Developing strain-level resolution metagenomic methods to profile the microbiome

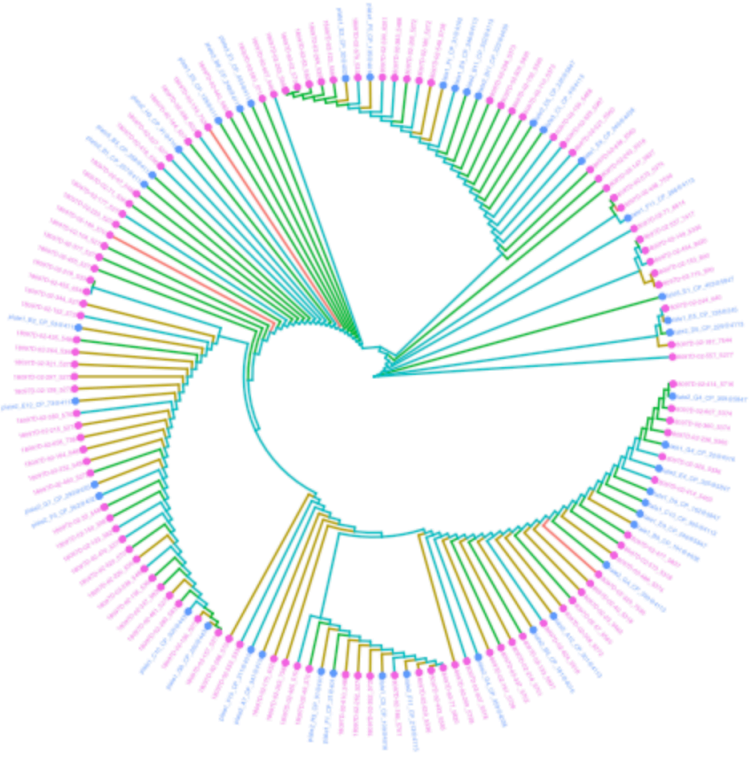
Zachariasen, Trine

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

DTU Health Tech
Department of Health Technology

Developing strain-level resolution metagenomic methods to profile the microbiome



Trine Zachariasen

PhD Thesis
August 2023

Developing strain-level resolution metagenomic methods to profile the microbiome

Trine Zachariasen

PhD Thesis

August 2023

If it was easy someone else would have done it.

from a supervisor's meeting

Preface

This PhD thesis was prepared as part of the requirements to obtain a PhD degree at the Technical University of Denmark (DTU). The work presented in this thesis was carried out in the groups of Metagenomics and Health Bioinformatics & Personal Medicine, section of Bioinformatics, institute of Health Technology, DTU. Part of the time was spent with collaborators at COPSAC in Gentofte and on a research stay with collaborators in the Knight Lab at University of California San Diego (UCSD).

The presented work was carried out from September 2020 to August 2023 under the supervision of Professor Ole Lund, Professor Anders Gorm, Associate Professor Gisle Alberg Vestergaard and co-supervision of Post Doc Asker Brejnrod.

Trine Zachariasen
Kongens Lyngby, August 2023

Publications Included in the Thesis

PAPER I

Identification of representative species-specific genes for abundance measurements

Trine Zachariasen, Anders Ø. Petersen, Asker Brejnrod, Gisle A. Vestergaard, Aron Eklund, Henrik B. Nielsen

Published in: Bioinformatics Advances, Volume: 3, Issue: 1, Year: 2023

PAPER II

MAGinator enables strain-level quantification of *de novo* MAGs

Trine Zachariasen, Jakob Russel, Charisse Petersen, Gisle A. Vestergaard, Shiraz Shah, Stuart E. Turvey, Søren J. Sørensen, Ole Lund, Jakob Stokholm, Asker Brejnrod, Jonathan Thorsen

Submitted to: Nature Biotechnology, August 2023

PAPER III

Differential responses of the gut microbiome and resistome to antibiotic exposures in infants and adults

Xuanji Li, Asker Brejnrod, Jonathan Thorsen, Trine Zachariasen, Jakob Russel, Urvish Trivedi, Gisle A. Vestergaard, Jakob Stokholm, Morten A. Rasmussen, Søren J. Sørensen

Submitted and in second review at: Nature Communications, March 2023

Summary

Microbes exist all around us and take part in shaping the world as we know it. Invisible to the naked eye, they co-inhabit all types of environmental niches and create vast and complex communities, termed *microbiomes*. They are essential for life on earth, where they play a central role in shaping the ecosystems and have a great impact on human health. The interplay of the microbes and our health is directly linked to the specific composition of the microbiome. To understand their impact it is crucial to be able to identify the finest-possible granularity, moving from identification at species-level to strain-level resolution.

By applying metagenomics this becomes possible. Metagenomics is the study of DNA extracted directly from the environment, bypassing the need for cultivation of the microbes. With this approach the entire genetic content of the microbes are analysed, enabling strain-level analysis. However, due to the complexity and variability found within the microbial world, this is a task that remains unsolved.

In this thesis efforts have been made to develop strain-level resolution metagenomic methods for accurately profiling the microbiome. In the first published work we proposed a method for selecting a set of signature genes, which can be used for accurate identification and abundance estimates of the bacteria found within the microbiomes. As the signature genes are unique for each biological entity, they can be used to profile the microbes even at very low abundance.

For the second project in this thesis, we use the signature genes in single nucleotide variant analysis, which facilitates sub-species level identification. Through this project we created the bioinformatic tool MAGinator, which enables *de novo* quantification and taxonomic annotation of the microbes found within the metagenomics sample. Through a combination of both gene- and contig-based techniques it offers insights into the genetic and functional content along with the bacterial origin.

Subsequently we explored the antimicrobial resistance gene (ARG) profiles of young adults and infants, to determine differences and

identify the specific bacteria harbouring them. The analysis revealed that bacterial composition, especially *Escherichia coli*, critically influences the ARG profile. Specific ARG clusters were identified and linked with certain strains of *Escherichia* and *Bifidobacterium*, highlighting the importance of strain-level identification.

The final project reported in this thesis investigated the spread and diversity of the opportunistic pathogen *Pseudomonas aeruginosa* across the globe. The results revealed no evolutionary differences in the genomes across different environmental niches. The metabolites produced by the microbes varied between the environments, however it remains to explore if this can also be found for the metabolites specific to *Pseudomonas aeruginosa*.

As a whole, the presented work has covered methods for strain-level analysis of the microbiome. Being able to identify strains opens a door to understand the interplay between the microbes, and also the effects that they have on the environment they occupy.

Resumé

Mikrober findes overalt omkring os og er med til at forme verden, som vi kender den. Usynlige for det blotte øje bebor de alle typer af miljøer og skaber komplekse samfund, kaldet *mikrobiomer*. De er afgørende for livet på jorden, hvor de spiller en central rolle i at forme økosystemer og har stor indflydelse på menneskers sundhed. Samspillet mellem mikroberne og vores sundhed er direkte knyttet til mikrobiomets specifikke sammensætning. For at forstå deres indvirkning er det afgørende at kunne identificere med den højeste detaljeringsgrad og gå fra identifikation på artsniveau til bakteriestammer.

Ved at anvende metagenomics bliver dette muligt. Metagenomics er studiet af DNA ekstraheret direkte fra miljøet, hvilket omgår dyrkning af mikroberne. Med denne tilgang analyseres hele det genetiske indhold af prøven, hvilket muliggør analyse på stammeniveau. På grund af kompleksiteten og variabiliteten der findes mellem mikroberne, er dette en opgave, der endnu ikke er løst.

I denne afhandling er der blevet gjort bestræbelser på at udvikle metagenomiske metoder på stammeniveau for nøjagtigt at kortlægge mikrobiomet. I det første projekt foreslår vi en metode til at vælge et sæt af signaturgener, som kan bruges til nøjagtig identifikation og kvantificering af de bakterier, der findes inden for mikrobiomerne. Da signaturgenerne er unikke for hver biologisk enhed, kan de bruges til at profilere mikroberne, selv ved meget lav tilstedeværelse.

I det næste projekt bruger vi signaturgenerne og undersøger forskellene i deres nucleotid-varianter (single nucleotide variants), hvilket muliggør identifikation på underartsniveau. Som en del af dette projekt skabte vi analyseværktøjet MAGinator, som muliggør *de novo* kvantificering og taksonomisk annotation af de mikrober, der findes i metagenomprøven. Gennem en kombination af både gen- og contig-baserede teknikker giver det indsigt i det genetiske og funktionelle indhold sammen med bakteriens oprindelse.

Derefter undersøgte vi de antimikrobielle resistensgen (ARG) profiler af unge voksne og spædbørn for at bestemme forskelle, samt identificere de specifikke bakterier, der bærer ARG. Analysen viste,

at bakteriesammensætningen, især *Escherichia coli* påvirker ARG-profilen. Specifikke ARG-clustre blev identificeret og knyttet til bestemte stammer af *Escherichia* og *Bifidobacterium*, hvilket fremhæver vigtigheden af identifikation på stammeniveau.

Det sidste projekt rapporteret i denne afhandling undersøgte spredningen og diversiteten af den opportunistiske patogen *Pseudomonas aeruginosa* over hele verden. Ingen evolutionære forskelle i genomet blev set mellem forskellige miljøer. De metabolitter, der produceres af mikroberne, varierede mellem miljøerne, men det skal stadig udforskes, om dette også er tilfædet for specifikke metabolitter produceret af *Pseudomonas aeruginosa*.

Som helhed har det præsenterede arbejde afdækket metoder til at bestemme bakteriestammer i mikrobiomer. At kunne identificere stammer åbner en dør for at forstå samspillet mellem mikroberne og også de effekter, de har på det miljø, de bebor.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisors, whose ideas and inspirational drive along with their invaluable feedback and support provided the foundation upon which this work stands. It has been a bumpy road, but I have grown with the challenges and that has led me to where I am today.

I extend my deep appreciation to my colleagues at COPSAC and the community at Copenhagen University for their cooperation, expertise, and for providing me with the necessary help and environment conducive to research.

To the members of my section, your insights, camaraderie, and shared moments of challenge and triumph have been part of shaping this journey.

To my friends and family, thank you for your consistent support and encouragement throughout this journey.

Lastly, I would like to acknowledge my own resilience, determination, and perseverance that kept me going, especially during the toughest times.

Thanks to all those people who made it possible for me to be here today.

Trine

Contents

Introduction

The human microbiome represents a diverse and complex consortium of microbial entities, that inhabits different areas of our bodies. The purpose and influence of the microbes are highly diverse, and some are performing crucial functions for us influencing our health. The human gut microbiome has been found to influence both our physiology and our immunity. The interplay between the microbes and our health are influenced by the specific composition of the microbiome. To understand the influence of the microbes, it is key to gain insights in the highest possible resolution. When going from species- to strain-level resolution we can identify detailed insights into the microbial phylogeny's, their adaptations and their unique metabolic profiles.

To explore the complex microbiomes, metagenomics can be employed. Metagenomic analysis allows us to bypass the cultivation of the microbes and investigate the DNA extracted directly from the sample. It enables the analysis of the genetic content of the complete microbial community allowing strain-level examinations. Moreover, microbial environments often display a large variability between sites and at different time points, which demands robust bioinformatics techniques to reliably being able to interpret the data generated from the microbiome studies.

The aim of this thesis was to enhance and expand the existing methods for microbiome profiling. The goal was to obtain higher resolution, even for previously unseen microbes, which is key to unlocking the potential of metagenomics to identify critical microbes for human health and environmental investigations. This facilitates precise integration of abundance, taxonomic and func-

tional annotations, empowering investigations within the microbiome field.

The thesis is structured in the following way:

Chapter 1 introduces the theoretical background for the projects comprised in this PhD. The background is divided into 3 sections: Section 1.1 covering the fundamentals of microbiomes, their ecological and physiological roles, and the significance of the gut microbiome and human health. Section 1.2 the fundamental principles of metagenomics is explained, and how it has revolutionized the study of microbial communities compared with traditional microbiological approaches. Section 1.3 briefly bridges the first two chapters and dives into the metrics used for characterizing the microbiome.

Chapter 2 covers the first scientific paper describing the use of signature genes to profile the microbiome. The method, validated with both simulated and real data, demonstrates that signature genes enhance species identification and improve abundance estimation.

Chapter 3 is based on the second paper, introducing the tool MAGinator. The aim of MAGinator is to achieve *de novo* subspecies level resolution in microbiome studies, enabling precise integration of abundance estimates and taxonomic- and functional annotation.

Chapter 4 briefly discuss the context of antibiotic resistance and the role of the gut microbiome in this issue. This is elaborated upon in the third paper, where we studied the variations in antibiotic resistance from infancy to adulthood.

Chapter 5 introducing a fourth project, which is ongoing. The project concerns the environmental spread of *Pseudomonas aeruginosa* and its associated metabolomic profile, with the hypothesis that the genomic variability increase in host-associated strains.

Chapter 6 concludes the thesis with an epilogue, discussing the key points from the presented work as well as future perspective.

CHAPTER 1

Theoretical Background

1.1 Microbiome

Microbes are ubiquitous and are found all around us. Despite their size, they play an important part in shaping ecosystems, influencing human health, driving biochemical processes and are essential for life on earth. The term *microbiome* covers the microbial community inhabiting a specific environment, including different microorganisms such as bacteria, fungi, and virus.

1.1.1 Human microbiome

The human microbiome, also known as the *microbiota* [1], is diverse and each body site harbours a distinct signature of microbes. The microbes inhabiting our body outnumbers our human cells with a factor 10, comprising 10-100 trillion microbes [2]. They live almost everywhere on our body, but most abundant and assorted is the microbiome found in our gut [3]. Collectively the microbiota comprises about 3.3 million genes, another number that vastly overshadows the 22,000 protein-coding genes found in the human genome. Additionally, the human genomes are 99.9% similar, where the difference between microbiomes has been found to be up to 80-90% between individuals. This highlights the huge diversity within our microbial inhabitants and underlines how,

through symbiosis, they provide traits that humans have not had to evolve on their own [1].

Human Gut Microbiome

The intricate ecosystem of the human gut microbiome, residing in the gastrointestinal tract, hosts thousands of microbial species. It is constituted predominantly by bacteria, which dependent on the health status and age of the host is dominated by different taxonomic groups. Other microbes such as archaea, fungi, viruses and protozoa also plays a large role in shaping the gut flora and together they aid the host by food digestion, e.g. by breaking down otherwise indigestible dietary polysaccharides. The metabolites produced by the microbes, such as short-chain fatty acids, are essential for maintaining a healthy gut and contribute to immune development and modulation [1, 4].

The human gut microbiome is shaped already during the first hours of our lives and is highly influenced by the mode of delivery. Once the microbiome has been formed it has a strong signature throughout the rest of our lifespan. It has been found to be influenced by host factors, such as genotype, lifestyle (including diet) and physiological status (such as aging) [5].

Impact on Human Health

The human gut microbes play a essential role in human health [1, 3, 6, 7]. When aiding with digestion, they alter the nutritional gain from the food, e.g., by degrading complex carbohydrates. They also aid by producing vitamins and facilitates the host with absorption of minerals. Additionally, the microbes has been found to play an important role in the energy metabolism of the host and influence the storage of fat, thus directly linking the microbiome with metabolic dysfunctions, such as diabetes and obesity [3].

Additionally, the microbial signature has been found to have an impact on a great variety of diseases, spanning from immune-related diseases, including allergies and chronic inflammation to mental disorders, such as autism and depression [3].

Whether the microbiome is the cause or the consequence of the diseases are still being investigated [8]. For some diseases, such

as Inflammatory Bowel Disease (IBD) a shift in the microbiome can be seen before any symptoms appear, hence the disease is a consequence of the microbiome [4]. Other physiological factors, such as a shift in temperature or pH can lead to improved conditions for certain microbes leading to an altered microbiome. And yet for some complex diseases, such as asthma, the intestinal microbes seem to act as an environmental factor, which is only one of many factors contributing to the disease status [6]. Other confounding factors such as genetics, pH, and nutrient availability are also highly important to consider when estimating the impacts of the microbiome on human health.

The microbes constituting the gut microbiome has been found to be of great importance for human health, however the specific mechanisms and processes for the systematic diseases are still largely unknown [8]. Though for gastrointestinal disorders it is more straightforward to identify the responsible pathogen, such as *Campylobacter jejuni* or *Salmonella* which are known to cause food poisoning.

1.1.2 Resistome

The term *resistome* comprises all the Antibiotic Resistance Genes (ARGs) found in an environment, covering genes originating from pathogenic and non-pathogenic bacteria and are thus interchangeably linked to the microbes that lives in the environment [9].

Antibiotic resistance initially emerged as part of the inherent defense mechanisms of bacteria. Yet, as medicine science advanced, antibiotics have been used to treat a range of bacterial infections, including those responsible for food poisoning. ARGs can provide resistance towards one or more types of antibiotics and can either be intrinsic to the bacteria or be acquired by Horizontal Gene Transfer (HGT) [10–12]. External environmental factors have previously been shown to play a large role in the spread of antibiotic resistance including pollution, inadequate sanitation, inappropriate waste disposal. And significantly the misuse of antibiotics in medicine, agriculture and livestock production has accelerated the spread [13].

This spread causes a large concern for the public health, as multi-resistance emerges in pathogens and can be extremely difficult to treat and control [9, 10].

1.1.3 Tools for Microbial Analysis

The human microbiome has been investigated since the 17th century, when initial observations revealed that there was a difference between the microbiomes in healthy and sick individuals [3]. Though the study of the microbiome is not a new invention, the bioinformatics methods used for analysing the microbes are.

Traditional methods were cultivation- dependent and allowed only to study one or a few bacteria at the time [14]. This was a limiting factor in the study of microbiomes, as many bacteria are 1) not cultivable under standard laboratory conditions and 2) only growing in concert with other specific microbes. With the advance of culture-independent techniques that can analyse the microbiome as a whole has led to better comprehension and more unbiased insights into the microbes and their interactions.

One of such methods is metagenomics, which opens the door for investigating all the DNA present in a microbiome sample.

1.2 Metagenomics

By understanding the complex landscape of the microbiome and its close connection to human health, the need for techniques that can accurately analyze it becomes apparent. One such powerful method is *metagenomics*. With metagenomics direct analysis of all genetic material from an environmental sample is analysed, without the need for cultivation or identification of the organisms. This approach offers an unprecedented window into the microbial world, providing insights into community structure, functional capabilities, and dynamic interactions.

