composition of the gut microbiome (GM). Microbes in the gut produce metabolites essential for enterocyte function. Many of these metabolites regulate the integrity of the intestinal barrier, and some were shown to exert an immunomodulatory role [1]. Because of this interplay between the GM metabolism and intestinal function, it has been proposed that changes in the GM may be involved in the genesis and/or evolution of several intestinal diseases. One of the most widely studied associations between GM changes and gut pathology is inflammatory bowel disease (IBD). IBD is a gastrointestinal disorder characterized by a chronic inflammation of the gastrointestinal tract, associated in some cases with extraintestinal signs of systemic inflammation. IBD comprises two clinical manifestations: Crohn's disease (CD) and ulcerative colitis (UC). Studies have repeatedly shown that the GM composition of IBD patients differs from healthy subjects [2]. However, the species involved and the magnitude and sign of the changes observed are highly variable, and most taxa appear to increase their numbers in some studies, but decreasing in others [3]. This variability complicates the elucidation of a causal link between GM alterations and the onset and progress of IBD. However, in recent years, substantial evidence has accumulated on the potential role of short chain fatty acids (SCFAs) as possible mediators between GM alterations and IBD [4].

SCFAs such as acetate, propionate, and butyrate are produced by the GM from the anaerobic fermentation of carbohydrates and amino acids present in dietary fiber [5]. SCFAs are important nutrients for enterocytes, which use them as a primary source of energy. They also promote the development of regulatory T lymphocytes, which modulate

tissue inflammation and cytokine production [6]. These functional roles may have a clinical interaction since dietary information from IBD patients correlated higher fiber intake with a lower risk of CD [7]. It was proposed that SCFA production from dietary fiber by the GM may be implicated in a protective effect on the intestinal epithelium [8]; however, the mechanisms of such an effect are still unclear. Whereas studies in germ-free mice showed that SCFAs promoted T regulatory cell development [9] and protected against T-cell-mediated colitis [10], multi-omic studies in humans have produced less conclusive results. In IBD patients, a depletion in certain butyrate-producing bacteria such as *Faecalibacterium* or *Roseburia* has been observed [11–13]. However, these studies failed to reveal a univocal relationship between changes in GM composition and SCFA concentrations.

One possibility for this discrepancy is that changes in overall species composition, as identified by 16S metataxonomic studies, do not directly correlate with variations in the metabolic capabilities of the GM. Recent studies showed that bacterial species within the human gut exhibit substantial genomic variation, each exhibiting up to hundreds of genomically different species [14]. If taxonomic labels correlate poorly with metabolic capability, it is possible that pathological changes in the GM may not be noticed by 16S metataxonomy. To test this possibility, we studied the correlation of SCFA abundances in IBD patients and healthy subjects with the abundance of GM genes directly involved in SCFA metabolism. Our results show that CD patients exhibit a decrease in propionate that coincides with lower abundances in the terminal genes of the metabolic pathways leading to propionate production. When comparing these decreases with 16S information, we obtained a poor correlation, suggesting that geno-centric approaches such as the one developed here may be better suited for identifying the causal links between GM alterations and gut disease.

2. Materials and Methods

2.1. IBD Cohort Data

For this project, we studied a cohort of 132 IBD patients integrated in the second part of the Human Microbiome Project [15]; specifically, 38 patients diagnosed with UC, 67 with CD and 27 with no IBD (H). As stated in [16], none of these subjects had been diagnosed with known bleeding disorders, acute gastrointestinal infections, hepatitis, or immune-mediated diseases. We analyzed the publicly available multi-omic data from these subjects, allocated in the Inflammatory Bowel Disease Multiomics Database [16]. Specifically, we retrieved two separate tables with 546 metabolic (265 CD, 146 UC, and 135 H) and 178 metataxonomic (86 CD, 46 UC, and 46 H) merged profiles. We also downloaded the processed sequencing files of 1638 metagenomic stool samples (583 CD, 353 UC, and 362 H) from these patients, which were collected every two weeks, processed as detailed in [16] and sequenced on an Illumina HiSeq2500. Finally, we retrieved the metadata with the associations between the stool samples extracted in the original study and their corresponding patients.

2.2. Analysis of Metabolomic and Metataxonomic Data

Metabolic and metataxonomic profiles from the available samples and metadata with the diagnosed condition of the patients and their corresponding samples were parsed and analyzed through in-house-made Bash and R scripts. Plots were elaborated with the R package ggplot2 [17]. Statistical analysis was performed through pairwise Mann–Whitney tests with the Benjamini–Hochberg false discovery rate correction to assess the significance of the differences regarding the metabolic and 16S levels between the three groups of samples corresponding to the CD, UC, and H conditions.

2.3. Extraction of the GM Genes Involved in the Formation of Propionate

We defined the enzymes catalyzing the terminal reactions involved in the formation of microbial propionate through fermentative pathways, namely propionate kinase, propionate CoA transferase, and propionate CoA ligase. Individual reactions and their

corresponding enzyme commission (EC) numbers were targeted through the metabolic-pathways databases MetaCyc, BRENDA, and KEGG [18–20]. After this, the EC numbers of target enzymes were linked to Pfam domains [21] by studying the associations established by ECDomainMiner [22].

The Unified Human Gastrointestinal Genome (UHGG) [14] is the most extensive database of sequenced GM genomes and microbial genes elaborated so far, and it was used in this study to retrieve the families of fermentative genes (FGs) coding for the enzymes involved in the formation of propionate. We specifically selected the 3205 pan-genomes of GM species with at least two characterized strains, which altogether contain more than 21 million genes. The profile hidden Markov model (pHMM) of every Pfam domain associated with a target EC number was queried with HMMER version 3.3 (Howard Hughes Medical Institute, MD, USA) [23] against the pan-genomes for homologous sequences using the *hmmsearch* function, only allowing hits with an e-value less than 0.001 ($-E$ 0.001). Genes retrieved multiple times with different Pfams associated to the same enzyme were deduplicated, and short sequences were eliminated from the resulting sets of genes to remove potential misannotations in the UHGG.

The resulting genes were concatenated to several protein-coding genes with experimentally validated enzymatic activity. The sequences of these genes were retrieved from MetaCyc, and they were used as controls to analyze their phylogenetic distance to the sets of genes defined previously. To do so, the genic groups were aligned with MAFFT version 7.271 [24] and represented in phylogenetic trees with IQ-TREE version 2.0.3 [25]. The trees were inspected to extract phylogenetically related sequences composing each family of FGs involved in the formation of propionate. As a result, we obtained four enzyme-coding FG clusters involved in the last steps of microbial propionate formation: *tdcD* and *pduW* (coding for the propionate kinase), *pct* (coding for the propionate CoA transferase), and *prpE* (coding for the propionate CoA ligase). A graphical representation of the complete workflow is provided in Figure S1.