Two main categories of metagenomics exist: targeted 16S rRNA sequencing and untargeted shotgun metagenomic sequencing [15]. In this PhD thesis, the method of analysis used is shotgun sequencing. Thus, for the purposes of this thesis, 'metagenomics' will refer specifically to that method.

1.2.1 Shotgun Sequencing & Data Characteristics

With shotgun metagenomic sequencing, short reads are generated by random sampling of DNA from the sample. Thereby offering the potential for the highest taxonomic identification and functional characterization, as no data must be filtered out in advance.

With each sequencing run producing up to billions of reads, shotgun sequencing generates significant amounts of data. Despite the large volumes it is still not certain that all parts of the microbiome are represented in the sequenced reads. The proportionality of the different constituents of the environment is often highly skewed, leading to sparse, overdispersed and heterogeneous data [5, 7, 15]. However, the sparsity can arise for two reasons, by either technical errors or biological variations. A technical zero could be caused if the microbe is present in the environment, but not present in the data due to low sequencing depth or sampling imbalance. A biological zero would be if the microbe is not present in the environment being sequenced [16].

For metagenomics studies it is relevant to be able to compare the microbiomes across samples. Given the inherent variability in sequence quantity among these samples, normalization is often

employed to adjust the read counts. This process often involves scaling the reads relative to the total sum of reads within each sample, thereby generating a relative and compositional measure that facilitates comparison [15, 16]. This can be done for various features, such as species or genes predicted in the samples, which can be stored with the counts of the feature in each sample. As the number of reads in each sample are dependent on the sequencer, the count constraint leads to strong dependencies regarding the abundance of the features in the sample, e.g. if the abundance of one species increase, this implies a decrease for another species as the total number of reads in the sample are fixed [15, 17].

1.2.2 Metagenomic Assembly

To gain information from the reads, it is relevant to piece together the shorter sequences into longer contiguous fragments (contigs). Assembly is carried out to regenerate the original genomic sequences of the microbes found in the sample. This can be challenging, as some parts of the genomes are very similar or repetitive and can be difficult to distinguish [18].

Different approaches to assembly exist, however the most common method is using the de Bruijn graph [19]. The reads are broken into smaller fragments of length k, termed k-mers. The overlap of the k-mers are found and linked in a graph, where the paths in the graph represent the tentative contigs. Depending on the data the optimal value of k varies. Smaller values will make the graph more tangled, and it will be hard to determine the optimal path through it. Larger values can erroneously miss overlaps between the reads, especially in areas with low coverage, making the graph more fragmented [19], however it will help with distinguishing very similar genomes, such as strains. Thus, the optimal approach would be to have a high k for high coverage regions and a low k for low coverage regions.

One way to accommodate this trade-off is to use multiple k-mer values, thus accommodating the complexity of metagenomics samples. The state of the art within the metagenomics assembly field include tools such as metaSPAdes[20] and MEGAHIT [21].

An example of a de Bruijn graphs i seen in Figure 1.1. As seen from the figure, de Bruijn graphs of a metagenomics sample can be complex and not straight-forward to interpret, leading to potential misassemblies. This occurs due to the presence of repeat regions, strain variation, uneven coverage across the genomes or errors in the input reads [19].

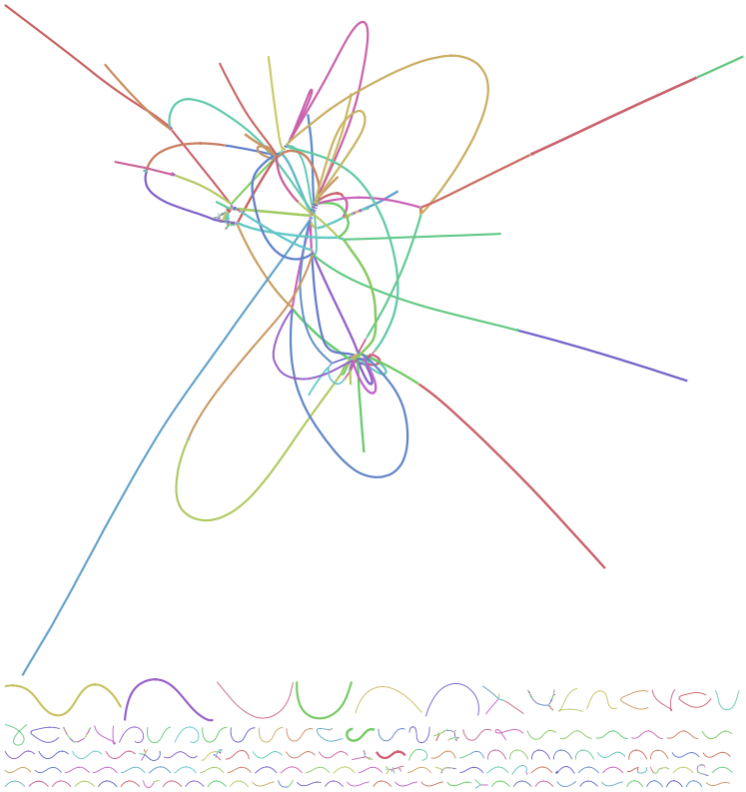


Figure 1.1: Assembly graph generated by metaSPAdes [20], showing the contigs of a metagenomic sample. Each node represents a contig and each edge represent an overlap between the contigs. Only the largest contigs are shown.

1.2.3 Metagenomic Binning

The contigs can be grouped together by their genome of origin, a process called *binning*. Each bin represents a microbial genome, termed a Metagenomic Assembled Genome (MAG). The contigs

are grouped based on shared characteristics found in the contigs. These characteristics include sequence composition, coverage across samples [22, 23] or identification of phylogenetic markers [24].

In general the metagenomics binning approaches can be divided into two groups, supervised or unsupervised. For the supervised methods already known information is used to guide the binning process, such as reference genomes or phylogenetic marker genes. The disadvantage of supervised binning approaches is that the results are limited to the information you already have in the references. The process is highly accurate in the cases where references are already available, however when investigating novel entities with no close relatives found in the references, the results are often of poor quality. Additionally these approaches also gives ambiguous results, in cases where closely related organisms have very similar genomes as they can be hard or impossible to distinguish [24, 25].

The unsupervised methods are based on complex mathematical models, leveraging the inherent information found in the data. They can group the contigs based on the sequence composition and coverage patterns [14]. The sequence composition is important as contigs belonging to the same microbial genome will display somewhat identical nucleotide frequency. Additionally read coverage of the contigs, are used to support the binning process, as genetic components originating from the same organism will have approximately the same abundance [20]. Multi-sample approaches have gained dominance, as it includes the strengths of co-abundance patterns amongst contigs and reads across samples. MaxBin2 showed that the difference in binning two samples individually yielded 19 and 26 bins, whereas co-abundance binning of the two yielded 84 bins [26]. Despite only having two samples, MaxBin was able to identify more than twice as many bins by using the co-assembly approach.

Another variant of the multi-sample approach is integrated by VAMB [23]. It employs a neural network in the form of a variable autoencoder, which learns the complex and high-dimensional structure of the data, through the sequence composition and co-

abundances. This is used for clustering of the contigs into bins. VAMB creates a bin for each sample in which the MAG is present, which can be combined across samples into a cluster. That way it is possible to identify even small differences between the MAGs of the samples [23].

The complexity, variation and sparsity of the metagenomics data challenges the precision and accuracy of the binning process. Despite the challenges, metagenomics binning provides the most detailed insight into the individual microbial inhabitants of the communities we are examining.

1.3 Characterizing the Microbiome

Understanding the complexities of a microbial environment requires more than being able to identify its microbial inhabitants. It also covers the functional profiles of the microbes, their interactions and dynamics affecting each other and their potential host. When characterizing the microbiome in high resolution, it covers the microbial diversity, their associated abundances, and functions.

Diversity Metrics

The diversity describes the variation of different microbes found in the environment. The diversity is traditionally divided into two categories, alpha- and beta diversity. Alpha-diversity is a measure of the diversity within a sample, where beta-diversity gives the difference in diversity between samples [17]. Beta diversity is calculated as the compositional dissimilarity between the samples, providing a measure of ecological distance. Beta diversity can be used to describe the diversity between samples from different environments or the same environment over time.

A higher diversity is often associated with better health status [27]. A more diverse microbiome implies a larger potential for functions and gives a higher resilience towards environmental changes and pressures.

Phylogenetic & taxonomic profiling

A method for characterizing the microbes found in the environment is by phylogenetic profiling. This covers the systematic arrangement of species based on their evolutionary relationships, determining how closely related certain species or groups are within a given community [28].

In phylogenetic analysis, the selection of appropriate genes or genetic regions for comparison is paramount, as factors like mutation rates and horizontal gene transfer can influence the observed relationships. While the accessory genes may provide nuanced insights for closely related strains, broader comparisons across diverse bacteria necessitate the examination of conserved marker

genes. One very conserved gene, the 16S rRNA gene, is commonly used for phylogenetic classification, due to its universal presence in prokaryotes [22]. As the 16s rRNA gene is extensively investigated and characterized with reference databases it can be used for taxonomic profiling. Taxonomic profiling assigns organisms to various taxonomic ranks (e.g., species, genus, family). Other genes can also be used for taxonomic assignments. A tool created specifically for annotation of metagenomic samples is GTDB-tk [29], which uses a combination of 120 bacterial and 100 archaeal marker genes for taxonomic classification.

Functional profiling

Beyond the taxonomic characterization of the microbial community another important aspect is the functions comprised by the microbes. From the genes found in the sample, the functions can be predicted by comparing the sequences with genes of known functions, with tools such as eggNOG-mapper [30]. This reveals the metabolic pathways and metabolic pathways that are present within the sample. This enables comparisons between samples with different conditions or environments, which can reveal metabolic functions which are up- or down-regulated. Various diseases has been examined by examining their microbiomes against healthy controls, enabling associations between metabolites and microbes and their impact on human health [4, 6, 31].

CHAPTER 2

Signature Genes used for Microbiome Profiling

Despite the advances within the metagenomics field, the task of accurately profiling the microbiome is still unresolved [14]. This is influenced by several factors, both technical and biological. The strains that we try to separate have very similar genetic composition and can therefore be hard to distinguish, even in cases with high abundance. With traditional reference-based metagenomic quantification methods, reads may align perfectly to more than one species, leading to misclassification or crossmapping [18]. Additionally, the biological understanding of strains is still developing. Previously it was believed that one strain would out-compete other similar strains and that only one would be present within a sample. This is to a large extend the case for some species such as *Escherichia coli* [12], however for other, such as *Bifidobacterium longum* we see, that multiple subspecies can coexist in the microbiome (described in PAPER II).

A way to overcome these problems is to select a set of representative genes for each microbial entity, termed signature genes, which can be used as markers. The genes have to be unique for the entity, and at the same time found within all members. If reads are present within the sample and mapping to the signature genes, the

entity is present within the sample. This facilitates quantification of species even at very low abundance.

PAPER I

In PAPER I we propose a method for identifying a set of species-specific genes, found *de novo* for that particular dataset. These representative genes can be used for identification of the microbes as well as for abundance estimations. The genes are found using a negative binomial model, ranking the genes across the samples. The signature gene set is evaluated according to how many reads that maps to the entity and how many signature genes we detect within the samples. Based on these metrics the signature gene set is iteratively improved by switching the worst performing genes. For each signature gene set a variant of the 'coupon collector's problem' is applied to calculate the probability that the full gene set is present within the samples, given the number of reads that maps to the signature genes. This process leaves a set of signature genes, which can be used for precise detection and more reliable abundance estimations of the microbial entities. The method is validated using a simulated dataset, which has been binned using MSPminer [32] and Bäkhed's First Year of Life dataset [33], which has been binned using VAMB. MSPminer creates the bins based on genes and VAMB based on contigs, which shows the flexibility of the method.

Another set of challenges within microbiome profiling is the technical complications. The metagenomics data displays bias between the genetic pool being selected and the actual genetic output from the sequencer, which can occur especially at the PCR-amplification or at the sequencing steps. The latter occurs is especially due to extreme values of GC-content [34].

Selecting the initial Signature Gene Set

When interested in identifying a set of signature genes for each microbial entity it is relevant to take these bias into consideration.

A practical solution to the problem is to initially filter the genes found within the biological entity according to their detection rate across the samples (Figure 2.1). The genes are initially sorted according their fit to the median gene abundance profile within the species (Figure 2.1A). If they display a skewed frequency of detection, the genes are reordered, so they are found more consistently and frequently across the samples (Figure 2.1B).

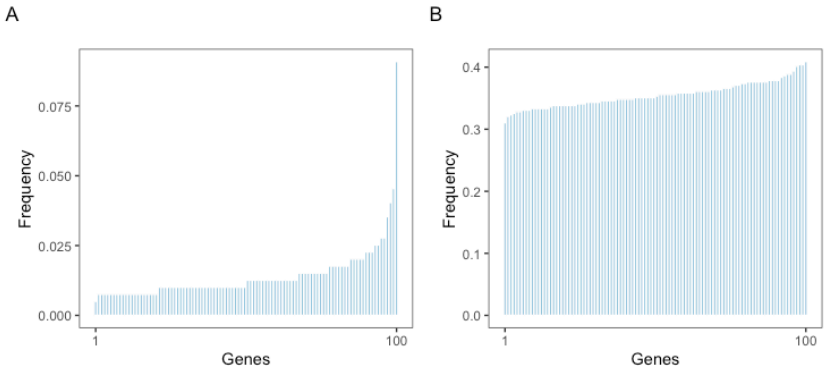


Figure 2.1: The frequency of the 100 initial genes for *Dialister sp.* identified in the First Year of Life data set from Bäkhed et al. A) Sorted according to median gene abundance profile and B) Sorted according to gene frequency across samples.

Refining the Signature Gene Set

In addition to the filtering across samples, the signature gene sets can be refined by modelling the read counts using the negative binomial (NB) distribution. Each gene is evaluated using the NB distribution to test whether an increase in sequencing depth reliably results in an increase in the counts of that gene. It has previously been described how gene counts follow a NB model [5] as it allows overdispersion, which is often seen in shotgun data [16]. The model is applied for each sample, where the read count of gene i in the jth sample is denoted y_{ij}, then

$$\begin{aligned} y_{ij} &\sim NB\left(y_{ij}|\mu_j, \sigma_j\right) \\ &= \frac{\Gamma\left(y_{ij}+\sigma_j\right)}{\Gamma\left(\sigma_j\right) y_{ij}!}\left(\frac{\sigma_j}{\mu_j+\sigma}\right)^{\sigma_j}\left(\frac{\mu_j}{\mu_j+\sigma_j}\right)^{y_{ij}}, \ \mu_j > 0 \end{aligned} \tag{2.1}$$

where μ_j is the average read count per gene, σ_j is the sample-specific dispersion parameter and $\Gamma(\cdot)$ is the gamma function (extension of the factorial function). Under this parameterization of the negative binomial model, the expected read count is denoted as $E[y_{ij}] = \mu_j = \lambda_{ij} N_j$, where λ_{ij} is the proportion of reads mapped to gene i in the jth sample, and N_j is the total number of reads mapped to sample j. Therefore, μ_j depends on both the sequencing depth and the abundance of the species in the sample. The variance of the read count is given by $\text{var}(y_{ij}) = \mu_j + \mu_j^2/\sigma_j$. The counts of each signature gene are evaluated based on this NB model and ranked within each sample by comparing the difference between the expected and observed count. The NB model is thus enabling us to rank the signature genes, leaving us with the possibility of changing the worst-performing genes.

Evaluating the signature gene sets

To estimate whether the switch of signature genes leaves a gene set, which is better for profiling the microbiome, the deviance between the actual number of detected signature genes versus the expected are calculated across samples.

Given the size of the signature gene set as n genes, we can calculate the number of expected signature genes, that has reads that map within a sample, d, by the number of reads mapping to that sample, k_j, as

$$d_j = (1 - (\frac{n-1}{n})^{k_j})n, \;\; j = 1, 2, .., m \tag{2.2}$$

where m is the number of samples and n has been set to 100 genes. The function is visualized Figure 2.3 A+B indicated by a blue line.

The optimal size of the signature gene set, n, was found by testing different gene set sizes in the range from 70-150 genes. Using the data from PAPER I the performance was estimated by comparing the relative error from the true abundance of the simulated gene set with the predicted abundances stemming from these different gene set sizes (Figure 2.2). From these tests we find a local minimum around $n = 100$. Additionally, it is seen, that the error is smaller for the refined signature gene sets, indicating that it is more suitable for abundance estimations.

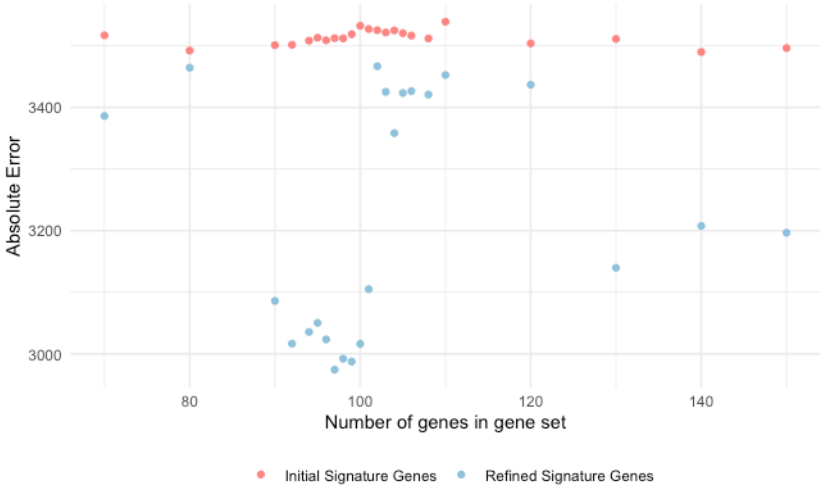


Figure 2.2: Using the data from PAPER I, different sizes of the signature gene set is evaluated. The absolute error between the true and calculated relative abundance are found with the different sizes of signature gene sets. This error is shown for the initial/filtered and refined signature gene sets.

In PAPER I a variant of the Coupon Collector's Problem (CCP) [35] is applied to estimate the likelihood of sequence reads that maps to a certain number of signature genes d, in relation to the quantity of reads k that correspond to the entire gene set. With this metric the gene set are evaluated within sample and the chance that the full gene set is present within the sample can be calculated. As we are interested in identifying signature gene sets, where all genes are present within all samples, this is a valuable metric to compare different gene sets.

Performance in simulated and real data

We evaluated the performance of the species-specific signature gene sets using two different data sets; A simulated gene catalogue, created by Borderes et al. [36] and the First Year of Life data set created by Bäkhed et al. [33].

Benchmarking of the method is possible using the simulated data set, as the predicted profiling of the environments can be compared with the truth. We compared the signature genes with the results from MSPminer, where we saw a significant improvement in how

well the signature genes followed the expected distribution across species Wilcoxon signed rank test (p-value of 4.0×10^6, paired). An example of the improvement of the signature gene set is seen for one of the Metagenomic Species Pangenome (MSP) predicted using MSPminer in Figure 2.3, where the refined signature gene set follows the expected distribution more closely leading to more samples being accepted by CCP (p-value < 0.05).

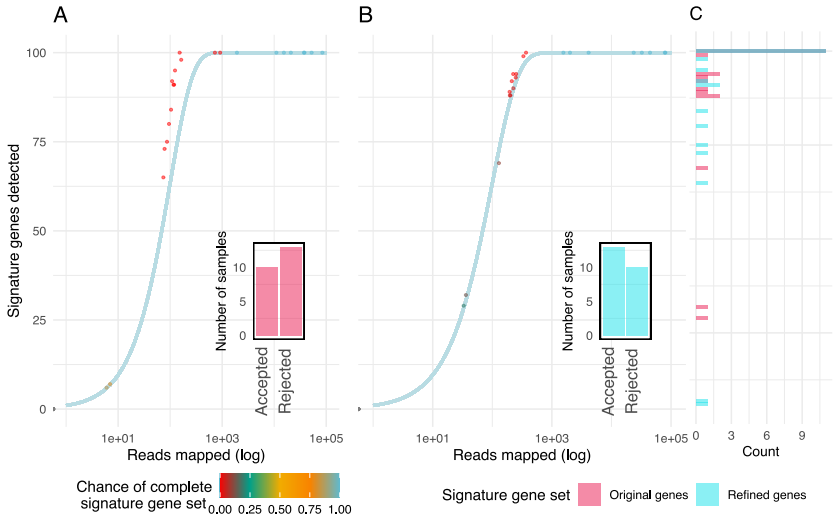


Figure 2.3: Major insights using the simulated data set in PAPER I. The detection of signature genes is displayed for each sample. The number of identified signature genes by the number of reads mapped to signature genes of MSP54. Colors indicate the chance of this sample containing 95 unique signature genes as described in methods. The bar plot indicates the number of samples that were rejected (P<0.05, CCP) and accepted (P>0.05, CCP). The expected distribution of samples for a metagenomic entity which contains 100 signature genes is indicated by a blue line. A) signature genes prior to refinement and B) after signature gene refinement. C) Accepted and rejected samples by CCP for the two gene sets. Figure adapted from PAPER I.

Another key finding in PAPER I was through the First Year of Life cohort. Using a combination of binning with VAMB and the signature genes we were able to identify 1843 MAG clusters, compared with the original study's 373 meta-Operational Taxonomic Units (mOTUs), yielding a more fine grained resolution of the microbiome. Additionally we were able to reproduce the results of

the original study, where we were able to identify the same presence/absence patterns of their *Signature Taxa*, for the taxonomic ranks where we identified the exact same taxonomies.

Conclusion

To summarize, in PAPER I we presented a method for *de novo* identification of signature genes for a dataset enabling more precise species-identification as well as improved abundance estimations. We successfully implemented method for both simulated and real data.

CHAPTER 3

Strain-level profiling of the microbiome

We know that the species composition of the microbiome alone does not explain the complex mechanisms and processes found in the microbial environment. For a more comprehensive understanding of the microbiome and its phylogenetic and functional relationships it can be relevant to dive into more detail with subspecies level resolution, including host associations of genes, leading to a broader understanding their metabolic fingerprints [36].

When diving into the microbiome at a deeper level than species, it gives us the ability to characterize specific strains or subspecies and link them to unique functionalities. This can be exemplified by *Bifidobacterium longum*, where a specific subspecies, *Bifidobacterium longum* subspecies *infantis* (*B. infantis*), has been shown to be able to breakdown specific types of human milk oligosaccharides [37, 38], which is the main energy source for breastfed human infants [39].

Being able to pair the microbial entities with their associated genes is an essential part of being able to obtain a comprehensive understanding of the community. This can be obtained by integrating contig and gene information. Binned contigs gives information about the genomic structure and the genes gives insights into the functional capabilities of the microbes. This can e.g., be used if the presence of specific genes within a contig indicates functional

capabilities of a microbe, such as its ability to metabolize certain substances or its antibiotic resistance.

PAPER II

In PAPER II we introduce our tool MAGinator, which is created to delve into the fine-scale biological differences within MAGs by the use of signature genes. MAGinator is a workflow, which processes the reads, contigs and bins of a metagenomics dataset. The key features of the tool is its capacity to identify subspecies-level microbes found *de novo* for the data set and additionally providing the user with relative abundance profiles, SNV-level phylogenetic trees and synteny clusters. To achieve this information both genome- and gene-based methods are combined, allowing us to determine the origin of the genes. Consequently, the functional profile can be predicted and associated with its host organism.

The strengths of MAGinator is validated using simulated data originating from the Critical Assessment of Metagenome Interpretation (CAMI) [40], benchmarked using data from a case-control study designed by Franzosa et al [4] and used for exploratory analysis of two infant cohorts from COPSAC_{2010} [6] and CHILD [41].

MAGinator is available at GitHub https://github.com/Russel88/MAGinator.

COPSAC_{2010} and data preparation

To illustrate the features and analysis created by MAGinator, the following sections will be based on the results of MAGinator run on the COPSAC_{2010} cohort [6, 9, 42].

The data consists of 662 samples collected from 1-year old infants. The data has been preprocessed, assembled with metaSPAdes [20] and binned using VAMB [23]. MAGinator has been run using default settings on the data set identifying 880 MAG clusters.

Strain tracking

With the context of the $COPSAC_{2010}$ cohort, we used MAGinator to obtain strain level resolution. We investigated this in more detail for the subspecies of *Bifidobacterium longum*, where we are particularly interested in being able to separate the two subspecies *B. infantis* from *Bifidobacterium longum* subsp. *longum* (*B. longum*) due to their different metabolic capabilities. Within the MAGinator framework 12 MAG clusters was identified and annotated within the *Bifidobacterium* genus including one MAG cluster for the subspecies *B. infantis* and one for *B. longum*. To benchmark the performance MetaPhlAn [43] was also run on the data, which produced a single abundance measure for the species *Bifidobacterium longum*. We summed the abundances of our *B. infantis* and *B. longum* clusters and compare with the abundance from the MetaPhlAn cluster and found that 87% of the variation was explained (PAPER II, Suppl. Figure 4). This indicates a higher level of stratification using the results from MAGinator.

Additionally, we analysed the samples using StrainPhlAn [31], which detects stains with predefined marker genes. Two clusters was identified, which correlates with the relative abundance of the two *Bifidobacterium longum* subspecies (Figure 3.1). The two StrainPhlAn clusters are mutually exclusive and are thus only able to identify one of the clusters in each sample. From Figure 3.1A we see that MAGinator is able to detect the subspecies even in samples with low abundance. This illustrates how MAGinator with *de novo* identification of MAG clusters and subsequent identification of signature genes enables better stratification of the microbiome.

Reusing the signature genes

The signature genes have been found *de novo* for the deeply sequenced COPSAC cohort. But in the case of having a shallower sequenced data set, would we be able to reuse the signature genes?

A subset of the CHILD cohort consisting of 2846 shallow-sequenced samples from infants was included in the analysis. We mapped the reads to the non-redundant gene catalogue found from the $COPSAC_{2010}$ cohort, yielding the read counts of the signature genes. The read mappings of the two cohorts to *B. infantis* is seen in Figure 3.2A, where it is clear that the strains from the COP-

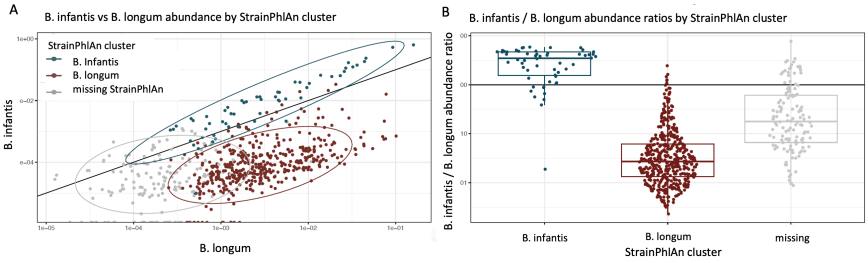


Figure 3.1: Stratification of StrainPhlAn and MAGinator clusters for COPSAC_{2010} using data from PAPER II. A) The relative abundance of the subspecies identified using MAGinator colored by StrainPhlAn cluster B) Ratio of relative abundance of MAGinator subspecies displayed for the StrainPhlAn clusters. Figure adapted from PAPER II.

SAC cohort follows the expected distribution (Equation 2.2) more closely (MSE=103.95 for COPSAC compared to MSE=878.09 for CHILD). A large subpopulation of the CHILD samples never reach more than 50 detected signature genes, despite having a large amount of reads that map to the MAG cluster. This indicates that part of the signature gene set are not found in the strain seen in the CHILD cohort, which can also be seen from the heatmap in Figure 3.2B.

Despite the cohorts having a large resemblance we see that the signature genes are not as specific for the strains found in the CHILD cohort. It is thus preferred to run MAGinator and find a relevant set of signature genes for the data set in question.

SNV-level phylogenetic trees

A method for elucidating genome-variations is by identifying Single Nucleotide Variants (SNVs). When applied to closely related genomes small differences are captured and they will be able to be distinguished. SNVs can be used for inferring the phylogenetic relationship between the samples, thus illuminating evolutionary relations [28]. SNV-profiles of marker genes have previously been shown to successfully divide strains from different environments[22] or conditions [31].

SNVs can be found for the signature genes, which can be used to infer the phylogeny, elucidating the smaller biological differences

found within the MAG clusters. An alignment for each signature gene is made for the samples that contain the signature genes. The clades of the tree can be associated with metadata to reach strain-level differences. As the SNVs are found based on the sequences of the signature genes, this allows placement of samples in the tree of the MAG cluster even when no MAG was found in the sample. This is illustrated by *Faecalibacterium Faecalibacterium* identified in the COPSAC cohort (Figure 3.3). The MAG is identified in 85 samples, and 13 additional samples are placed in the tree.

Gene synteny

Gene synteny refers to the physical co-localization of genes on a chromosome and are thus genes located adjacent to each other. Genes found in synteny, referred to as synteny clusters, can be used to provide a deeper understanding of the genomic organization of the genes and help us gain insights into their shared

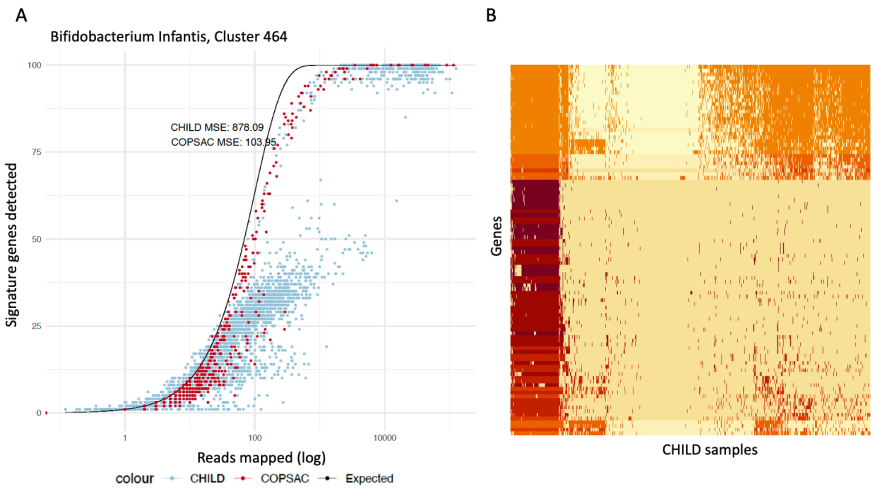


Figure 3.2: Reuse of the signature genes identified from the COPSAC_{2010} on the CHILD cohort from PAPER II. Read mappings of *B. infantis* signature genes. A) The number of reads mapped to the signature genes presented with the number of signature genes detected. Each dot is a sample. The red colour indicates COPSAC_{2010} samples, the blue color indicates CHILD samples. The black line indicates the expected distribution (Equation 2.2). B) Heatmap of the read mappings to *B. infantis* signature genes. Figure adapted from PAPER II.

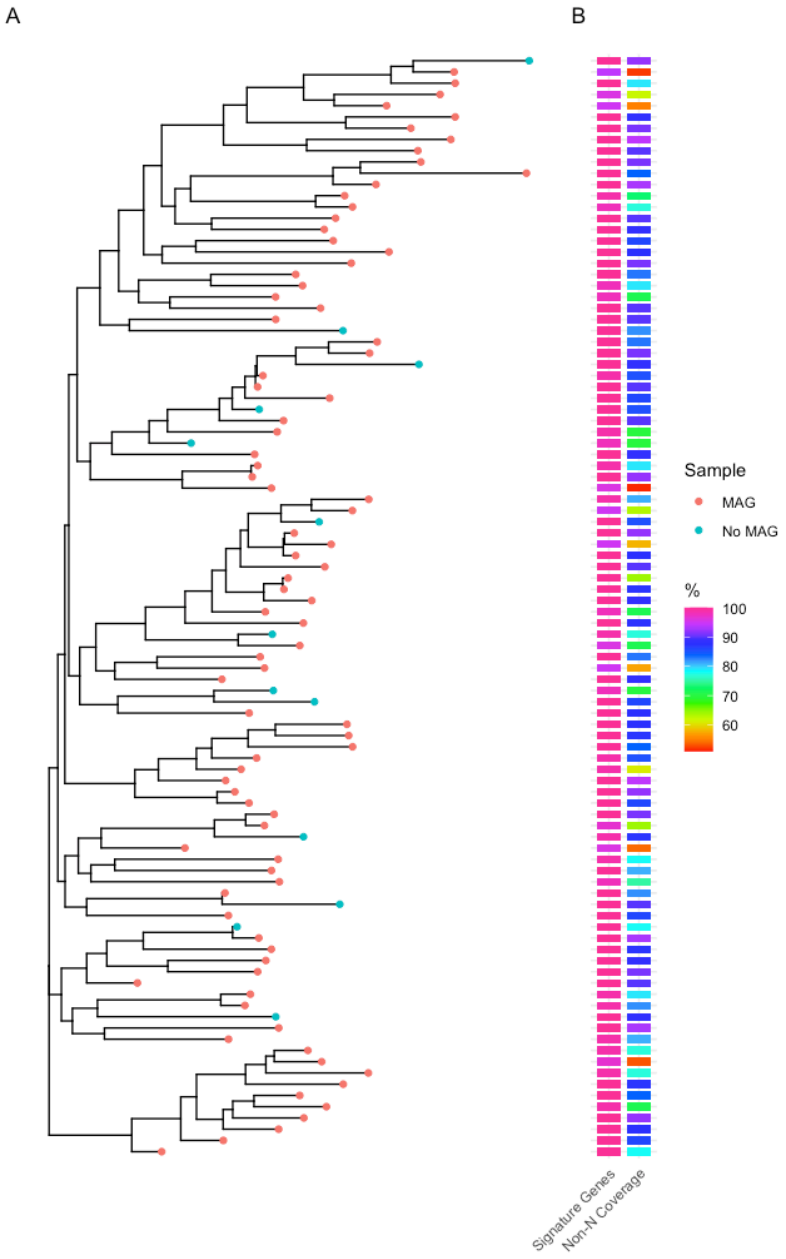


Figure 3.3: Phylogenetic tree created with output generated by MAGinator using the $COPSAC_{2010}$ data from PAPER II. A) SNV-level phylogenetic tree based on the signature gene of the MAG cluster *Faecalibacterium Faecalibacterium* sp900758465. The tip color indicates whether the sample contains a MAG. B) Heatmap of number of signature genes detected in the sample and proportion of signature gene sequence covered by read mappings in the alignment (%). Figure adapted from PAPER II.

pathways [44].
MAGinator has been developed to identify synteny clusters. This is done by creating a weighted graph of the adjacency of the genes on the contigs. If the genes are close enough in the graph, they will be categorized to be part of the same synteny cluster. The clustering of the graph is done using mcl-clustering [45] and only immediate adjacency is used. As genes found in the same synteny cluster is believed to be part of the same metabolic pathway [46], the synteny clusters predicted by MAGinator have been evaluated by examining the functional annotation of the genes found in synteny. MAGinator was run on the $COPSAC_{2010}$ cohort, producing in 746,251 synteny clusters with an average of 3 genes per cluster (Figure 3.4 B). The predicted synteny clusters were functionally annotated using eggNOG mapper [30, 47, 48]. For each cluster the KEGG module [49] with the highest occurrence was identified and the proportion of the genes within the cluster with this annotation was calculated (Figure 3.4 A+C). 92.8% of the clusters was found to have over 80% agreement in assigned KEGG module, indicating that the genes of the synteny clusters are part of the same metabolic pathway.

Software development
While the innovative capabilities of MAGinator are undeniably its core strength, its design for reproducibility and user-friendliness ensures that it stands out as an asset in the microbiome research toolkit. As the field continues to grow, tools like MAGinator, which prioritize both scientific content and user experience, will be instrumental in driving forward our understanding of complex microbial communities.

The software has been setup as a Python module and based on a set of Snakemake [50] workflows. The dependencies of running MAGinator is mamba [51] and Snakemake, the rest of the dependencies are installed automatically once MAGinator is run. Additionally, the database for GTDB-tk [29] has to be downloaded for taxonomic annotation.