2.4. Analysis of Gene Abundance in the Metagenomic Samples

We inspected the distribution of FG clusters in the metagenomic samples of the IBD cohort using DIAMOND version 2.0.2 [26]. Briefly, every metagenomic sample was aligned against each FG cluster, allowing only one alignment per sequencing read ($-\text{max-hsps}$ 1) with an e-value less than 0.001 ($-e$ value 0.001) and a percentage of sequence identity between the read and each FG higher than 80% ($-id$ 80). The sum of metagenomic reads aligned to each FG was considered an indicator of the genic abundance in each sample. The samples were posteriorly analyzed for differential abundance between the three conditions through a pairwise Mann–Whitney test with the Benjamini–Hochberg false discovery rate correction to assess the significance of the differences regarding the genic abundance between the three groups of samples corresponding to the CD, UC and H conditions.

3. Results

3.1. Propionate Is Depleted in Some Manifestations of IBD

The abundance of SCFAs in healthy subjects, and CD and UC patients was obtained from the fecal samples of the IBDMDB cohort. Overall abundances were found to be not statistically significant for all groups and SCFAs analyzed, except propionate. Propionate levels were found to be significantly decreased in UC patients compared with healthy subjects (Figure 1). As shown in the figure, average levels were also lower in CD patients, although this difference was not found to be statistically significant.

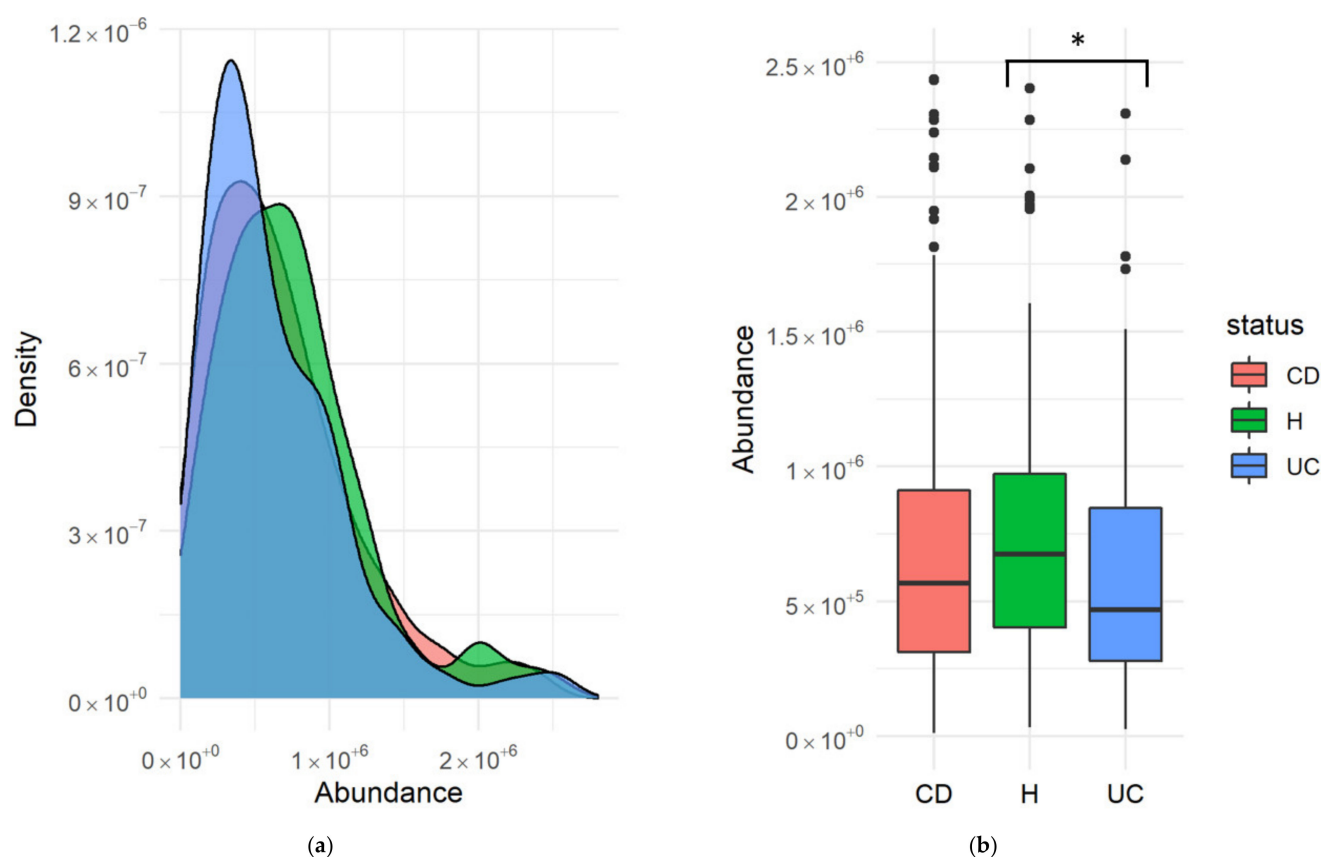
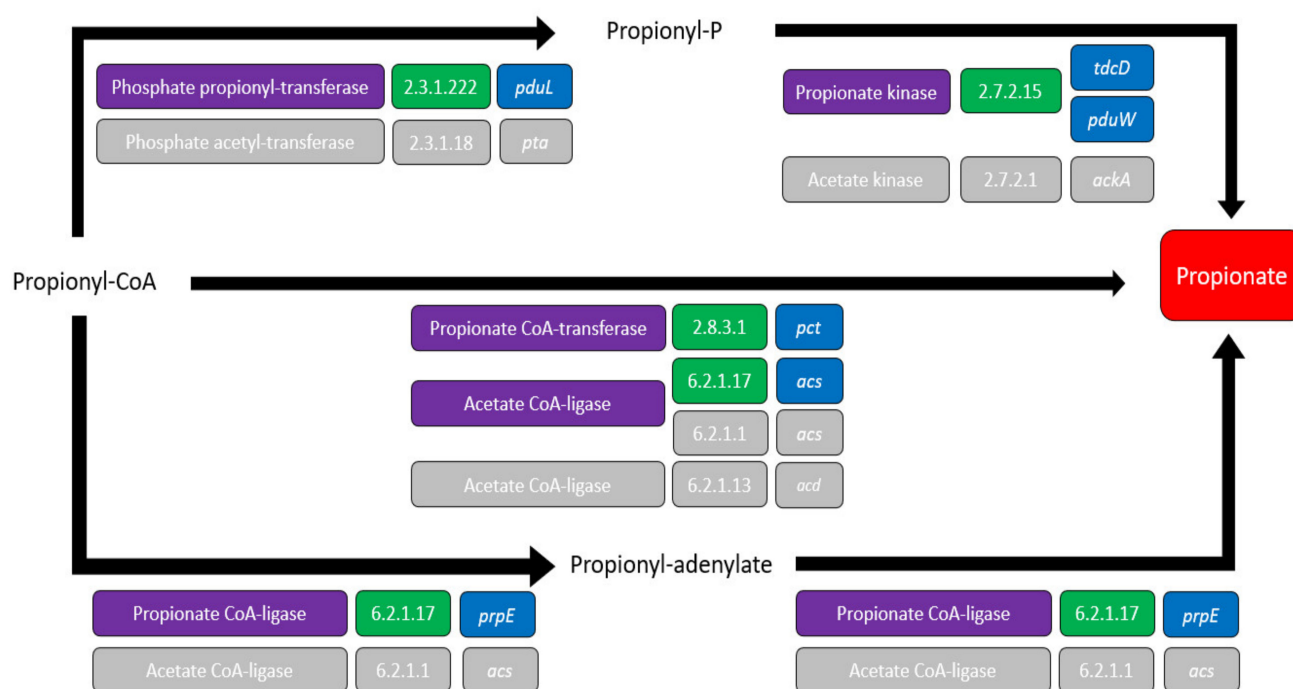