MAGinator has been developed, so that it can be run on a server or a compute cluster systems such as qsub (torque) and sbatch (slurm). Additionally, we have implemented the workflow to be

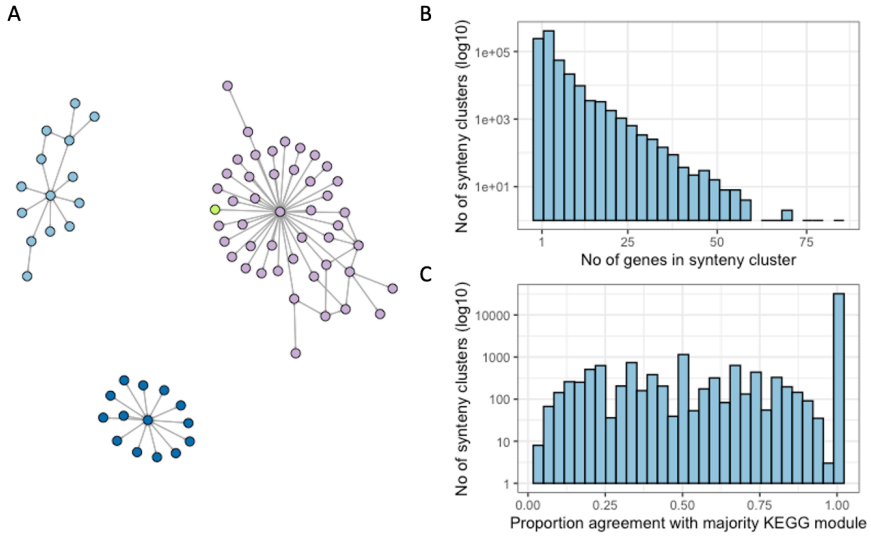


Figure 3.4: Central result from PAPER II. Synteny clusters and associated functional annotation created with output generated by MAGinator using the $COPSAC_{2010}$ data. A) Graph network of 3 synteny clusters. The colors represent KEGG modules. Green indicates that no KEGG module was annotated B) Distribution of synteny cluster size C) Proportion of genes annotated with the most common KEGG module in the cluster. Only clusters of >=5 genes are included. Figure adapted from PAPER II.

as versatile as possible, allowing for the user to input parameters for the different tools and analyses. This ensures that MAGinator can be tailored to address specific scientific questions.

Conclusion

PAPER II presents the tool MAGinator, which is a freely accessible tool, designed to obtain *de novo* strain-level resolution of metagenomics shotgun data sets. It provides precise abundance estimates, even in samples containing the microbe in low abundance. MAGinator combines information from gene- and contig-based methods, enabling merge of information about taxonomic profiles and the origin of the genes and genetic content, which can be used for functional understanding of the organisms found within the samples.

We have tested it on several data sets, including the $COPSAC_{2010}$ cohort. This covered 880 high quality MAG clusters, for which we

have identified signature genes. We have shown that the signature genes can be used as a basis for subspecies-level analysis, providing information regarding their functionality, their internal relatedness between the samples.

CHAPTER 4

Resistome of different age groups

As described in section 1.1.2, the human gut microbiome contains a large reservoir of ARGs. They play a crucial role in the response to pathogens and antibiotics and thereby on human health. The ARG profile found in the microbiome is highly influenced by the bacterial composition, as certain genera or species are more prone to exchange ARGs due to selective or competitive pressure [12, 52]. Some taxonomic groups are also more prone to carry certain types of ARGs, such as β-lactamases, which is most often found in Enterobacteriaceae (including *E. coli*) [53].

Despite the importance for human health, the influence of age on the ARG profiles and its response to antibiotic exposure remain largely unknown. We wanted to explore these mechanisms in more detail by examining the resistome in 1-year-old infants and young adults.

PAPER III

In PAPER III we investigated how the resistome change with age or in response to antibiotic treatment. The ARG profiles from metagenomics samples from 662 infants and 217 adults were identified. The samples used for the study originated from the $COPSAC_{2010}$ and $COPSAC_{2000}$ cohort, comprising samples of one-year-old infants and 18-year-old young adults respectively [42, 54, 55]. A bimodal pattern is seen in the ARG abundance for both cohorts, with peaks indicating high and low richness. The duality is mainly driven by *Escherichia coli* (*E.coli*). A significant correlation between the cohort and the ARG profile of *E.coli* was seen. Additionally, we found that antibiotic treatment enhances ARG and MGE abundance and decrease the bacterial richness. The infant gut was found to recover faster from antibiotic treatment, despite harboring more plasmids than the adults. For both cohorts an increase of ARGs was seen after intake of antibiotics. The adult microbiome was found to harbor a lower diversity and abundance of ARGs as well as fewer bacteria carrying high abundance of ARGs such as *E.coli* compared to the infant cohort.

ARG profiles of *Escherichia*

Previous findings in the $COPSAC_{2010}$ cohort, showed that *Escherichia* and especially *E. coli* play a crucial role in shaping the ARG profiles [9]. The same pattern was observed in the $COPSAC_{2000}$ cohort. To gain more insights into the ARGs from the *Escherichia* genus, we constructed a phylogenetic tree from the MAGs annotated as *Escherichia*. 127 MAGs were found in the young adults and 513 MAGs were found in the infants. The Average Nucleotide Identity (ANI) was used to assess the similarity between the genomes (Figure 4.1). We tested whether the MAGs differed between the cohorts by creating a cophenetic distance matrix from the tree and testing the cluster-membership of the cohorts. From a phylogenetic perspective the *Escherichia* MAGs differed between the two cohorts (PERMANOVA; $P = 0.02$).

Additionally we wanted to include the information about the ARG

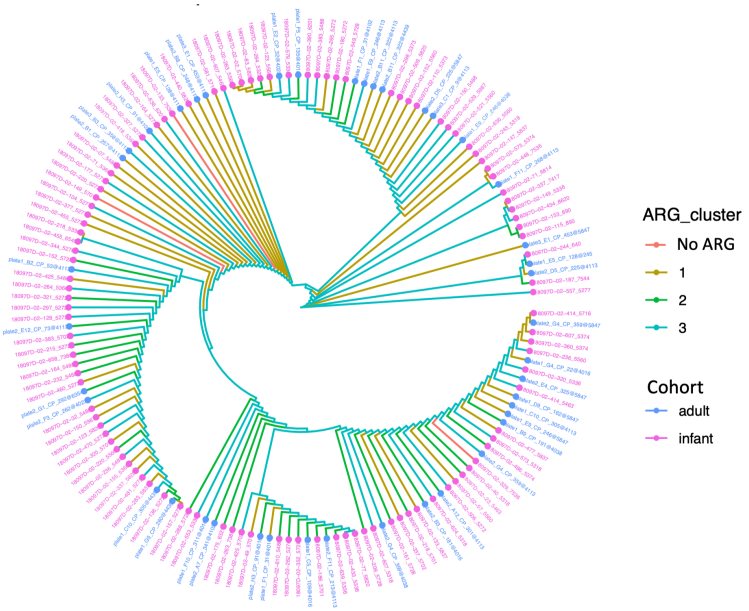


Figure 4.1: Phylogenetic tree of *Escherichia* MAGs in adult and infant gut based on 99% ANI analysis with data from PAPER III. The *Escherichia* MAGs are grouped into four categories using PAM clustering. The colored branches represent the four ARG profiles (red indicating no ARG in the MAG). The coloring of the tips indicate the cohort of origin. Figure adapted from PAPER III.

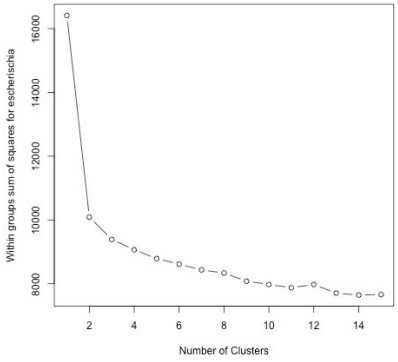


Figure 4.2: Dividing the ARG profiles (based on presence/absence) into optimal number of clusters. With PAM-clustering, different cluster-sizes are tested using wss for the *Escherichia* MAGs. Clusters between 2-15 are tested. The figure is created using data from PAPER III.

profiles in order to examine, whether they differed between the cohorts. Based on presence/absence of ARGs on the contigs we clustered MAGs with PAM clustering and found the optimal number of clusters to be 4 using the within-sum-of-squares (wss) method (Figure 4.2). We tested whether the *Escherichia* MAGs correlated with the ARG clusters and found a significant correlation (PERMANOVA; P = 0.01) (Figure 4.3 A). No ARG cluster was found to be exclusive for either adult or infant.

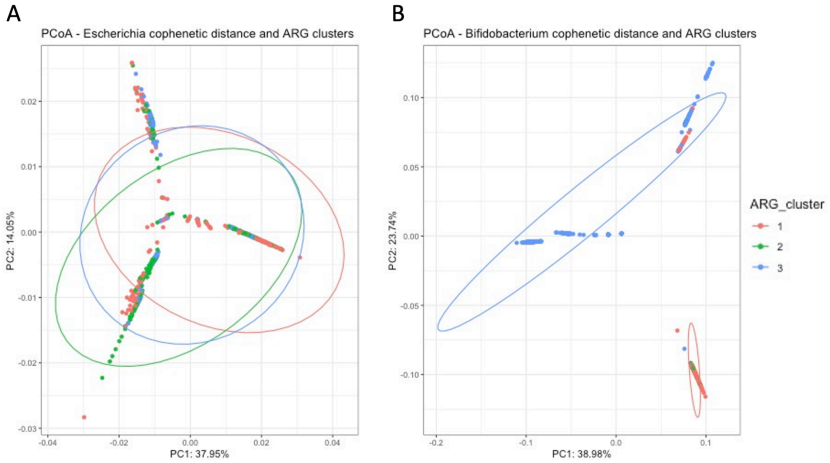


Figure 4.3: Correlation between taxonomy and ARG cluster membership. PCoA for the ARG cluster and cophenetic distance of the MAGs for A) *Escherichia* B) *Bifidobacterium*. Figure is created using data from PAPER III.

ARG profiles of *Bifidobacterium*

The method was repeated for *Bifidobacterium*, as this genus is found to play an important role in human health (described in Chapter 3). *Bifidobacterium* is occurring frequently and 2044 MAGs are identified across the two cohorts. From the MAGs a phylogenetic tree was constructed (Figure 4.4). From a phylogenetic perspective, *Bifidobacterium* MAGs differed between the two cohorts (PERMANOVA; P = 0.001). We also tested whether the *Bifidobacterium* MAGs correlated with the ARG clusters, using PAM clustering and wss, also resulting in 4 clusters. We found a significant correlation between MAGs and ARG cluster (PERMANOVA; P = 0.001) (Figure 4.3B). Furthermore, we found one ARG profile (cluster 3) to be almost exclusively present in infants

that was also predominantly distributed in one specific MAG cluster. Additionally, many MAGs from *Bifidobacterium* did not carry any ARGs.

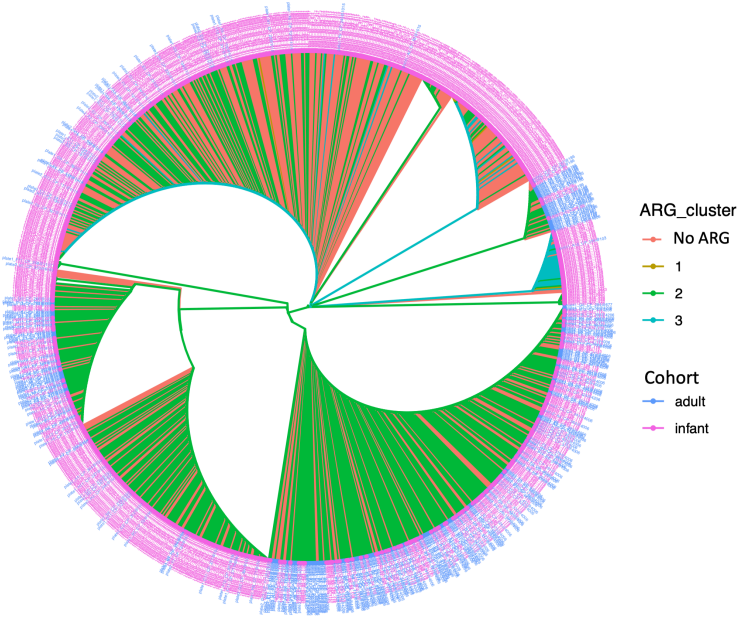


Figure 4.4: Phylogenetic tree of *Bifidobacterium* MAGs in adult and infant gut based on 99% ANI analysis using data from PAPER III. The *Bifidobacterium* MAGs are grouped into four categories using PAM clustering. The colored branches represent the four ARG profiles (red indicating no ARG in the MAG). The coloring of the tips indicate the cohort of origin. Figure adapted from PAPER III.

Conclusion

Based on the metagenomics sequencing of infants and young adults we were able to describe age-related patterns of the ARG profiles in terms of abundance and distribution in the gut. We were able to identify ARG clusters for *Escherichia* that were significantly correlated to the cohort. From a phylogenetic perspective, the *Escherichia* MAGs were also found to differ between the two cohorts.

CHAPTER 5

Strain level resolution used for environmental profiling

Pseudomonas aeruginosa (*P. aeruginosa*) is a common and widespread microbe found in various environments, and also an opportunistic pathogen in humans that can cause a range of different infections. Its ability to form biofilms [56] and its intrinsic and acquired resistance to a range of antibiotics makes it a challenging pathogen to combat [57, 58].

In the environment *P. aeruginosa* plays a crucial role in nutrient cycling and can be found in various ecological niches, such as soil, water (fresh and saline) and on plant surfaces. Its versatility allows it to adapt to a broad range of environments, making it an important bacterium in many microbial communities. Despite being found in most environments, the highest abundance of *P. aeruginosa* is found in humans or areas associated with human activity [59]. Some of the mechanisms, which could be influencing the large spread of the species is firstly its ability to form biofilms. Other contributing factors include its motility mechanisms such as flagella, and its ability to utilize a wide range of organic compounds as energy sources [57].

The metabolomic profile of *P. aeruginosa* is highly diverse and reflects its adaptability [59]. However there exists several *P. aerugi-*

nosa-specific metabolites (such as phenazines, rhamnolipids, quinolones and pyoverdin), which in symphony can be used as a molecular signature of the species, despite not being exclusively produced by *P. aeruginosa* [56, 60]. Additionally certain strains of *P. aeruginosa* produce toxins, which can be used to describe its virulence and pathogenicity [61].

Ongoing project

The scope of this project is to identify differences between host-associated and environmental *P. aeruginosa* strains. *P. aeruginosa* is nearly omnipresent in the environment and an opportunistic pathogen in humans. Our hypothesis suggests that genomic variability increases in host-associated strains due to the selective pressures exerted by the host immune system, antibiotic treatments, or competition with other microbes in host environments. In contrast, environmental strains are exposed to less stress, leading to a more conserved genome.

To examine the spread of *P. aeruginosa* we have used metagenomics, metabolomics and associated metadata from from the Earth Microbiome Project (EMP) [62, 63]. Additionally, we downloaded all reference genomes annotated as *aeruginosa* from NCBI [64].

MAGinator was used to identify signature genes and relative abundance of the reference strains in the EMP data. The bacterial abundance was correlated with the metabolomics data and corresponding environmental origin.

The project was carried out in collaboration with University of California San Diego (UCSD), as part of my external research stay at the Knight Lab.

Earth Microbiome Project (EMP)

The EMP is a systematic effort to characterize the global microbial world. The project aims at uncovering the taxonomic and

functional diversity of the microbes for the benefit of the whole planet. The data has been collected by research groups across the globe using standard protocols [62]. The data is comprehensive and includes metagenomics 16s rRNA and shogun sequencing, metabolomics and a broad selection of metadata, such as information regarding sampling environment, storage of the sample etc.

In this project 817 samples were included originating from a broad range of environments (Table 5.1), including 16 controls. The environments have been collapsed into two categories "Free-living" or "Host-associated" ((Table 5.2).

Environment	Sample Count
Animal corpus	67
Animal distal gut	182
Animal proximal gut	30
Animal secretion	20
Fungus corpus	12
Plant corpus	28
Plant surface	57
Sediment (non-saline)	47
Sediment (saline)	66
Soil (non-saline)	215
Subsurface (non-saline)	10
Surface (non-saline)	2
Surface (saline)	2
Water (non-saline)	24
Water (saline)	39

Table 5.1: Sample count for each environment, high stratification (controls have been removed).

Type	Count
Control	16
Free-living	405
Host-associated	396

Table 5.2: Sample count for each environment, low stratification.

The samples contained an average of 7.4 $\pm$ 8.8 million reads (Fig-

ure 5.1). An adequate sequencing depth rely on the taxonomic diversity of the community where the sample is extracted. However even for low-diversity habitats, this will be considered as shallow sequenced [65, 66].

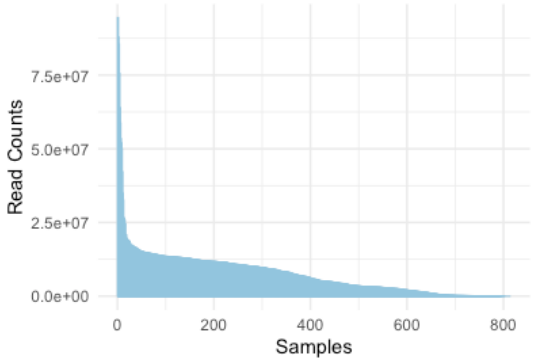


Figure 5.1: Characteristics for the EMP data included in this analysis. Read counts for the 817 samples (R1 counts shown).

MAGs in the EMP

Despite the shallow sequencing we tried to assemble and bin the samples. The samples were assembled using SPAdes [19], leaving a total of 2,4 million contigs. The contigs were binned using VAMB run with default settings, however due to the shallow sequencing and high diversity of the data only 6 MAG clusters were produced. Various settings for both assembly and binning was tested and the produced MAGs were examined using CheckM [67] (data not shown). With the sparse metagenomics data it was not possible to generate any results, where the spread of *P. aeruginosa* could be examined using the MAGs.

Using *P. aeruginosa* as reference catalogue

To circumvent the issues with the shallow sequencing we decided to create a reference catalogue of *P. aeruginosa*. Instead of using MAGs as input for MAGinator we used the reference catalogue, thus identifying signature genes for each strain, from which abundance estimates was found.

All 863 reference genomes annotated as *aeruginosa* were downloaded from NCBI [64]. As the signature genes must be unique

for the strains, the redundant genes were identified using MMseqs2 [68] and removed, leaving a nonredundant gene catalogue. The reads were mapped to the genes using bwa-mem2 [69] and counted using Samtools [70]. This resulted in an average of 360,580 ± 563,699 reads mapping from each sample to the gene catalogue (Figure 5.2 A). However, 104 samples have fewer than 1,000 reads that map. Per gene we find an average 3,081 ± 9,049 reads that map (Figure 5.2 B).