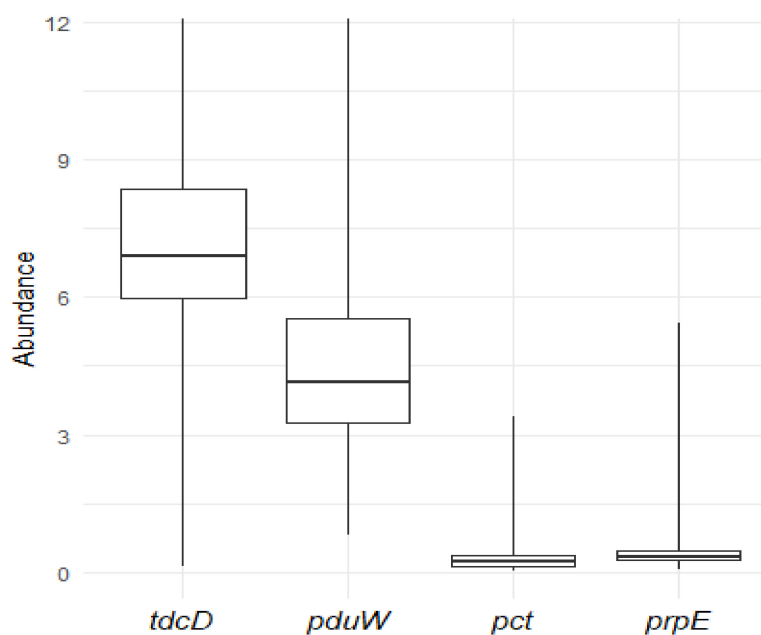
Figure 1. Levels of propionate in healthy subjects (H) and patients with Crohn’s disease (CD) and ulcerative colitis (UC) in absolute concentration units. (a) A kernel density estimate plot represents the abundance of propionate from the metabolomic profiles measured in fecal samples from the cohort analyzed in Lloyd-Price [19]. (b) A boxplot represents the abundance of propionate measured in fecal samples. Horizontal black bars represent median levels, while boxes and whiskers represent the data from first to third quartiles and from the quartiles to the minimum and maximum, respectively. Black dots correspond to outlier values outside the interquartile range. Statistical significance of the differences between groups was evaluated using a pairwise Mann–Whitney test with the Benjamini–Hochberg false discovery rate correction. No significant differences among groups were obtained for CD and H, while significant differences between H and UC were observed with $p = 0.029$ (*).

3.2. Propionate Kinase Is the Most Abundant Terminal Enzyme Involved in the Formation of Microbial Propionate

Once propionate was identified as a potential SCFA altered in UC, we focused on the metabolic pathways leading to its formation. The enzymes catalyzing the terminal reactions involved in the production of propionate and the genes coding for these enzymes, were characterized through metabolic-pathways databases, as detailed in the Materials and Methods. As shown in Figure 2a, propionate can be formed through three different reactions. The first reaction is a dephosphorylation of propionyl-P that yields one ATP. This energetically favorable reaction is catalyzed by a propionate kinase (EC 2.7.2.15) that can be coded by two genes: *tdcD* and *pduW*. Propionate can also be formed through the transference of CoA cofactor in propionyl CoA to another metabolite through the enzyme propionate CoA transferase (EC 2.8.3.1) or acetate CoA ligase (coded respectively by *pct* and *acs*). This CoA-transferase route also conserves the energy of the CoA bond in the newly formed CoA-moiety of the co-substrate. Finally, propionate can be formed from propionyl-adenylate through a propionate CoA ligase (EC 6.2.1.17) encoded by the gene *prpE*. A complete diagram depicting propionate formation from pyruvate is provided in Figure S2.



(a)



(b)

Figure 2. Characterization of the terminal genes involved in the formation of bacterial propionate. (a) Scheme with the three terminal reactions involved in the formation of propionate. Enzymes, EC numbers, and coding genes are represented in purple, green, and blue, respectively. As seen, some of the participating enzymes have substrate broadness between acetate and propionate (depicted in light grey). For example, the enzyme encoded by the *acs* gene forms acetate (EC 6.2.1.1) and propionate (EC 6.2.1.17); (b) boxplot representing the relative abundance of the four terminal genes involved in propionate formation in all the GM metagenomic samples. Horizontal black bars represent median levels, while boxes and whiskers represent the data from first to third quartiles and from the quartiles to the minimum and maximum, respectively. Abundances were measured as indicated in the Materials in Methods and are represented in thousands of reads.

The abundance of genes *tdcD* and *pduW* (coding for the propionate kinase), *pct* (coding for the propionate CoA transferase), and *prpE* (coding for the propionate CoA ligase) was measured in the metagenomic sequencing data from the stool samples of the IBD cohort, as described in the Materials and Methods. The results showed that the relative abundance of propionate kinase genes *tdcD* and *pduW* was higher than both CoA-mediating enzymatic genes (Figure 2b). This suggests that dephosphorylation is the major pathway employed by the GM to produce propionate.

3.3. Terminal Genes Involved in the Synthesis of Propionate Are Differentially Abundant in IBD

The abundance of genes *tdcD* and *pduW* (coding for the propionate kinase), *pct* (coding for the propionate CoA transferase), and *prpE* (coding for the propionate CoA ligase) was compared between the H, CD, and UC conditions. Pairwise Mann–Whitney tests with the Benjamini–Hochberg false discovery rate correction were used to assess the significance of the differences regarding the genic abundance between the three conditions. Abundances are presented as plots of their kernel density estimates. As shown in Figure 3a,b, propionate kinase coding genes *tdcD* and *pduW* are significantly more abundant in healthy samples when compared to CD ($p = 0.0004$ for *tdcD* and $p = 0$ for *pduW*, respectively). No significant differences among groups were obtained for UC and H. As shown in Figure 3c,d, no differences were found in *pct* or *prpE* either, probably due to their low relative abundance (Figure 2b). Figures S3 and Figure S4 depict boxplots representing these differences.

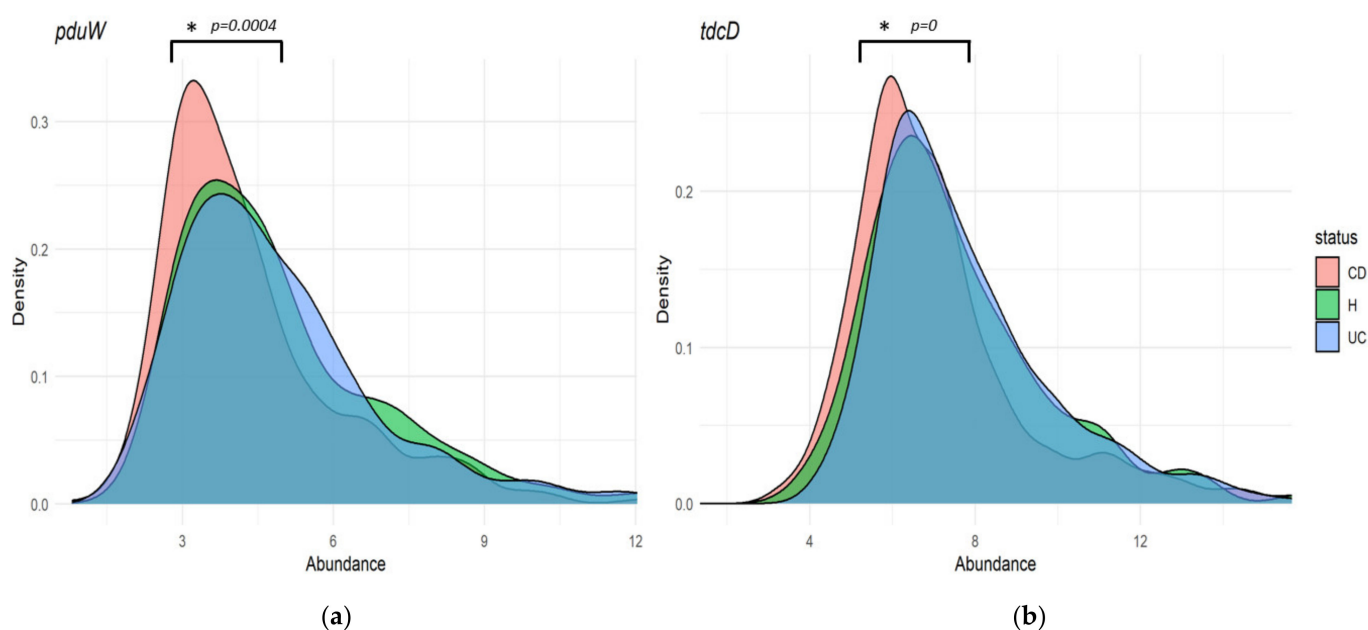


Figure 3. Cont.

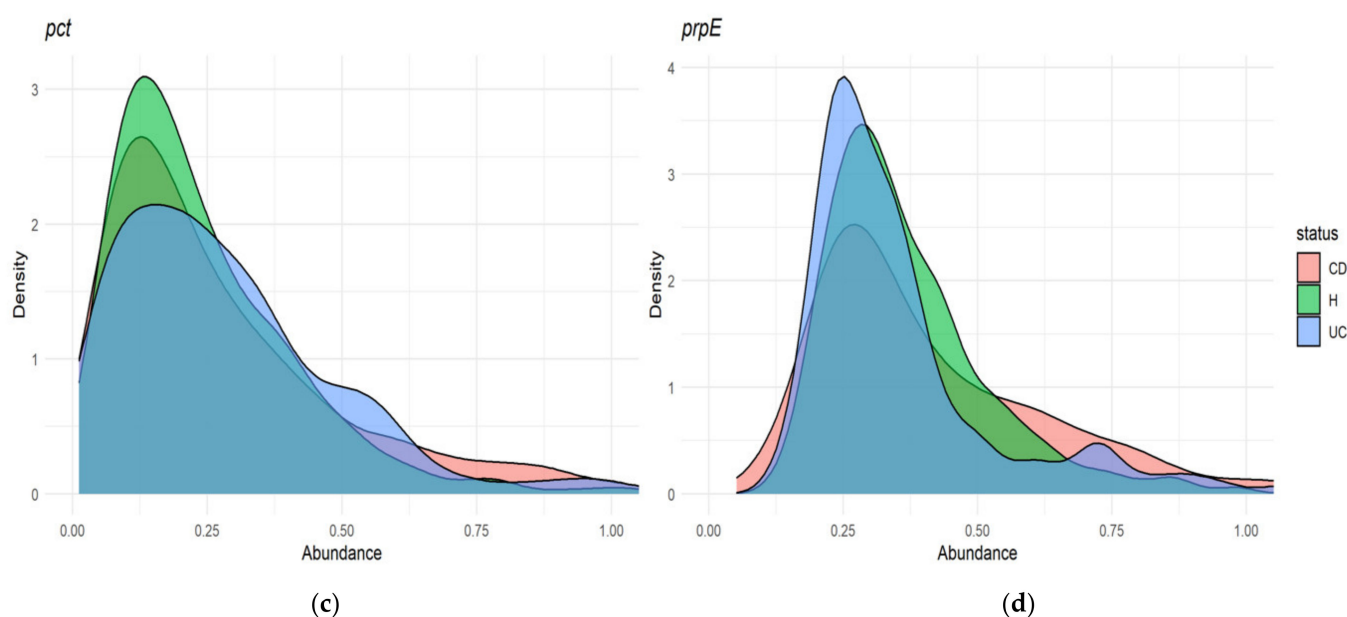
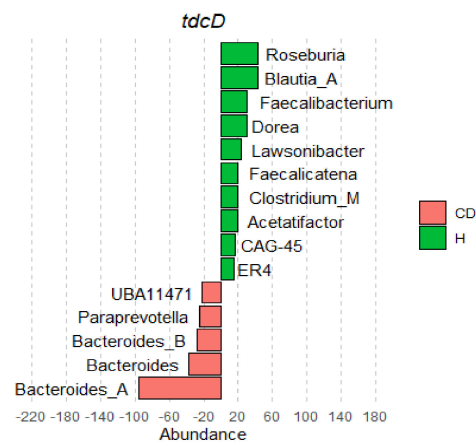