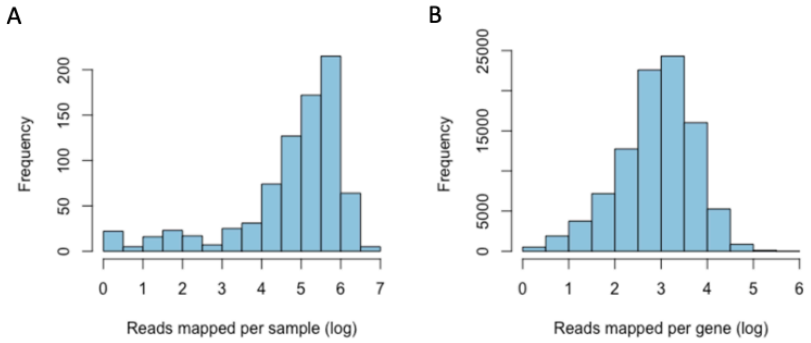


Figure 5.2: Read counts of EMP to the *aeruginosa* genes A) The number of reads that map to the genes per sample B) The number of reads that maps to each gene.

Signature genes of *P. aeruginosa* in EMP

The signature genes of each strain was found according to the methods presented in PAPER I. Based on the read mappings to the signature genes, the relative abundance was found. As the signature genes have to be unique for the strain, and the catalogue consists of 863 strains of the same species, the pool of genes is limited. We therefore suspected it to be hard to identify 100 unique signature genes for each strain. We tested gene set sizes of 100, 80, 60 and 20 genes. From Table 5.3 we see that the smaller the signature gene set is, the more strains is captured with more than 1‰ of the reads mapping. Having more strains identified could be a sign of noise of the data, indicating that 20 signature genes might be too few to capture the true presence/absence of the strains. However, 100 signature genes could be too many, disqualifying the strains with fewer unique genes.

Evolution of *P. aeruginosa* in different enviroments

Gene Set Size	Strains > 1‰ abund.
100	75
80	116
60	195
20	696

Table 5.3: Gene Set Size and the resulting number of strains harboring more than 1‰ of the mapped reads.

Using the relative abundances of the *P. aeruginosa* strains it is possible to associate it with the environments where it originated from (Table 5.1 and Table 5.2). No significant differences were found in alpha-diversity (data not shown).

The beta diversity and the proportion of variance explained was found (by scaling of the results, subsequent PCoA and PERMANOVA analysis (Figure 5.3 A+B), Table 5.4). None of the results are significant. However, we see that higher stratification of the environment yields higher proportion of explained variance.

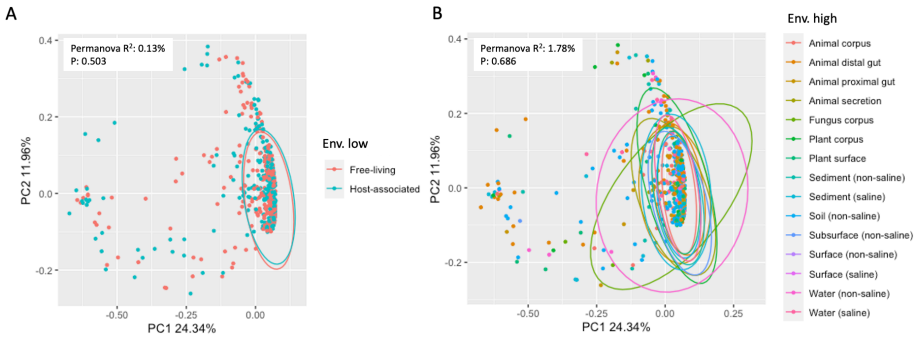


Figure 5.3: Beta diversity analysis of *aeruginosa* in the EMP data. PCoA and PERMANOVA (999 permutations) for A) Low stratification of environments B) High stratification of environments. Colors indicate environment.

Gene Set Size		Env. low	Env. high
100	R^2	0.12%	16.37%
	P	0.58	0.80
80	R^2	0.16%	1.71%
	P	0.29	0.75
60	R^2	0.13%	1.78%
	P	0.50	0.69
20	R^2	0.11%	1.93%
	P	0.67	0.52

Table 5.4: Correlation between beta diversity when using different gene set sizes versus low and high stratification of environmental conditions. Showing the PERMANOVA results (999 permutations, Bray-Curtis dissimilarity metric).

The algae *Microcystis aeruginosa*

When inspecting the results one of the MAGs did not belong to *P. aeruginosa*, but was instead the algae *Microcystis aeruginosa* (Figure 5.4). The algae is found primarily in fresh and brackish water [71]. When examining the beta diversity with the environmental variables "saline" and "non-saline", the algae is found to be correlated with salinity (Table 5.5). Interestingly the strongest correlated strains were 3 clinically isolated strains originating from the same study [72], which was found to be negatively correlated with "saline". Additionally, a strain sampled from the Indian Ocean [73] was found to be positively correlated with salinity.

	Saline	Non-saline	R^2	P
Human strain1	-0.43788	0.89903	0.0090	0.034 *
Human strain2	-0.83996	0.54265	0.0010	0.699
Human strain3	-0.43081	0.90244	0.0021	0.490
Algae	0.02568	-0.99967	0.0026	0.395
Marine strain	0.90529	-0.42480	0.0018	0.543

Table 5.5: Relationship between strains and environmental variables.

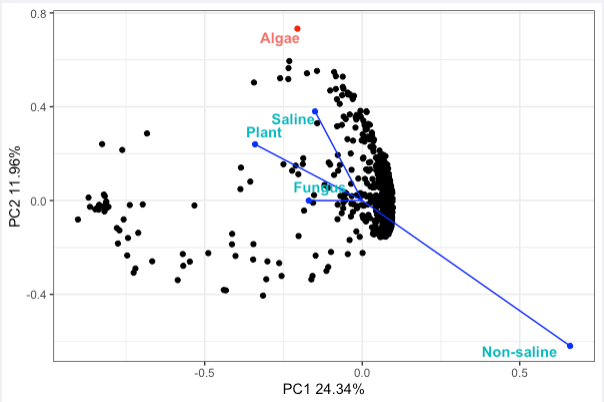


Figure 5.4: Beta diversity analysis displayed with directions of environmental variables. *Microcystis aeruginosa*, the algae, is marked with red.

Despite the algae not being part of the initial investigation, the observed results serve as a validation of the methodologies employed.

Metabolomic profile of *P. aeruginosa*

As part of the EMP collection, the metabolomic profile of the samples was found with untargeted LC-MS analysis. LC-MS is a combination of liquid chromatography (LC) and mass spectrometry (MS). Metabolomics is used to determine the products/metabolites that are found in the environment. The metabolites, which are microbially related have been identified and preprocessed according to the methods described by J. Shaffer et al. [63] and are used for the subsequent analysis.

The metabolite diversity was analyzed by applying identical methodology used for the beta-diversity assessment of the metagenomics data (Figure 5.5). Presence/absence of the metabolites in the samples have been used. The analysis shows a significant influence of the environmental origin of the sample, which can aid in explaining the diversity found within the metabolites.

This analysis is based on all metabolites related to microbes, however none of the *P. aeruginosa*-specific metabolites are present within the data.

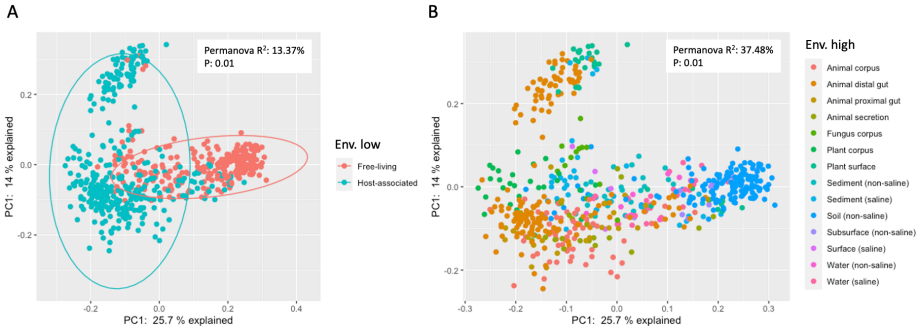


Figure 5.5: Diversity analysis of the EMP metabolomics data. Presence/absence of the metabolites have been used. PCoA and PERMANOVA (999 permutations) for A) Low stratification of environments B) High Stratification of environments. Colors indicate environment.

Next steps for uncovering diversity of *P. aeruginosa*

The initial investigations of the abundance of the *P. aeruginosa* did not lead to any conclusions regarding the genomic variability. This is most likely due to too many similar strains being present within the reference catalogue. If the reference genomes are too closely related, it can lead to problems in accurately selecting a set of signature genes unique for each strain but also lead to cross mapping of the reads. The latter was suspected, as the alpha diversity indicated that most strains was present with similar abundance in all environments. One approach to solve this is to cluster similar strains. This can be done based on their phylogeny.

However, we did successfully run MAGinator on a reference catalogue consisting of genomes instead of MAG clusters, which we believe can be a valuable asset in the future. This also opens for the possibility of combining the strengths of reference based and *de novo* identified MAG clusters for profiling of the metagenomics data sets.

Additionally, the metabolomic profile of the strains would have to be inferred by mining other metabolomics data sets, linking with the metabolites found in the current study. These can be found at public databases such as GNPS (Global Natural Products Social Molecular Networking) [74]. However this can pose a challenge as limited data exists for environmental strains of *P. aeruginosa*.

CHAPTER 6

Conclusion

The research presented in this thesis was centered around metagnomics profiling of the microbiome. The main scope was to refine and expand the current methods to allow strain-level resolution. This resolution allows for accurate integration of abundance, taxonomic and functional annotation in microbiome studies, which is needed to empower investigations in the microbiome field.

The initial research objective was to develop a method for selection of a set of signature genes, which can be used for precise detection and more reliable abundance estimations of the microbial entities. This work is demonstrated in PAPER I.

The signature genes was found to accurately profile the species in the samples. Additionally, by analysing the SNV-profiles of these genes we are able to further stratify the diversity of the bacteria in the samples to reach sub-species level identification as presented in PAPER II. The tool MAGinator was developed, covering a pipeline for *de novo* quantification and annotation of MAGs at sub-species level. It links the information from gene- and contig-based methods, allowing insights into both taxonomic profiles and the origin of the genes as well as their genetic content. This can be used for inference of the functional capabilities linked to the host organism and their presence within each sample.

In PAPER III the ARG profiles of young adults and infants are examined and compared using metagenomics. We identified and described age-related patterns of the ARG-profiles based on the composition and distribution. The bacterial composition was found to play a pivotal role in shaping the ARG profile. Especially *E. coli* was found to influence the ARG composition and certain ARG clusters was found to corretage with the cohort. This study displayed the importance of species and strain-specific profiling of the genomes found within the cohorts.

Lastly we investigated the spread and diversity of the opportunistic pathogen *P. aeruginosa* in different ecological niches using the EMP data. As described in Chapter V, we identified signature genes from a reference database of genomes, leading to abundance estimates of the strains in the different types of environments. The initial analysis did not leave us with any confident conclusion regarding the evolutionary differences of *P. aeruginosa* across environments. From the metabolites a clear pattern in diversity was seen between the environments, however as no *P. aeruginosa*-specific metabolites was present within the data it was not possible to directly link it with this species.

Collectively the research presented in this thesis has explored methods for profiling and characterizing the microbiome. The profound diversity and variability among the inhabitants of the microbiome make this a task that is still not fully solved. An additional avenue to achieve higher resolution is through the addition of long-read sequencing. As long-read sequencing can span span entire genomic regions, in combination with short-reads have been found to yield high quality hybrid assemblies [75]. Another interesting aspect to further characterize the microbiome is the presence and functions of other biological entities, such as vira and archaea, which has also been shown to influence the bacterial composition [76].

Bibliography

[1] P. J. Turnbaugh et al. “The Human Microbiome Project”. en. In: *Nature* 449.7164 (Oct. 2007), pp. 804–810.

[2] B. Stecher et al. “Gut inflammation can boost horizontal gene transfer between pathogenic and commensal *Enterobacteriaceae*”. en. In: *Proceedings of the National Academy of Sciences* 109.4 (Jan. 2012), pp. 1269–1274.

[3] L. K. Ursell et al. “Defining the human microbiome”. en. In: *Nutrition Reviews* 70 (Aug. 2012), S38–S44.

[4] E. A. Franzosa et al. “Gut microbiome structure and metabolic activity in inflammatory bowel disease”. en. In: *Nature Microbiology* 4.2 (Dec. 2018), pp. 293–305.

[5] X. Zhang et al. “Negative binomial mixed models for analyzing microbiome count data”. en. In: *BMC Bioinformatics* 18.1 (Dec. 2017), p. 4.

[6] J. Stokholm et al. “Maturation of the gut microbiome and risk of asthma in childhood”. en. In: *Nature Communications* 9.1 (Jan. 2018), p. 141.

[7] M. L. Calle, M. Pujolassos, and A. Susin. “coda4microbiome: compositional data analysis for microbiome cross-sectional and longitudinal studies”. en. In: *BMC Bioinformatics* 24.1 (Mar. 2023), p. 82.

[8] P. Han et al. “The Association Between Intestinal Bacteria and Allergic Diseases—Cause or Consequence?” en. In: *Frontiers in Cellular and Infection Microbiology* 11 (Apr. 2021), p. 650893.

[9] X. Li et al. “The infant gut resistome associates with E. coli, environmental exposures, gut microbiome maturity, and asthma-associated bacterial composition”. en. In: *Cell Host & Microbe* 29.6 (June 2021), 975–987.e4.

[10] C. Michaelis and E. Grohmann. “Horizontal Gene Transfer of Antibiotic Resistance Genes in Biofilms”. en. In: *Antibiotics* 12.2 (Feb. 2023), p. 328.

[11] S. C. Forster et al. “Strain-level characterization of broad host range mobile genetic elements transferring antibiotic resistance from the human microbiome”. en. In: *Nature Communications* 13.1 (Mar. 2022), p. 1445.

[12] S.-J. Paquette et al. “Competition among Escherichia coli Strains for Space and Resources”. en. In: *Veterinary Sciences* 5.4 (Nov. 2018), p. 93.

[13] P. Vikesland et al. “Differential Drivers of Antimicrobial Resistance across the World”. en. In: *Accounts of Chemical Research* 52.4 (Apr. 2019), pp. 916–924.

[14] K. Arikawa et al. “Recovery of strain-resolved genomes from human microbiome through an integration framework of single-cell genomics and metagenomics”. en. In: *Microbiome* 9.1 (Dec. 2021), p. 202.

[15] E. Z. Chen and H. Li. “A two-part mixed-effects model for analyzing longitudinal microbiome compositional data”. en. In: *Bioinformatics* 32.17 (Sept. 2016), pp. 2611–2617.

[16] Y. Zeng et al. “mbDenoise: microbiome data denoising using zero-inflated probabilistic principal components analysis”. en. In: *Genome Biology* 23.1 (Apr. 2022), p. 94.

[17] M. L. Calle. “Statistical Analysis of Metagenomics Data”. en. In: *Genomics & Informatics* 17.1 (Mar. 2019), e6.

[18] N. Sangwan, F. Xia, and J. A. Gilbert. “Recovering complete and draft population genomes from metagenome datasets”. en. In: *Microbiome* 4.1 (Dec. 2016), p. 8.

[19] A. Bankevich et al. “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing”. en. In: *Journal of Computational Biology* 19.5 (May 2012), pp. 455–477.

[20] S. Nurk et al. “metaSPAdes: a new versatile metagenomic assembler”. en. In: *Genome Research* 27.5 (May 2017), pp. 824–834.

[21] D. Li et al. “MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph”. en. In: *Bioinformatics* 31.10 (May 2015), pp. 1674–1676.

[22] A. Milanese et al. “Microbial abundance, activity and population genomic profiling with mOTUs2”. en. In: *Nature Communications* 10.1 (Mar. 2019), p. 1014.

[23] J. N. Nissen et al. “Improved metagenome binning and assembly using deep variational autoencoders”. en. In: *Nature Biotechnology* 39.5 (May 2021), pp. 555–560.

[24] T. Zachariasen et al. “Identification of representative species-specific genes for abundance measurements”. en. In: *Bioinformatics Advances* 3.1 (Jan. 2023). Ed. by S. Forslund, vbad060.

[25] H. B. Nielsen et al. “Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes”. en. In: *Nature Biotechnology* 32.8 (Aug. 2014), pp. 822–828.

[26] Y.-W. Wu, B. A. Simmons, and S. W. Singer. “MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets”. en. In: *Bioinformatics* 32.4 (Feb. 2016), pp. 605–607.

[27] S. Zouiouich et al. “Markers of metabolic health and gut microbiome diversity: findings from two population-based cohort studies”. en. In: *Diabetologia* 64.8 (Aug. 2021), pp. 1749–1759.

[28] M. N. Price, P. S. Dehal, and A. P. Arkin. “FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments”. en. In: *PLoS ONE* 5.3 (Mar. 2010). Ed. by A. F. Y. Poon, e9490.

[29] P.-A. Chaumeil et al. “GTDB-Tk v2: memory friendly classification with the genome taxonomy database”. en. In: *Bioinformatics* 38.23 (Nov. 2022). Ed. by K. Borgwardt, pp. 5315–5316.

[30] C. P. Cantalapiedra et al. “eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale”. en. In: *Molecular Biology and Evolution* 38.12 (Dec. 2021). Ed. by K. Tamura, pp. 5825–5829.

[31] F. Beghini et al. “Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3”. en. In: *eLife* 10 (May 2021), e65088.

[32] F. Plaza Oñate et al. “MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data”. en. In: *Bioinformatics* 35.9 (May 2019). Ed. by J. Wren, pp. 1544–1552.

[33] F. Bäckhed et al. “Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life”. en. In: *Cell Host & Microbe* 17.5 (May 2015), pp. 690–703.

[34] M. P. Sato et al. “Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes”. en. In: *DNA Research* 26.5 (Oct. 2019), pp. 391–398.

[35] D. J. C. Fowler. “Attacking variations of the coupon collector problem with Maple”. en. In: *27th International Conference on Technology in Collegiate Mathematics* (2016).

[36] M. Borderes et al. “A comprehensive evaluation of binning methods to recover human gut microbial species from a non-redundant reference gene catalog”. en. In: *NAR Genomics and Bioinformatics* 3.1 (Jan. 2021), lqab009.

[37] D. L. Dai et al. “Breastfeeding enrichment of B. longum subsp. infantis mitigates the effect of antibiotics on the microbiota and childhood asthma risk”. en. In: *Med* 4.2 (Feb. 2023), 92–112.e5.