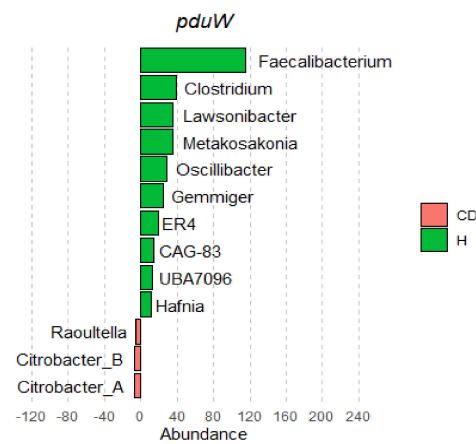
Figure 3. Abundance of genes involved in the terminal formation of propionate in healthy subjects (H) and patients with Crohn’s disease (CD) and ulcerative colitis (UC). (a,b) Propionate kinase genes *tdcD* and *pduW*. (c,d) Propionate CoA transferase gene *pct* and propionate CoA ligase gene *prpE*. Density plots represent the abundances of terminal microbial genes involved in the formation of propionate, from the metagenomic measures of the fecal samples from the cohort analyzed in [16]. Differences in abundances were evaluated using a pairwise Mann–Whitney test with the Benjamini–Hochberg false discovery rate correction. Significant differences among groups were obtained for *tdcD* and *pduW* between CD and H (*).

3.4. Genes Involved in the Last Steps of Microbial Propionate Formation Have Taxonomic Shifts in IBD That Do Not Always Correlate with 16S Abundances

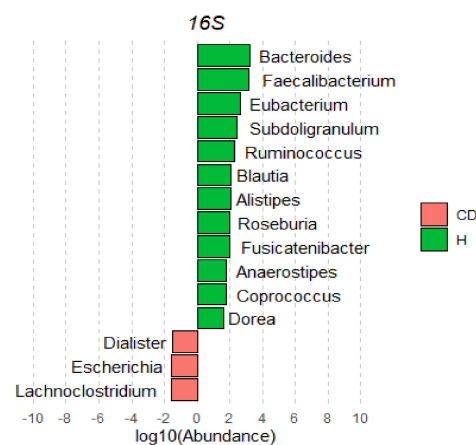
The increases in *tdcD* and *pduW* abundances in the healthy condition compared with CD led us to try to determine whether this differential abundance was caused by changes in specific bacterial taxa. To do so, we calculated the average abundance of each kinase-coding gene in each genus and condition (UC, CD, and H). To compare these changes with absolute bacterial abundances, we retrieved the abundance of each genus in each condition in the same cohort of patients. This was achieved by plotting the 16S metataxonomic profiles obtained from fecal IBDMDDB samples. Average gene levels were normalized, and the values for each genus were compared between H and CD patients (Figure 4a,b) and between H and UC patients (Figure S5). As shown in Figure 4a,b, H subjects showed an increased abundance of both *tdcD* and *pduW* kinase genes in multiple genus from class *Clostridia*. Specifically, kinase genes from members of family *Lachnospiraceae*, such as *Roseburia*, *Blautia*, or *Dorea* spp., as well as family *Ruminococcaceae* such as *Faecalibacterium* spp. were increased in the H condition. In contrast, kinase genes from *Bacteroidales* and, to a lesser extent, *Enterobacterales* were more abundant in CD patients.



(a)



(b)



(c)

Figure 4. *tdcD*, *pduW*, and 16S average gene-abundance differences between healthy and CD conditions. (a,b) Differences in kinase abundances. Horizontal bars in the plots represent the difference in total average gene abundances between H and CD groups for *tdcD* and *pduW*. (c) Differences in average 16S abundance between the CD and H groups. Bars indicate the difference in log₁₀ average abundances between H and CD groups, inferred from the 16S data in [16]. Only genera with the highest net variation (positive or negative) are represented.

These results are in sharp contrast with 16S-derived abundances, which showed that H individuals had an overall increase in *Bacteroides* (Figure 4c). Changes in the abundance of propionate kinase genes thus correlated poorly with changes in taxon abundance, as determined by 16S. This discrepancy may be caused by *tdcD* not being core gene in many of the most significant taxa of the human GM (Table S1). The large genotypic variation observed for species in the GM [14] means that many biochemical pathways are present only in a fraction of the strains of a given species. Thus, changes in *tdcD* abundance may be shifts in certain strains rather than in particular species. However, the discrepancy between *tdcD* and 16S abundances may also be attributed to a methodological artifact. Although *tdcD* is quantified directly from metagenomic data, 16S counts are retrieved from metataxonomic analysis, which imply significant differences in DNA amplification and sequencing. To check whether this discrepancy between the abundances of *tdcD* and 16S counts could be ascribed merely to technical reasons, we measured the overall levels of the core gene *rpoB* in the most significant taxa. *rpoB* codes for the beta subunit of the RNA polymerase, and it was quantified in the metagenomic data from the stool samples, as described in the Materials and Methods. By comparing *tdcD* and *rpoB* abundances, we could thus rule out discrepancies due to different methodological procedures. As presented in Figure 5, *rpoB* and *tdcD* abundances showed a poor correlation, indicating that overall bacterial abundance cannot be used as a proxy to estimate *tdcD* abundance.

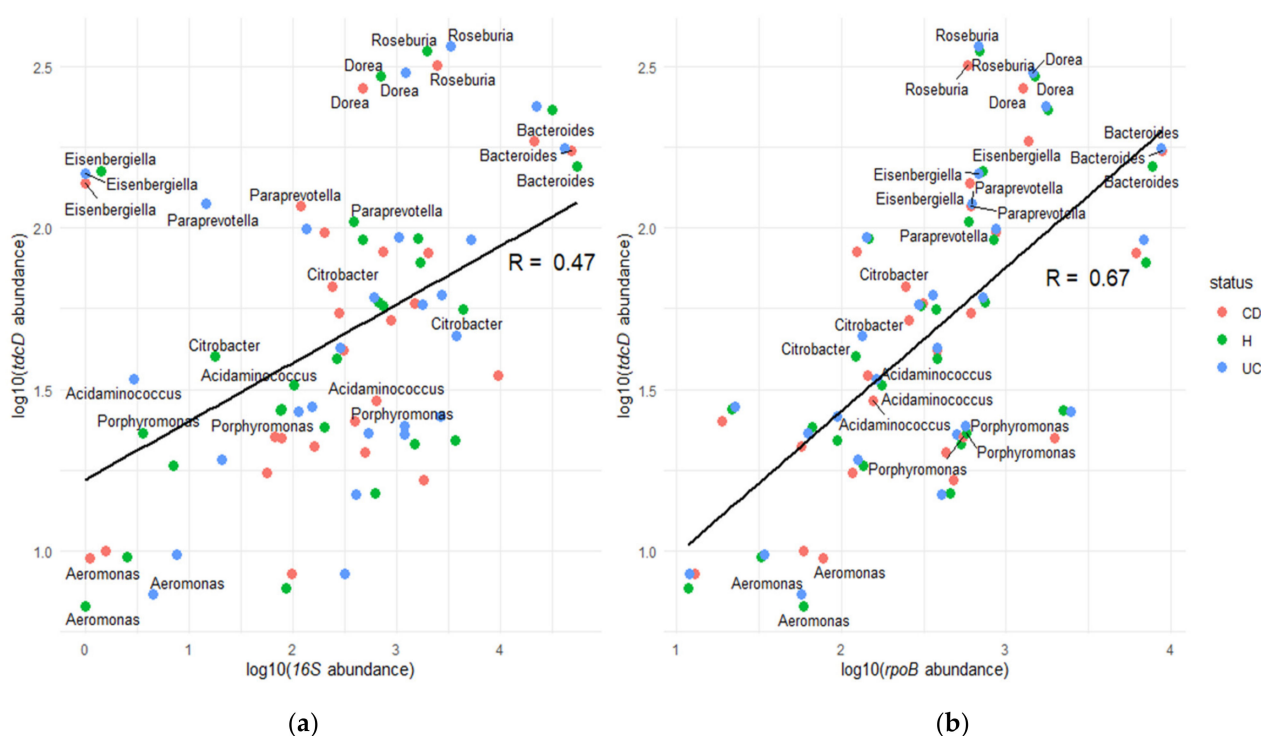


Figure 5. Dot plots between (a) *tdcD* and 16S, and (b) *tdcD* and *rpoB* average gene abundance. Dots in the plot represent the average gene abundance of GM bacterial genera in CD, UC, and healthy conditions on a log10 scale. Only genera with consistent taxonomic annotation between data from [14,16] are presented. Labels indicate bacterial taxa with higher differences between the 16S- and *rpoB*-based quantifications. The regression lines with the correlation coefficient between the corresponding genes are presented in both plots.

4. Discussion

Inflammatory diseases like IBD are complex, multi-trait disorders in which causal links are difficult to identify. There is evidence pointing to the involvement of the GM in development of IBD, and the role of SCFAs in homeostasis and immunomodulation in the gut has been indicated as a possible cause. However, comprehensive multi-omic

studies have failed to univocally ascribe a role for the GM in IBD [13,16]. A major problem arising is the lack of consistency in the relative abundances of different taxa, which appear uncorrelated in healthy and IBD-affected patients in different studies. One possible reason for this discrepancy is the large genomic diversity within the bacterial species in the GM [14]. Since species may hide large genomic variations, alterations in the GM may be due to changes in genes rather than changes in species. From this perspective, we re-analyzed multi-omic results for an IBD cohort [16], trying to link changes in SCFA abundance with differential abundances in the genes responsible for the metabolization of these genes. As shown in Figure 1, propionate was found to be significantly decreased in UC patients compared with healthy subjects. It was also decreased in CD patients, although not significantly.

Propionate is produced by a variety of GM species through several different pathways, including lactate fermentation; succinate degradation; the degradative pathways of certain amino acids, such as alanine and threonine; and the fermentation of pyruvate to propionate through the succinate and acrylate pathways, among others [5]. Analyzing the variety of enzymes involved in these and other relevant pathways would be daunting, but by focusing on the common terminal reactions that directly lead to propionate, we were able to narrow our search. There are two known terminal reactions leading to propionate (Figure 2a). It can be produced from propionyl CoA by propionate CoA transferases and ligases, such as in the acrylate pathway, succinate/propionate conversion, pyruvate and lactate fermentations, and alanine degradation. Alternatively, it can be produced from propionyl phosphate via the reverse activity of propionate kinases, such as in the threonine degradation pathway and the methylcitrate cycle. Our results indicated that, from these, propionate kinases are more abundant in the GM metagenome (Figure 2b). This is a surprising finding, as propionyl-CoA-mediated reactions are much more frequent in fermentative and degradative pathways, and even routes that have a kinase in their final steps (threonine degradation and the methylcitrate pathway) also include a propionyl CoA intermediary. The higher abundance of *tdcD* and *pduW* is thus puzzling, yet it is known that SCFAs, CoA ligases, and transferases show extensive substrate promiscuity. Acetate CoA enzymes can metabolize acetate and propionate in multiple situations [5], as shown by some GM genera such as *Phascolarctobacterium* [27]. Further research is thus required to analyze the extent to which substrate promiscuity plays a role in SCFA production in the GM.