[38] M. N. Ojima et al. “Priority effects shape the structure of infant-type Bifidobacterium communities on human milk oligosaccharides”. en. In: *The ISME Journal* 16.9 (Sept. 2022), pp. 2265–2279.

[39] S. Asakuma et al. “Physiology of Consumption of Human Milk Oligosaccharides by Infant Gut-associated Bifidobacteria”. en. In: *Journal of Biological Chemistry* 286.40 (Oct. 2011), pp. 34583–34592.

[40] F. Meyer et al. “Critical Assessment of Metagenome Interpretation: the second round of challenges”. en. In: *Nature Methods* 19.4 (Apr. 2022), pp. 429–440.

[41] T. J. Moraes et al. “The Canadian Healthy Infant Longitudinal Development Birth Cohort Study: Biological Samples and Biobanking: The CHILD study: biological samples”. en. In: *Paediatric and Perinatal Epidemiology* 29.1 (Jan. 2015), pp. 84–92.

[42] H. Bisgaard et al. “Deep phenotyping of the unselected COPSAC 2010 birth cohort study”. en. In: *Clinical & Experimental Allergy* 43.12 (Dec. 2013), pp. 1384–1394.

[43] D. T. Truong et al. “MetaPhlAn2 for enhanced metagenomic taxonomic profiling”. en. In: *Nature Methods* 12.10 (Oct. 2015), pp. 902–903.

[44] J. B. Ahrens, K. J. Wade, and D. D. Pollock. *A fast, general synteny detection engine*. en. preprint. Evolutionary Biology, June 2021.

[45] S. Van Dongen. “Graph Clustering Via a Discrete Uncoupling Process”. en. In: *SIAM Journal on Matrix Analysis and Applications* 30.1 (Jan. 2008), pp. 121–141.

[46] D. Vallenet. “MaGe: a microbial genome annotation system supported by synteny results”. en. In: *Nucleic Acids Research* 34.1 (Jan. 2006), pp. 53–65.

[47] J. Huerta-Cepas et al. “eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses”. en. In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D309–D314.

[48] B. Buchfink, C. Xie, and D. H. Huson. “Fast and sensitive protein alignment using DIAMOND”. en. In: *Nature Methods* 12.1 (Jan. 2015), pp. 59–60.

[49] M. Kanehisa and S. Goto. “KEGG: Kyoto Encyclopedia of Genes and Genomes”. en. In: *Nucleic Acids Research* 28.1 (2000).

[50] F. Mölder et al. “Sustainable data analysis with Snakemake”. en. In: *F1000Research* 10 (Jan. 2021), p. 33.

[51] QuantStack and Mamba contributers. *Mamba*. 2020.

[52] H. Darmancier et al. “Are Virulence and Antibiotic Resistance Genes Linked? A Comprehensive Analysis of Bacterial Chromosomes and Plasmids”. en. In: *Antibiotics* 11.6 (May 2022), p. 706.

[53] G. Zarfel et al. “Comparison of extended-spectrum- -lactamase (ESBL) carrying Escherichia coli from sewage sludge and human urinary tract infection”. en. In: *Environmental Pollution* 173 (Feb. 2013), pp. 192–199.

[54] H. Bisgaard et al. “Fish Oil–Derived Fatty Acids in Pregnancy and Wheeze and Asthma in Offspring”. en. In: *New England Journal of Medicine* 375.26 (Dec. 2016), pp. 2530–2539.

[55] H. Bisgaard. “The Copenhagen Prospective Study on Asthma in Childhood (COPSAC): design, rationale, and baseline data from a longitudinal birth cohort study”. In: (2004).

[56] L. A. Meirelles and D. K. Newman. “Both toxic and beneficial effects of pyocyanin contribute to the lifecycle of *Pseudomonas aeruginosa*”. en. In: *Molecular Microbiology* 110.6 (Dec. 2018), pp. 995–1010.

[57] M. Ratajczak et al. “Relationship between antibiotic resistance, biofilm formation, genes coding virulence factors and source of origin of *Pseudomonas aeruginosa* clinical strains”. en. In: *Annals of Agricultural and Environmental Medicine* 28.2 (June 2021), pp. 306–313.

[58] K. C. Costa et al. “Pyocyanin degradation by a tautomerizing demethylase inhibits *Pseudomonas aeruginosa* biofilms”. en. In: *Science* 355.6321 (Jan. 2017), pp. 170–173.

[59] S. Crone et al. “The environmental occurrence of *Pseudomonas aeruginosa*”. en. In: *APMIS* 128.3 (Mar. 2020), pp. 220–231.

[60] F. A. Alatraktchi, W. E. Svendsen, and S. Molin. “Electrochemical Detection of Pyocyanin as a Biomarker for Pseudomonas aeruginosa: A Focused Review”. en. In: *Sensors* 20.18 (Sept. 2020), p. 5218.

[61] D. D. Nguyen et al. “Indexing the Pseudomonas specialized metabolome enabled the discovery of poaeamide B and the bananamides”. en. In: *Nature Microbiology* 2.1 (Oct. 2016), p. 16197.

[62] L. R. Thompson et al. “A communal catalogue reveals Earth’s multiscale microbial diversity”. en. In: *Nature* 551.7681 (Nov. 2017), pp. 457–463.

[63] J. P. Shaffer et al. “Standardized multi-omics of Earth’s microbiomes reveals microbial and metabolite diversity”. en. In: *Nature Microbiology* 7.12 (Nov. 2022), pp. 2128–2150.

[64] E. W. Sayers et al. “Database resources of the national center for biotechnology information”. en. In: *Nucleic Acids Research* 50.D1 (Jan. 2022), pp. D20–D26.

[65] on behalf of the REHAB consortium et al. “The impact of sequencing depth on the inferred taxonomic composition and AMR gene content of metagenomic samples”. en. In: *Environmental Microbiome* 14.1 (Dec. 2019), p. 7.

[66] C. Pal et al. “The structure and diversity of human, animal and environmental resistomes”. en. In: *Microbiome* 4.1 (Dec. 2016), p. 54.

[67] D. H. Parks et al. “CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes”. en. In: *Genome Research* 25.7 (July 2015), pp. 1043–1055.

[68] M. Steinegger and J. Söding. “MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets”. en. In: *Nature Biotechnology* 35.11 (Nov. 2017), pp. 1026–1028.

[69] M. Vasimuddin et al. “Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems”. en. In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. Rio de Janeiro, Brazil: IEEE, May 2019, pp. 314–324.

[70] H. Li et al. “The Sequence Alignment/Map format and SAMtools”. en. In: *Bioinformatics* 25.16 (Aug. 2009), pp. 2078–2079.

[71] M. Georges Des Aulnois et al. “Salt Shock Responses of Microcystis Revealed through Physiological, Transcript, and Metabolomic Analyses”. en. In: *Toxins* 12.3 (Mar. 2020), p. 192.

[72] N. Murugan et al. “Comparative Genomic Analysis of Multidrug-Resistant Pseudomonas aeruginosa Clinical Isolates VRFPA06 and VRFPA08 with VRFPA07”. en. In: *Genome Announcements* 2.2 (May 2014), e00140–14.

[73] M.-V. Grosso-Becerra et al. “Pseudomonas aeruginosa clinical and environmental isolates constitute a single population with high phenotypic diversity”. en. In: *BMC Genomics* 15.1 (Dec. 2014), p. 318.

[74] M. Wang et al. “Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking”. en. In: *Nature Biotechnology* 34.8 (Aug. 2016), pp. 828–837.

[75] D. Bertrand et al. “Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes”. en. In: *Nature Biotechnology* 37.8 (Aug. 2019), pp. 937–944.

[76] G. Liang and F. D. Bushman. “The human virome: assembly, composition and host interactions”. en. In: *Nature Reviews Microbiology* 19.8 (Aug. 2021), pp. 514–527.

PAPER I

Identification of representative species-specific genes for abundance measurements

Zachariasen, T., Petersen, A. Ø., Brejnrod, A., Vestergaard, G. A., Eklund, A., Nielsen, H. B.

Published in Bioinformatics Advances, 2023

The work carried out in relation to this thesis included an exploratory analysis of the data. The work comprised investigation of ways of modelling the read count distributions, assessing the performance of a gene set, identifying a suitable method for testing new genes and interpretation of the results. As the corresponding author I have been in charge of the submission and subsequent review of the paper.

Bioinformatics Advances, 2023, vbad060
https://doi.org/10.1093/bioadv/vbad060
Advance Access Publication Date: 8 May 2023
Original Paper

OXFORD

Metagenomics

Identification of representative species-specific genes for abundance measurements

Trine Zachariasen [1,*], Anders Østergaard Petersen[1], Asker Brejnrod[1], Gisle Alberg Vestergaard[1], Aron Eklund [2] and Henrik Bjørn Nielsen[2]

[1]Department of Health and Technology, Technical University of Denmark, Lyngby 2800, Denmark and [2]Clinical Microbiomics A/S, Copenhagen 2100, Denmark

*To whom correspondence should be addressed.

Associate Editor: Sofia Forslund

Received on August 26, 2022; revised on April 14, 2023; editorial decision on April 29, 2023; accepted on May 5, 2023

Abstract

Motivation: Metagenomic binning facilitates the reconstruction of genomes and identification of Metagenomic Species Pan-genomes or Metagenomic Assembled Genomes. We propose a method for identifying a set of *de novo* representative genes, termed signature genes, which can be used to measure the relative abundance and used as markers of each metagenomic species with high accuracy.

Results: An initial set of the 100 genes that correlate with the median gene abundance profile of the entity is selected. A variant of the coupon collector's problem was utilized to evaluate the probability of identifying a certain number of unique genes in a sample. This allows us to reject the abundance measurements of strains exhibiting a significantly skewed gene representation. A rank-based negative binomial model is employed to assess the performance of different gene sets across a large set of samples, facilitating identification of an optimal signature gene set for the entity. When benchmarked the method on a synthetic gene catalog, our optimized signature gene sets estimate relative abundance significantly closer to the true relative abundance compared to the starting gene sets extracted from the metagenomic species. The method was able to replicate results from a study with real data and identify around three times as many metagenomic entities.

Availability and implementation: The code used for the analysis is available on GitHub: https://github.com/trinezac/SG_optimization.

Contact: trizac@dtu.dk

Supplementary information: Supplementary data are available at *Bioinformatics Advances* online.

1 Introduction

Metagenomic binning tools, such as MetaBAT 2 (Kang *et al.*, 2019), VAMB (Nissen *et al.*, 2021) and MSPminer (Plaza Oñate *et al.*, 2019), facilitate the reconstruction of genomes and identification of metagenomic entities, such as Metagenomic Species Pan-genomes (MSPs) or Metagenomic Assembled Genomes (MAGs), by gathering groups of genetic components, such as genes or contigs, that are believed to originate from a clade. The clade is typically at the species or subspecies level, where the gene composition is relatively conserved (Nielsen *et al.*, 2014). The composition of a typical metagenomic sample is a priori unknown and may contain novel organisms, new variants of already characterized organisms, and closely related but distinct organisms. This challenges the metagenomic detection and quantification of the microbiome, since sequence reads from one species may map perfectly to the reference sequences of another species (Sangwan *et al.*, 2016). Stringent mapping may reduce the cross-species mapping, but this may come at the expense of robustness in quantifying variants of a species. Genes that are specific to a given strain, yet present in all members of that clade, are ideally suited for measuring the abundance of the species by eliminating cross-mapping of reads while allowing for accurate and precise measure of a given strain. Additionally, SGs should not be duplicated within a strain to avoid biasing the abundances of strains whose genes have high copy numbers. Such a set of genes is referred to as a signature gene set (Segata *et al.*, 2012). SGs have previously been identified by comparing reference genomes from species-level clades (Milanese *et al.*, 2019). This approach works when sufficiently many reference sequences are available from a given species as well as from the species from which the reference is to be distinguished from. However, for species with few available reference genomes, or few genomes from related species, it is difficult to define signature gene sets. Additionally, certain genomic sequences are easier to sequence, yielding more reads and a skewed read distribution throughout the genome. A selection of signature genes that does not account for this has the potential to artificially inflate the relative abundance of certain species. When no references are available, a set of SGs can be identified based on their ability to

quantify a species (or any clade of interest) in a given context, e.g. the human microbiome. We propose a method that relies entirely on a statistical analysis of the distributions of readmappings to the genes and that is entirely agnostic to bias in read generation, gene duplication, etc. This method searches for gene sets that produce robust and even mapping across natural population variability and minimize signal noise. Within each sample, the expected number of mapped reads per gene can be approximated by the discrete negative binomial (NB) distribution (Zhang *et al.*, 2017), as the reads are assumed to map in proportion to the gene length and exhibit some degree of variability. As the gene lengths are known, the total number of sequence reads that map to a good signature gene set should predict the number of genes in the set that the reads map to. In other words, the reads should appear to be drawn randomly from across the gene set. Large deviance from the expected model could be due to violations of the aforementioned characteristics of a good signature gene, i.e. genes that are not omnipresent to a given strain or not present in all members of that strain. Here, we illustrate the necessity for such an approach and propose a method for defining optimal gene sets and for estimating the likelihood that the observed read mappings only originate from a population that comprised the complete SG set in equal quantities. In this article, we propose (i) a method for selecting optimal signature gene sets and (ii) the use of a special case of the 'coupon collector's problem' (CCP) to assess the likelihood that sequence reads will map to a specific number of genes (d) given the number of reads (k) that map to the entire gene set.

Binning is typically divided into two major approaches, gene based and contig based. Contig-based binning is especially useful when trying to reconstruct whole genomes, while gene-based binning is useful for identifying and characterizing microbial communities at a higher taxonomic level. The method has been created to aid in *de novo* identification of species, for both gene- and contig-based methods as well as for profiling of species defined by reference genomes. The following results stem from the analysis of a simulated gene catalog (SGC) from Borderes *et al.* (2021) as well as a case study performed on the First Year of Life Dataset from Bäckhed *et al.* (2015).

2 Methods

2.1 Input data and formatting

The input for our method is a gene count matrix (comprising information about the number of mapped reads to each gene within each sample) as well as information linking the genes to their corresponding biological entity. In this study, we illustrate two different methods for creating these data structures.

2.1.1 The SGC

The non-redundant SGC used in this study is meticulously designed by Borderes *et al.* (2021). The short reads of the SGC are created by GenSIM (v.1.6) (McElroy *et al.*, 2012) and constructed based on the genomes of 47 strains belonging to 41 species and theoretical abundance profiles of 40 samples. Borderes et al. mapped the reads using MOCAT (v.2.0) (Kultima *et al.*, 2012) and SOAPALIGNER2 (Li *et al.*, 2009) with the 'allbest' mapping mode, and to generate gene abundance profiles for all samples (Kultima et al., 2012; Li et al., 2009). As the genes of the species and their corresponding abundances are known in the SGC, a golden standard has been created containing the gene identifiers and their associated species.

2.1.2 The MSPs

MSPminer (v. updated 2018-04-25) (Plaza Oñate *et al.*, 2019) has been applied to the SGC to identify the MSPs. Each MSP is a collection of clustered genes belonging to a biological entity. MSPminer is run with default parameters and the results are summarized in a tab-separated file, containing the genes and its corresponding MSPid. MSPminer divides the reads into a total of 54 bins with the number of clustered genes ranging from 575 to 5957. Each MSP is on average present in 36 samples, ranging from a minimum of 26 samples to a maximum of 40.

2.1.3 A case study using Bäkhed's First Year of Life dataset

The First Year of Life study has been used as a case study, to illustrate the SG method in a well-designed study with high-quality data. The dataset constructed by Bäckhed *et al.* (2015) comprises a total of 392 short-read samples from 98 infants [3 different timepoints (Newborn, 4M, 12M) and a sample from their mother]. The samples have been shotgun sequenced with an average of 3.99 Gb of reads per sample.

2.1.4 Creation of VAMB clusters

The samples have been through preprocessing including adapter removal using BBDuk (v. 38.96 http://jgi.doe.gov/data-and-tools/bb-tools/), removal of low-quality reads and reads shorter than 75 base pairs using Sickle (v. 1.33) (Joshi and Fass, 2011) and removal of human contamination (reference UCSC version hg19, GRCh37.p13) with BBmap (http://jgi.doe.gov/data-and-tools/bb-tools/).

De novo assembly was carried out per sample with Spades (v. 3.15.5) run with the meta-option (15) and kmer sizes of 213 355 and 77. Contigs <1500 bp were discarded. BWA-mem2 (v.2.2.1) (Vasimuddin *et al.*, 2019) and SAMTOOLS (v.1.10) (17) were used for mapping the reads to the assemblies. Metabat2's jgi_summarize_bam_contig_depths (v.2.12) (Kang *et al.*, 2019) was used to assess the depths of the contigs. VAMB (v.3.0.8) (Nissen *et al.*, 2021) was run with default parameters to bin the contigs. Annotation of the bins along with gene predictions was done using GTDK-tk (v.2.1.1) (Chaumeil *et al.*, 2022). The genes were clustered using MMseqs2 (v. 13.45111) (Steinegger and Söding, 2017) with a sequence identity threshold for the clustering of genes of 0.8. The remaining genes are used for the construction of a gene count matrix for each VAMB cluster, containing the read counts of each gene within each sample.

2.2 Data preparation

The statistical analysis and data handling have been performed in R (v.4.1.2) (R Core Team, 2021). The genes of the entity are sorted according to co-abundance with the genes with highest intra-species abundance correlation as the first genes within the entity. Different sizes of gene sets were evaluated by comparing the absolute difference in abundance between the predicted and the true abundance using the SGC. We tested the gene set sizes in the range between 70 and 150 and found a local minimum of 100 genes. A metagenomic species is considered detected in a sample if it contains reads that map to three or more signature genes. The read counts, which are normalized according to gene length, are multiplied by 1000 to avoid small numbers and rounded to the closest integer.