Regarding their relative abundances in H, CD, and UC groups, *prpE* was decreased in UC patients (Figure 3d). This result correlates with the reduction in propionate levels in this IBD manifestation. Propionate kinases were found to be significantly under-represented in CD patients (Figure 3a,b). By ascribing each propionate kinase count to its cognate species, we were able to obtain taxon-specific abundances. These abundances showed that CD patients have an under-representation of counts from *Faecalibacterium*, *Roseburia*, *Blautia*, and *Clostridium*, which coincides with an increased abundance of these genera in healthy subjects, according to 16S (Figure 4c). Previous studies have reported decreased levels of these genera in CD, many of which are known SCFA producers [28,29]. *R. hominis*, for example, was shown to promote gut barrier function and immunity in murine and in-vitro models [30] and it was associated with a protective role against IBD [27]. Similarly, *Faecalibacterium prausnitzii*, one of the most abundant bacteria in GM, was proposed as a potential biomarker given its depletion in CD and UC patients [29]. Our results showed that *tdcD* from *Roseburia* and *tdcD* and *pduW* from *Faecalibacterium* are increased in healthy samples (Figure 4a,b). This implies that the protective role of these taxa is due to the production of propionate through the kinases encoded by these genes.

In other taxa, however, comparing taxon-specific counts and 16S-derived abundances yielded conflicting results (Figure 4). The most obvious discrepancy occurred in *Bacteroides*, which showed a decrease in CD patients, yet *tdcD* genes from *Bacteroides* were increased. This discrepancy may arise from the genomic plasticity associated with many GM species [14]. As metabolic genes are not part of the core genome in many taxa (Table S1), shifts in the strain composition of the GM may alter its metabolic capabilities without re-

flecting in the 16S profile. This may be observed when total abundances and gene-specific abundances are compared (Figure 5). The lack of correlation between these two indicators strongly suggests that changes in strain composition are also a key player in alterations of the GM. This also means that 16S abundances probably yield a skewed view of GM changes. Geno-centric approaches, such as the one developed here, may help to better understand GM alterations associated with pathological conditions. Characterizing the genomic diversity of the GM is thus fundamental to understanding the metabolic activity encoded in the core and accessory genomes of individual species. This will contribute to refining the genic target that constitutes the basis of therapeutic strategies, such as stool transplant protocols, which are developed only from a taxonomic point of view. By linking metabolites and genomic abundances, they may also help us understand the causal links, if any, between changes in the microbiome and the onset and evolution of IBD and other diseases.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/jcm10102176/s1>, Figure S1: Graphical representation of the computational workflow, Figure S2: Propionate formation from pyruvate, Figure S3: Boxplots showing abundance of propionate kinase genes *tdcD* and *pduW*, Figure S4: Boxplots showing abundance of propionate CoA transferase gene *pct* and propionate CoA ligase gene *prpE*, Figure S5: Differences in *tdcD* and *pduW* abundances between H and UC, Table S1: Presence of *tdcD* gene in the strains of the most representative GM species.

Author Contributions: Conceptualization, J.M.M., R.F.-L., J.C., and F.d.l.C.; methodology, J.M.M.; software, J.M.M.; validation, J.M.M. and R.F.-L.; formal analysis, J.M.M., R.F.-L., J.C., and F.d.l.C.; investigation, J.M.M., R.F.-L., J.C., and F.d.l.C.; resources, J.C. and F.d.l.C.; data curation, J.M.M.; writing—original draft preparation, J.M.M., R.F.-L., J.C., and F.d.l.C.; writing—review and editing, J.M.M., R.F.-L., J.C., and F.d.l.C.; visualization, J.M.M.; supervision, J.C. and F.d.l.C.; project administration, J.C. and F.d.l.C.; funding acquisition, J.C. and F.d.l.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by project BFU2017-86378-P from the Spanish Ministry of Science and Innovation (MCINN) to F.d.l.C.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed for this study can be found in the following repositories: UHGG is allocated at <http://ftp.ebi.ac.uk/pub/databases/metagenomics/> (accessed date on 5 January 2021); IBD cohort data is allocated at <https://ibdmdb.org/> (accessed date on 5 January 2021). Analysis scripts elaborated are available at https://github.com/JuanmaMedina/GM_IBD/ (accessed date on 5 January 2021).

Acknowledgments: This computational study was based on the available data from IBDMDb and UHGG. The authors kindly thank the researchers involved in the creation and maintenance of both databases, as well as the software developers involved in the creation and maintenance of all the bioinformatic resources cited in the article.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Włodarska, M.; Luo, C.; Kolde, R.; D’Hennezel, E.; Annand, J.W.; Heim, C.E.; Krastel, P.; Schmitt, E.K.; Omar, A.S.; Creasey, E.A.; et al. Indoleacrylic Acid Produced by Commensal *Peptostreptococcus* Species Suppresses Inflammation. *Cell Host Microbe* **2017**, *22*, 25–37.e6. [CrossRef] [PubMed]
2. Schirmer, M.; Garner, A.; Vlamakis, H.; Xavier, R.J. Microbial genes and pathways in inflammatory bowel disease. *Nat. Rev. Microbiol.* **2019**, *17*, 497–511. [CrossRef]
3. Khan, I.; Ullah, N.; Zha, L.; Bai, Y.; Khan, A.; Zhao, T.; Che, T.; Zhang, C. Alteration of Gut Microbiota in Inflammatory Bowel Disease (IBD): Cause or Consequence? IBD Treatment Targeting the Gut Microbiome. *Pathogens* **2019**, *8*, 126. [CrossRef] [PubMed]