2.3 Development of benchmark

To assess the chance of identifying d out of n SGs given k reads assigned to signature genes, we use an analytical solution to a variant of the CCP described in a 2008 conference summary published from the *27th International Conference on Technology in Collegiate Mathematics*, by Fowler (2015). The variant of the CCP tackled in this article is the chance of drawing exactly d different balls out of an urn containing n different balls given k draws of one ball with replacement. The solution to this problem is

$$P(d,k,n) = \frac{\binom{n}{d}}{n^k} S(k,d)d! \tag{1}$$

As the generation of Stirling numbers of the second kind $S(k, d)$ is computationally intensive, pre-computed values of $S(k, d)$ are obtained from www.planetcalc.com, a resource for solutions to common mathematical problems. This resource lists solutions for $k = \{0, 1, 2, \ldots 176\}$, $d = \{0, 1, 2, \ldots k\}$. Eighty-six percent of P-values encountered in the dataset can be computed in this manner for $n = 100$. For the estimation of $P(d, k, n)$ in cases where $k > 176$ and/or n^k is evaluated as Inf by R, utilization of pre-computed

bootstrapping results is carried out. Bootstrapping is carried out by randomly sampling n genes k times and evaluating the number of different genes obtained, d, 10^5 times. The chance of obtaining exactly d out of n genes given k reads is evaluated as the number of times d unique genes was obtained out of 10^5 iterations. Bootstrapping is carried out for $k = \{0, 1, 2, \ldots 3000\}$, to $n = \{0, 1, 2, \ldots k\}$ for a given value of n. Solutions to $P(d, k, n)$ where $k > 3000$ are thus approximated as

$$P(d = n, k, n) \sim 1,\ k \geq 3000 \tag{2}$$

and

$$P(d < n, k, n) \sim 0,\ k \geq 3000,\ n \geq 3000 \tag{3}$$

To assess the degree of accuracy of bootstrapping, P-values obtained by bootstrapping were compared to P-values obtained by analytical evaluation across the entire dataset tested, the Pearson correlation coefficient (PCC) was evaluated as 0.99. Thus, this variation in the CCP is made to evaluate the performance of a set of SGs as a whole, ensuring that none of the genes are disproportionately easy or difficult to identify.

2.4 Signature gene refinement

2.4.1 Introduction

The performance of individual genes is evaluated using an NB model, which evaluates whether higher sequencing depth also reliably leads to higher read counts. The ranking of the genes enables the detection and removal of SG, which are found inconsistently across the samples. As the method utilizes the power across samples, one limitation to the method is that it requires at least three samples. Genes are ranked according to how well they fit this model in each sample and replaced with genes evaluated with CCP. Initially, all samples that contain three or more reads towards a certain set of SGs are identified. In the first step, SGs are exchanged if their mean rank across samples is above a certain threshold, t. In this way, we only replace genes that consistently underperform across multiple samples. Genes are ranked by how well they fit the NB model in each sample by the size of the residual, and the mean is taken across samples resulting in the mean rank. If the exchanged SG set has a lower mean squared error (MSE) than the previous set, the SG set is kept and reruns until the MSE no longer decreases. The method is repeated, this time assessing whether any genes are outlying in a subset of the samples. If the MSE improves for multiple thresholds, the refined SG set is selected as the one with the lowest MSE (Fig. 1).

2.4.2 Frequency-based filtering

The ranking of genes using an NB model ensures removal of genes from the original SG set whose detection is inconsistent across samples. However, it does not ensure that selected genes have a similar ease of detection across samples. Ideally, the genes within an SG set are found with an equal probability; however, it is expected that biological and technical bias will lead to a skewed sampling of the genes. This can lead to systemic biases in abundance estimation that favour the abundances of strains with sensitive SGs. Additionally, a good set of SGs should be sensitive, i.e. be part of genomic sequences that are easily sequenced, to detect low-abundant strains in a sample. To accommodate this prior to the refinement of the SGs, the genes need to be prescreened and ordered according to their sensitivity, such that an increase in k will entail an increase in d for the samples. The over- or undersampling of genes is alleviated using systematic replacement of genes, implementing a pre-filtering step in which a set of SGs with similar, high sensitivities were selected for replacing poorly performing genes, while avoiding genes whose sensitivities were very different from the other SGs. The genes assigned to the respective metagenomic entity are sorted in order of decreasing frequency of detection across all samples. A set of 700 genes are selected, which have the highest overall frequency of detection, excluding genes whose frequencies were outside the 1.2 interquartile range of the rest of the set. Thereby selecting genes with a high frequency of detection, but at the same time are also found in a consistent manner, ensuring the SGs that are used for replacement are all easy to detect. The genes used for the replacement of the SGs are found within this pool of 700 genes leading to a more heterogeneous frequency of detection of the genes included in the final SG set. In the case of an entity with <700 genes, all genes are used.

2.4.3 Ranking of genes

As part of identification of genes that should be removed, first, we must evaluate genes based on the consistency of detection. To assess the performances of each gene, an NB distribution is used to test whether increased sequencing depth reliably leads to additional counts of that gene. How consistently a gene is detected has previously been shown to follow an NB distribution (Zhang *et al.*, 2017) as the NB model is known to handle overdispersion that is frequently observed in sequencing data. The mentioned model is applied for each sample, where the read count of gene i in the jth sample is denoted y_{ij}, then

$$\begin{aligned} y_{ij} &\sim NB\left(\, y_{ij} | \mu_j, \sigma_j\right) \\ &= \Gamma(y_{ij} + \sigma_j)/\Gamma(\sigma_j) y_{ij}! \cdot \left(\sigma_j/\mu_j + \sigma\right)^{\sigma_j} \cdot \left(\mu_j/\mu_j + \sigma_j\right)^{y_{ij}},\ \mu_j > 0 \end{aligned} \tag{4}$$

where μ_j is the average read count per gene, σ_j is the sample-specific NB dispersion parameter and $\mathbf{\Gamma}(\cdot)$ denotes the gamma function. The NB model can be seen as a compounded Poisson-Gamma distribution, in which the rate parameter of the Poisson model itself is a random variable distributed according to a Gamma distribution (Zhang *et al.*, 2017). When the distribution approaches the Poisson distribution with equal mean and variance. From this parametrization of the NB model, the expected read count is given as $E[y_{ij}] = \mu_j = \lambda_{ij} N_j$, where λ_{ij} is the proportion of reads mapped to gene i in the jth sample and N_j is the total number of reads mapped to sample j; thus, μ_j depends on the sequencing depth as well as the abundance of the species in the sample. The variance of the read count is given as $var(y_{ij}) = \mu_j + \mu_j{}^2 1/\sigma_j$. The counts of each SG are evaluated according to this NB model and are ranked within each sample by evaluating the difference between the expected count and observed count.

2.4.4 Rejection and replacement

We use the mean rank of each gene across the samples to evaluate the performance of the SG, which enables the detection of SG with persistent discrepancies according to the NB model. If the genes have a lower average rank than a given threshold, consequently underperforming, the genes will be removed from the SG set, thereby leaving a smaller SG set, in which we have higher confidence that the genes are consistently found across samples. A range of thresholds are tested to obtain the best possible gene set. If the remaining SG set maps to <10 samples, the refined SG set will not be considered for further analysis, as the data are too scarce for reliably ranking of the SGs. The NB model is reapplied to the retained SG set to exclude potential noise caused by the already removed genes. The NB distribution is fitted exclusively on the genes, which we believe to reliably lead to an increase in SG detection as sequencing depth increases. A subset of the frequency-based filtered pool of genes are introduced to the SG set, leaving a complete SG set of 100 genes. The introduced subset is selected as the genes with the highest co-abundance, which were also accepted in the filtering step and have not already been included in the SG set. When assuming that each read has an equal change of mapping to each signature gene and that the mapping process of each read is independent of the previous reads, the probability of a gene not being detected can be described by

$$P_0 = \frac{(n-1)^k}{n} \tag{5}$$

where n is the number of signature genes and k is the number of reads that map to the SGs in that sample. By taking the complement of P_0, we can calculate the probability of an SG being detected can thus be calculated as $P_1 = 1 - P_0$. This can be utilized to calculate

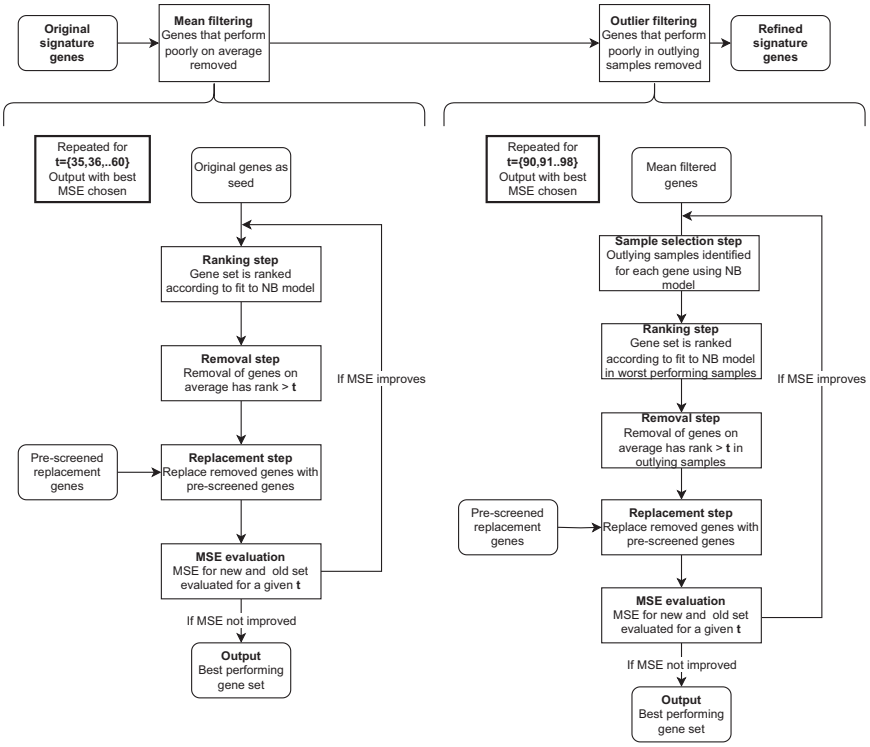


Fig. 1. Two-step signature gene refinement algorithm as described in methods. NB: negative binomial

the expected number of detected signature genes d, which for each sample j as

$$d_j = \left(1 - (n-1)^{k_j}/n\right)n,\ j = 1, 2, \ldots, m \qquad (6)$$

where k_j is the total number of reads mapped to the SGs and m is the number of samples. We assume that each read has an equal chance of mapping to each signature gene (after gene length normalization) and that the mapping process of each read is independent of the previous reads. The effect of SG replacement is evaluated based on the deviation from the expected distribution [Equation (5)]. Only if the deviation has been reduced, the changes to the SG are kept. The process of ranking, removing and replacing is repeated until the MSE is reduced by less than 1% from the previous iteration. The result is kept for the threshold that performs the best. By iteratively improving the SG set and reevaluating the NB approximation the gene set are continuously improved, leading to a gene set that is more reliable for abundance estimation of the species, as the genes are more often present within the majority of the strains.

The optimum threshold varies between the metagenomic entities, as each set of SG deviates from the expected distribution [Equation (5)] differently, leading to a different spread of the mean rankings of the SG. If the SGs are detected consistently across the samples, the SGs will have a small spread in average rank. In case the detection of the SGs is inconsistent, a large spread in the average ranking of the genes will be observed. In the first step, referred to as mean filtering, all integers in the range 35–60 are tested to identify the threshold for mean-rank leading to the genes, which are most reliably detected across samples to accommodate the differences of the metagenomic entities. The range of thresholds is selected, as testing indicates that the majority of optimal thresholds for obtaining the best possible SG set falls well within this range.

In some cases, one or more genes will be consistently missing in a smaller subset of samples, which the mean-based filtering across all samples cannot alleviate. To capture these genes, the ranking method is re-implemented, but where the average ranking of the SG was previously used, we are now evaluating the genes based on the 95th percentile according to adherence to the expected distribution [Equation (5)]. By selecting the 95th percentile, we are considering only the 5% of the samples that perform the worst. If SGs are persistently diverging from the NB model in this subset of samples, they are tried and replaced after which it is examined whether the refined SG set follows the expected distribution more closely. The optimal threshold for the removal of SGs based on the rank of the 95th percentile is found in the range 90–98.

2.5 Benchmark calibration

The goal of identification of SGs is to obtain a sizable set of genes that are shared by all members of a strain. Ideally, all 100 SGs are identified in all samples with high sequencing depth; however, it is very difficult to select a set of SGs that are all identified synchronistically when many reads are assigned to SGs, indicated by high k values. Any sample with very high read depth that contains the

metagenomic entity at an adequate abundance would result in a k that approaches the number of SGs chosen for that strain and at very high sequencing depths the chance of not finding all n SGs approaches zero. Any biological variation in SGs in deeply sequenced samples would be rejected and assigned a very low probability of occurring due to the assumption that all SGs are present without biological variation. This limits the applicability of this method. To allow for a biologically unsubstantial degree of variation in SGs, we will consider the chances of obtaining d or fewer unique SGs out if a species contains *at least* n SGs given k assigned reads instead of all n. We must arrive at a value of n that does not unfairly reject samples that are missing an inconsequential amount of SGs due to unimportant degrees of biological variation, while still rejecting samples whose metagenomic entity shows a clear lack of SGs, e.g. due to strain differences. A fair threshold should be able to distinguish samples in which strain differences show from the SGs but still allow for smaller biological differences to appear. We select $n = 95$, such that for a random distribution, one would expect approximately 5% of the samples to fall below $P < 0.05$. MSP07 is the best-performing initial gene set in our catalog with an MSE of 0.7 and setting an n of 95 rejects 1/40 of samples, which we consider to be a sufficient approximation. The null hypothesis we test against is that a sample contains 95 different SGs with an equal chance of finding each SG. The choice of rejecting or accepting a sample will be evaluated as the chance of obtaining d or fewer different SGs out of 95 SGs, given k reads assigned to SGs.

3 Results

3.1 Identifying the optimal signature gene sets

Typically, a strong relationship is observed between the number of reads mapped to the metagenomic entity and the number of detected genes within the gene set of each sample. However, for some MSPs, part of the initial gene set is rarely detected despite high coverage of the remaining genes. The gene set is selected as the 100 genes with the highest co-abundance correlation (Pearson's correlation coefficient) to the median abundance for all genes. An NB model was used to assess the probability that all of the genes in a given gene set are present, in exactly one copy per genome, in a sample, given the observed read mappings. The probability that a sample contains the complete gene set is dependent on the number of reads that map to the gene set (k), the number of different genes from the SGs detected in the sample (d) and the threshold for what is considered a complete SG set (n). Using this statistical framework, we can evaluate the expected number of detected genes from the set, d, given a number of mapped reads, k. From this expected distribution, we can evaluate the performance of an SG set by its MSE between the observed and expected numbers of identified genes from the set across a series of samples. During refinement of the gene set, the MSE was used to reject changes to the SG set that led to an increase and accept changes that reduced it. During the evaluation of individual samples, some biological variation is allowed by setting the n to 95; hence, a species, with a given SG, is considered detected in a sample if the observed reads mapped fit with an SG set with at least 95 genes. When evaluating MSEs, for refinement or otherwise, the expected distribution is derived from a distribution with $n = 100$ to avoid optimizing for an incomplete set of SGs.

Refinement was done using a two-step approach as described in methods, which relies on replacement of genes that perform consistently poorly across multiple samples. Poor performing genes were replaced until improvement of MSE was negligible. The performance of the improved signature gene sets will be assessed on four parameters: model fit (MSE), amount of initial SGs retained, PCC of counts between the initial and refined SG sets and change in the number of samples where the observed reads mapped fit with significantly reduced number of genes in the SG. To allow for a negligible amount of biologically variation in SG evaluation, samples thought to contain 95 or more different SGs will be accepted as wholly.

The method was applied to the SGC created from 40 simulated metagenomics samples with reads from 47 reference strains (Borderes *et al.*, 2021). To assess how close a set of SGs follow the expected distribution, the MSE between the observed and expected numbers of different SGs was evaluated for the pre- and post-refinement SG sets. To clearly illustrate the issues with original signature gene sets and the changes occurring in each species between pre- and post-refinement SGs, *Buchnera aphidicola* (MSP54) will be used to exemplify the changes, after which summary statistics will be given for the improvement of all MSPs (Fig. 2). *B.aphidicola* was chosen because the original SG set exhibits some of the issues we are addressing in this article, namely large amounts of samples with a shifted distribution, indicating a heterogeneous ease of detection of SGs. MSP54 initially exhibited an MSE of 110.57, which after refinement was reduced to 2.13, showing a set of SGs that follow the expected distribution much closer after refinement. The MSP is mapped with at least three reads in 23 samples. Prior to refinement, 10 of these samples were accepted ($P > 0.05$, CCP), indicating that the amount of detected SGs is coherent to the number of reads mapped to the MSP. After refinement, 13 of the samples are accepted ($P > 0.05$, CCP), indicating that the replacement genes are more compliant with the probabilistic model. Across all MSPs, the MSE was evaluated for the initial and refined SG sets. We observe a decrease in MSE for 28/54 MSPs (Supplementary Fig. S1). A significant lowering of MSE is observed between the initial SG set and the refined SG set by a Wilcoxon signed rank test (P-value of 4.0e−06 paired).

To assess the degree of SG exchange, the fraction of original SGs retained and the ratio of MSE before and after SG refinement were compared (Supplementary Fig. S2). In MSP54, 20 out of the 100 initial signature genes were retained; hence, a large proportion of the initial SGs were exchanged in favour of other genes. Across all MSPs, we observed a correlation between the relative MSE ($MSE_{before} - MSE_{after}$) and the number of signature genes retained; however, no significant correlation was found. We observe the largest improvements in MSE in MSPs in which a large fraction of SGs have been replaced. The MSPs are having 38 genes replaced on average, with 18 of the MSPs replacing 75 or more of their SGs. Conversely, 10 MSPs change between 25 and 75 of its original SGs, while the remaining 26 MSPs experience no change in SGs.

A sample is rejected if the chance of obtaining d out of 95 signature genes given k reads assigned to signature genes is below 5% ($P < 0.05$, CCP). Samples that contain fewer than three reads ($k > 3$) assigned to the SGs were not considered as we are not confident in the detection of this metagenomic entity. Samples with fewer than three assigned reads were neither accepted nor rejected to avoid influence of samples that would otherwise not be considered for abundance measurement. For example, 23 samples were found to have three or more reads for MSP54 prior to refinement, 13 of which were significantly depleted in SGs. After refinement, 10 out of 23 samples with more than three reads assigned to SGs were significantly ($P < 0.05$, CCP) depleted in SGs. Across all MSPs, prior to refinement, a significant depletion ($P < 0.05$, CCP) in SGs was found in 18% of instances in which three or more of the initial SGs were mapped and were rejected. This rate is lowered after refinement, as 15% of instances were rejected.