4. Gonçalves, P.; Araújo, J.R.; Di Santo, J.P. A Cross-Talk Between Microbiota-Derived Short-Chain Fatty Acids and the Host Mucosal Immune System Regulates Intestinal Homeostasis and Inflammatory Bowel Disease. *Inflamm. Bowel Dis.* **2018**, *24*, 558–572. [\[CrossRef\]](#)
5. Louis, P.; Flint, H.J. Formation of propionate and butyrate by the human colonic microbiota. *Environ. Microbiol.* **2017**, *19*, 29–41. [\[CrossRef\]](#)
6. Kim, M.H.; Kang, S.G.; Park, J.H.; Yanagisawa, M.; Kim, C.H. Short-Chain Fatty Acids Activate GPR41 and GPR43 on Intestinal Epithelial Cells to Promote Inflammatory Responses in Mice. *Gastroenterology* **2013**, *145*, 396–406.e10. [\[CrossRef\]](#)
7. Ananthakrishnan, A.N.; Khalili, H.; Konijeti, G.G.; Higuchi, L.M.; de Silva, P.; Korzenik, J.R.; Fuchs, C.S.; Willett, W.C.; Richter, J.M.; Chan, A.T. A Prospective Study of Long-term Intake of Dietary Fiber and Risk of Crohn’s Disease and Ulcerative Colitis. *Gastroenterology* **2013**, *145*, 970–977. [\[CrossRef\]](#)
8. Gaudier, E.; Jarry, A.; Blottière, H.M.; De Coppet, P.; Buisine, M.P.; Aubert, J.P.; Labois, C.; Cherbut, C.; Hoebler, C. Butyrate specifically modulates MUC gene expression in intestinal epithelial goblet cells deprived of glucose. *Am. J. Physiol. Liver Physiol.* **2004**, *287*, G1168–G1174. [\[CrossRef\]](#)
9. Furusawa, Y.; Obata, Y.; Fukuda, S.; Endo, T.A.; Nakato, G.; Takahashi, D.; Nakanishi, Y.; Uetake, C.; Kato, K.; Kato, T.; et al. Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. *Nature* **2013**, *504*, 446–450. [\[CrossRef\]](#)
10. Smith, P.M.; Howitt, M.R.; Panikov, N.; Michaud, M.; Gallini, C.A.; Bohlooly-Y, M.; Glickman, J.N.; Garrett, W.S. The Microbial Metabolites, Short-Chain Fatty Acids, Regulate Colonic Treg Cell Homeostasis. *Science* **2013**, *341*, 569–573. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Imhann, F.; Vila, A.V.; Bonder, M.J.; Fu, J.; Gevers, D.; Visschedijk, M.C.; Spekhorst, L.M.; Alberts, R.; Franke, L.; Van Dullemen, H.M.; et al. Interplay of host genetics and gut microbiota underlying the onset and clinical presentation of inflammatory bowel disease. *Gut* **2018**, *67*, 108–119. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Vila, A.V.; Imhann, F.; Collij, V.; Jankipersadsing, S.A.; Gurry, T.; Mujagic, Z.; Kurilshikov, A.; Bonder, M.J.; Jiang, X.; Tigchelaar, E.F.; et al. Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Sci. Transl. Med.* **2018**, *10*, eaap8914. [\[CrossRef\]](#)
13. Franzosa, E.A.; Sirota-Madi, A.; Avila-Pacheco, J.; Fornelos, N.; Haiser, H.J.; Reinker, S.; Vatanen, T.; Hall, A.B.; Mallick, H.; McIver, L.J.; et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* **2019**, *4*, 293–305. [\[CrossRef\]](#)
14. Almeida, A.; Nayfach, S.; Boland, M.; Strozzi, F.; Beracochea, M.; Shi, Z.J.; Pollard, K.S.; Sakharova, E.; Parks, D.H.; Hugenholtz, P.; et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* **2021**, *39*, 105–114. [\[CrossRef\]](#)
15. The Integrative HMP (iHMP). Research Network Consortium the Integrative Human Microbiome Project. *Nat. Cell Biol.* **2019**, *569*, 641–648. [\[CrossRef\]](#)
16. Lloyd-Price, J.; Arze, C.; Ananthakrishnan, A.N.; Schirmer, M.; Avila-Pacheco, J.; Poon, T.W.; Andrews, E.; Ajami, N.J.; Bonham, K.S.; Brislawn, C.J.; et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nat. Cell Biol.* **2019**, *569*, 655–662. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016; ISBN 978-3-319-24277-4.
18. Caspi, R.; Billington, R.; Fulcher, C.A.; Keseler, I.M.; Kothari, A.; Krummenacker, M.; Latendresse, M.; Midford, P.E.; Ong, Q.; Ong, W.K.; et al. The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.* **2018**, *46*, D633–D639. [\[CrossRef\]](#)
19. Jeske, L.; Placzek, S.; Schomburg, I.; Chang, A.; Schomburg, D. BRENDA in 2019: A European ELIXIR core data resource. *Nucleic Acids Res.* **2019**, *47*, D542–D549. [\[CrossRef\]](#)
20. Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [\[CrossRef\]](#)
21. El-Gebali, S.; Mistry, J.; Bateman, A.; Eddy, S.R.; Luciani, A.; Potter, S.C.; Qureshi, M.; Richardson, L.J.; Salazar, G.A.; Smart, A.; et al. The Pfam protein families database in 2019. *Nucleic Acids Res.* **2018**, *47*, D427–D432. [\[CrossRef\]](#)
22. Alborzi, S.Z.; Devignes, M.-D.; Ritchie, D.W. ECDomainMiner: Discovering hidden associations between enzyme commission numbers and Pfam domains. *BMC Bioinform.* **2017**, *18*, 1–11. [\[CrossRef\]](#)
23. Eddy, S.R. Profile hidden Markov models. *Bioinformatics* **1998**, *14*, 755–763. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Katoh, K.; Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Minh, B.Q.; Schmidt, H.A.; Chernomor, O.; Schrempf, D.; Woodhams, M.D.; von Haeseler, A.; Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **2020**, *37*, 1530–1534. [\[CrossRef\]](#)
26. Buchfink, B.; Xie, C.; Huson, D.H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **2015**, *12*, 59–60. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Yilmaz, B.; Juillerat, P.; Øyås, O.; Ramon, C.; Bravo, F.D.; Franc, Y.; Fournier, N.; Michetti, P.; Mueller, C.; Geuking, M.; et al. Microbial network disturbances in relapsing refractory Crohn’s disease. *Nat. Med.* **2019**, *25*, 323–336. [\[CrossRef\]](#)
28. Lewis, J.D.; Chen, E.Z.; Baldassano, R.N.; Otley, A.R.; Griffiths, A.M.; Lee, D.; Bittinger, K.; Bailey, A.; Friedman, E.S.; Hoffmann, C.; et al. Inflammation, Antibiotics, and Diet as Environmental Stressors of the Gut Microbiome in Pediatric Crohn’s Disease. *Cell Host Microbe* **2015**, *18*, 489–500. [\[CrossRef\]](#)

29. Lopez-Siles, M.; Duncan, S.H.; Garcia-Gil, L.J.; Martinez-Medina, M. Faecalibacterium prausnitzii: From microbiology to diagnostics and prognostics. *ISME J.* **2017**, *11*, 841–852. [[CrossRef](#)] [[PubMed](#)]
30. Patterson, A.M.; Mulder, I.E.; Travis, A.J.; Lan, A.; Cerf-Bensussan, N.; Gaboriau-Routhiau, V.; Garden, K.; Logan, E.; Delday, M.I.; Coutts, A.G.P.; et al. Human Gut Symbiont Roseburia hominis Promotes and Regulates Innate Immunity. *Front. Immunol.* **2017**, *8*, 1166. [[CrossRef](#)]