Finally, we wish to evaluate the change in the number of samples with more than three reads assigned to SGs ($k > 3$), which we consider to be an indicator that the organism is present in the sample. There were concerns that a reduction in MSE could be achieved by selection of a set of SGs that were exclusively found in a rare strain but not present in the vast majority of samples. To assess this, the degree of change in mapped samples ($k > 3$) was evaluated. This number was correlated with the degree of change in MSE between initial and final SGs (Supplementary Fig. S3).

Of the 54 MSPs, only 11 of them display a change in number of mapped samples after the SG refinement. The average share of genes found across samples for the MSPs ranges from 0.69 to 0.99, indicating that once an MSP is identified within a sample, most of its SGs are detected and are thus estimated to be present in higher abundance. For MSP54, an average of 85 of the 100 initial SGs are identified in the samples, which are also seen in Figure 2C.

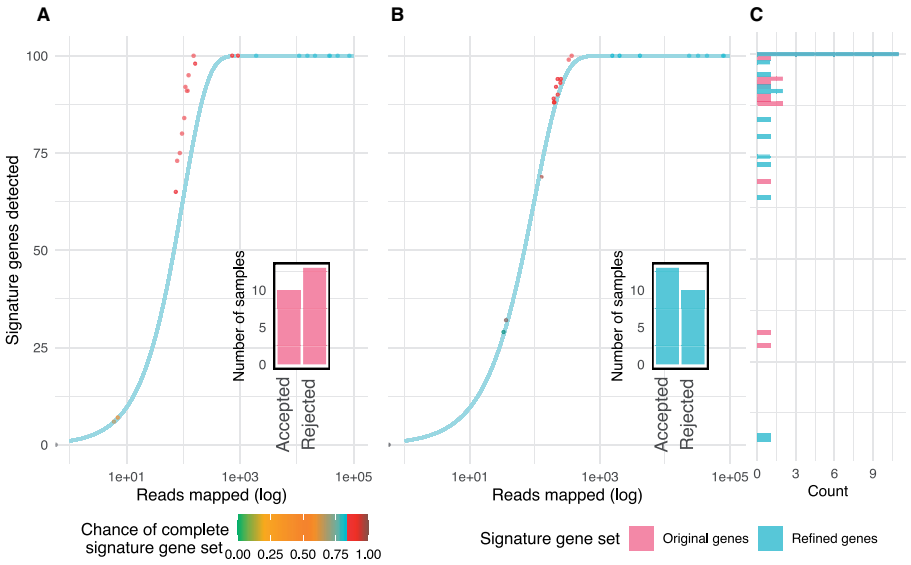


Fig. 2. Detection of signature genes. (**A** and **B**) Distributions of the number of different identified signature genes for a given number of reads mapped to signature genes for each sample for MSP54. Colours indicate the chance of this sample containing 95 unique signature genes as described in methods. The bar plot indicates the number of samples that were rejected ($P < 0.05$, CCP) and accepted ($P > 0.05$, CCP). The expected distribution of samples for a metagenomic entity which contains 100 SGs is indicated by a blue line. Panels (A) and (B) are for SGs prior to and after SG refinement, respectively. Panel (**C**) indicates the distribution of uniquely identified SGs across samples, red indicating pre-refinement and blue post-refinement

3.2 Relative abundance measures of the initial and refined signature genes

The sample-specific relative abundance is found by dividing the reads that map to the SGs by the total number of reads mapping to the SGs across the MSPs. MSPs with identical taxonomic annotations are collapsed into a single biological entity. The taxonomies are extracted per gene from the Golden Standard Single Assignment binning results from Borderes *et al.* (2021). If the SGs are assigned conflicting taxonomies, the most abundant taxonomy of the refined SG set is assigned to the MSP. Of the 47 genomes used to construct the SGC, 43 of the entities were represented by MSPs, yielding a precision of 0.81 and a recall of 0.91. The read counts of the genes are normalized according to the length of the genes. The error between the calculated relative abundances from the gene set of MSPminer and the refined SGs are computed (Fig. 3). The error in relative abundance tends to increase with an increase in the true relative abundance (Supplementary Fig. S4).

It was examined whether using the full gene set of the MSPs would yield better relative abundance estimates. When taking the reads that map to all of the genes comprised in the MSPs and comparing with the true relative abundance, the discrepancy is on average larger than for the initial SG set from MSPminer.

The relative abundance quantifications were compared with the true relative abundance by subtracting the true versus the calculated. The closer the abundance prediction is to the ground truth, the closer the value will be to zero. The differences in predicted and true abundance for both the initial SGs from MSPminer and the refined SGs the difference was evaluated (Fig. 4). The relative abundance of the refined SGs is found to be closest to the true abundance, when tested with a paired Wilcoxon signed rank test with a *P*-value of 2.2×10^{-16} and an effect size of 0.13. Especially the MSPs estimated by MSPminer to be in higher abundance than the ground truth are found in abundances closer to the ground truth after refinement.

3.3 Case study using Bäkhed's First Year of Life dataset

Bäkhed's dataset was used to demonstrate the applicability of SG in a real and well-structured metagenomics study. The study comprises samples obtained from infants and their mothers (Bäckhed *et al.*, 2015). The data were pre-processed followed by *de novo* assembly of contigs. Of the 392 samples, only 389 samples were successfully assembled after 20 days of runtime on our HPC (40 cores and 180 GB per job). The contigs were binned across samples with VAMB, and subsequent filtering (discarding bins <200 000 base pairs) resulted in 9763 bins from 2672 VAMB clusters. For reference, Bäkhed et al. identify 690 meta Operational Taxonomic Units (mOTUs). Further annotation of these clusters using GTDB-tk (v.2.1.1) was successful for 1843 clusters, where the original study annotated 373 species. Genes were predicted through GTDB-tk using Prodigal (v.2.6.3) (Hyatt *et al.*, 2010), 763 clusters were found in fewer than 3 samples or contained less than 100 genes and, consequently, the abundance was set to 0. For this dataset, an improvement in MSE between the SGs from VAMB and the refined SG's is obtained for 587 of the 1080 clusters, with an average improvement of 42.6 ± 27.2%. Cluster5004 (annotated as Parabacteroides distasonis) is one of the clusters, which displays a large improvement in MSE (Fig. 5A), from an MSE of 984.17 to 121.05. In the initial SG set, 6 samples had above 61 detected signature genes despite a mean number of reads mapping across the samples of 7581 ± 10 851.56. The detection of the initial SGs is shown as a heatmap, where the

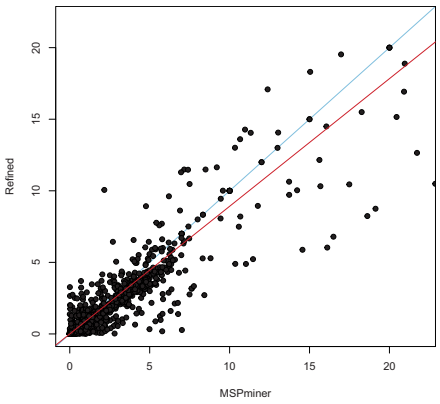


Fig. 3. Error in relative abundance of species-level taxa. Error from true abundance predicted by MSPminer and true abundance to the refined SGs. The error is calculated as the absolute difference from the predicted relative abundance and the true relative abundance. Each dot represents a species-level taxa. The red line indicates the linear relationship between the two methods. The blue line indicates the identity line ($y = x$). For visualization purposes the seven taxa with the largest discrepancy from the true relative abundance has not been included on the figure

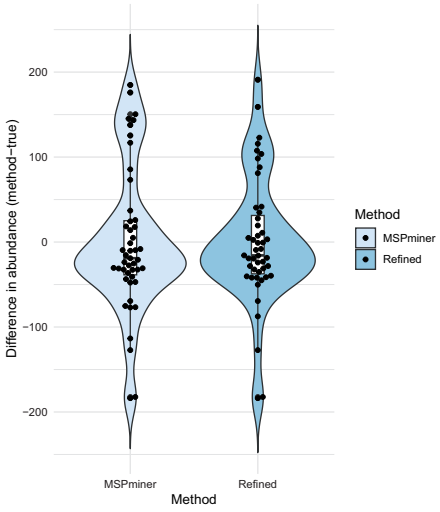


Fig. 4. Difference in relative abundance for the truth subtracted the abundance given from the initial SGs from MSPminer and the refined SGs. Combination violin and boxplot of the biological entities

genes and samples are grouped using hierarchical clustering (R Core Team, 2021) (Supplementary Fig. S5). The 6 samples with >61 genes detected cluster together to the left side of the figure. Fifty of the SGs are seen in less than 50 samples, despite the Cluster being present in 319 samples (Fig. 5B), with an average of detection of 111 reads per sample. For the refined SG, the average number of detections is 193 reads per sample (Fig. 5C).

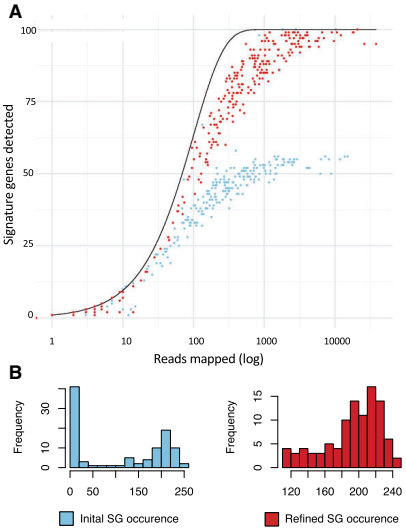


Fig. 5. Read counts of Cluster5004 from the case study using data from Bäkhed et al. (**A**) Distributions of detected SGs displayed as a function of number of reads mapped to the SGs per sample for the initial SG found with VAMB and the refined SG set. The expected detection of SGs given by the number of mapped reads is indicated by a black line [Equation (6)]. (**B**) Histogram of the detection of each SG for the initial and the refined SGs

A beta-diversity analysis of the relative abundances was carried out using Bray–Curtis distances and visualized with a principal coordinate analysis plot (Supplementary Fig. S6).

To demonstrate the use of SG for abundance measurements, a replication of their *Signature Taxa* from each stage of the study was performed (Fig. 6). The taxa from their Supplementary Table S5 were used for the creation of the heatmap. Of the 57 Signature Genera found for vaginally born infants, 37 were reidentified and annotated. For the infants born with C-section, 24 of 37 Signature Genera was reidentified. Three additional Signature Genera were found in our results; however, GTDB-tk had them classified as 'Family' level (Erysipelotrichaceae, Lachnospiraceae and Ruminococcaceae).

4 Discussion

We utilized a variant of the CCP as a theoretical framework to implement SG refinement. The CCP appears to be a good approximation for the majority of refined SG sets, although certain SGs with heterogeneous sensitivities do not follow the initial assumptions of this method, as the CCP assumes uniform probabilities for sampling.

Despite the prescreening and ordering of SG according to sensitivity, we still find certain refined SG sets where samples appear to detect numbers of SGs deviating from the expected, given the number of mapped reads, especially samples that have had an inflated number of identified SGs. This is in accordance with the findings of Borderes *et al.* (2021), where MSPminer is found to overestimate the number of binned genes of the SGC. This could be due to the SGC representing a simplistic and not necessarily representative version of the human gut microbiome, being unable to capture nature's variability, leading to an underrepresentation of cross-mapping of reads between species and genes being mapped more often than expected. We successfully implemented the SG method on the First Year of Life study, where we were able to reconstruct and annotate

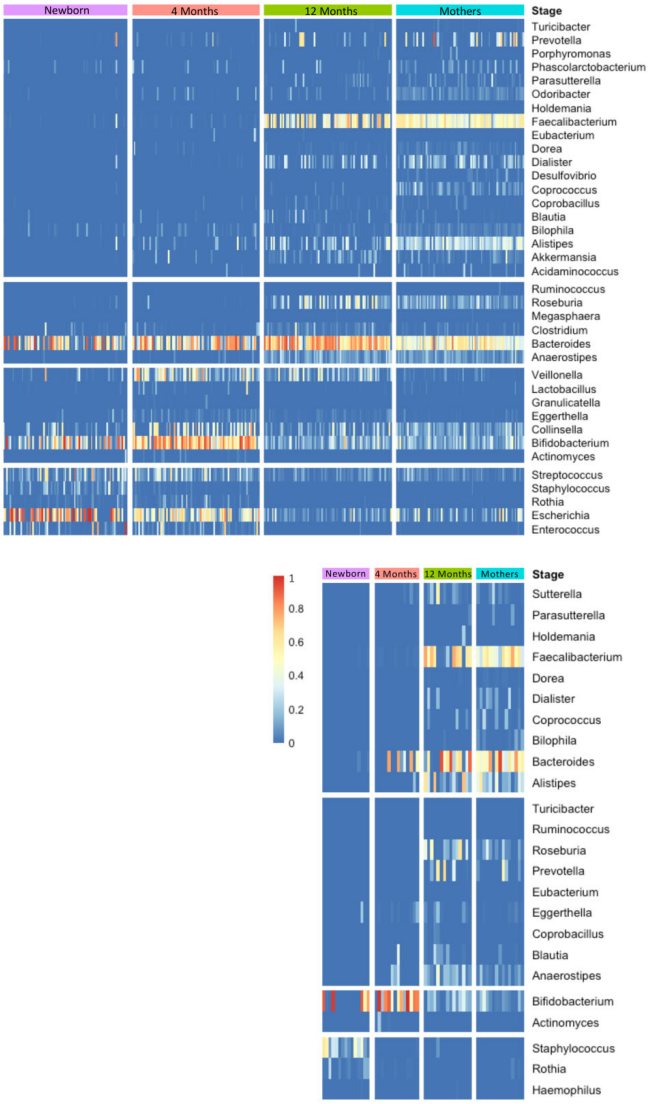


Fig. 6. Heatmap of relative abundance of the Signature Genera found by Bäkhed et al. in infants born (**A**) vaginally or (**B**) by C-section at the stages newborn, 4 months, 12 months, and their mothers. Columns indicate samples. The vertical coloured blocks indicate Signature Genera at each stage

1843 clusters, compared to the 373 mOTUs found in the original study.

While the two-step refinement appears to be very good at detecting genes with low detection rates, outlying SGs with very high detection frequencies were not removed, which could contribute to this problem. This is to some extent alleviated by pre-filtering of genes and could potentially be further alleviated by starting with an alternate set of initial SGs. No explicit criteria were given to select genes that had similar sensitivities for initial SG sets, as this would narrow the number of genes that could be searched to an extent that could end up hampering MSE improvement and complicate the assessment of improvement between refinement steps. However, for

Cluster5004 from The First Year of Life study, we were able to identify a suitable gene set, even when only 6 of the samples had >61 of the initial SGs detected. From the heatmap, it is clear that the six samples cluster and that there is a clear trend amongst which genes are consistently being detected. With half of the SGs having reads mapping in fewer than 50 out of 319 samples, this indicates that they are not part of the core genome of this cluster but are more likely strain specific. Or that the 6 samples having >61 detected SGs are part of a higher-level taxa than the rest. However, after the refinement, the SGs are detected more consistently across the samples.

In the absence of a large number of samples with adequate read counts, it is difficult to select suitable SGs for the metagenomic entity due to the fragmented nature of metagenomic count data. For rare species found in low abundance, the count data of the metagenomic entity will predominantly be zero inflated. The proposed model is not developed to explicitly deal with zero inflation. However, if the metagenomic entity is present in low abundance within a subpopulation of the samples, the model will utilize the information from the higher abundance samples to optimize the signature gene set. If the signature genes are truly specific for the species, they facilitate quantification even at a very low abundance.

We find that treating SG selection as a variant of the CCP allows us to identify SGs that are easier to detect uniformly across samples. We can identify genes that do not act as expected using an NB model, which allows us to replace these with genes that are more consistently found across samples. This leads to more reliable species identification and an improved abundance estimation, since abundance is less reliant in genes that are not likely to be present or seems to be oversampled in a majority of samples. When tested on the simulated data set, the refined SG sets were found to significantly improve the relative abundance estimates compared with the initial SG sets. The MSE between the distribution that one would expect from the model was reduced for approximately 52% of all sets of signature genes. The number of SGs identified in samples with significant ($P < 0.05$, CCP) depletion in SGs changed between pre- and post-refinement, rejecting fewer samples that otherwise showed a large representation of SGs, while still rejecting samples that had very few SGs in a given sample, which indicates the selection of a set of SGs that are more likely to be identified in unison. From the real dataset, it was clear that even in cases of low abundance species, the method was able to identify a set of SGs that are found more consistently across the samples.

Author contributions

Trine Zachariasen (Credit contribution not specified), Anders Østergaard Petersen (Credit contribution not specified), Asker Brejnrod (Credit contribution not specified), Gisle Alberg Vestergaard (Credit contribution not specified), Aron Eklund (Credit contribution not specified), and Henrik Bjørn Nielsen (Credit contribution not specified).

Funding

None declared.

Conflict of Interest: none declared.

References

Bäckhed,F. *et al.* (2015) Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe*, **17**, 690–703.

Borderes,M. *et al.* (2021) A comprehensive evaluation of binning methods to recover human gut microbial species from a non-redundant reference gene catalog. *NAR Genom. Bioinform.*, **3**, lqab009.

Chaumeil,P.-A. *et al.* (2022) GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics*, **38**, 5315–5316.

Fowler,J.C. (2015) Fowler J. pdf. In: *Attacking Variations of the Coupon Collectors Problem with Maple. 27th International Conference on Technology in Collegiate Mathematics*, pp. 1–9.

Hyatt,D. *et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.

Joshi,N.A. and Fass,J.N. (2011) Sickle*: A Sliding-Window, Adaptive, Quality-Based Trimming Tool for FastQ Files github*.

Kang,D.D. *et al.* (2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, **7**, e7359.

Kultima,J.R. *et al.* (2012) MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS One*, **7**, e47656.

Li,R. *et al.* (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.

McElroy,K.E. *et al.* (2012) GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics*, **13**, 74.

Milanese,A. *et al.* (2019) Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.*, **10**, 1014.

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

Nissen,J.N. *et al.* (2021) Improved metagenome binning and assembly using deep variational autoencoders. *Nat. Biotechnol.*, **39**, 555–560.

Plaza Oñate,F. *et al.* (2019) MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics*, **35**, 1544–1552.

R Core Team (2021) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sangwan,N. *et al.* (2016) Recovering complete and draft population genomes from metagenome datasets. *Microbiome*, **4**, 8.

Segata,N. *et al.* (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods*, **9**, 811–814.

Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.

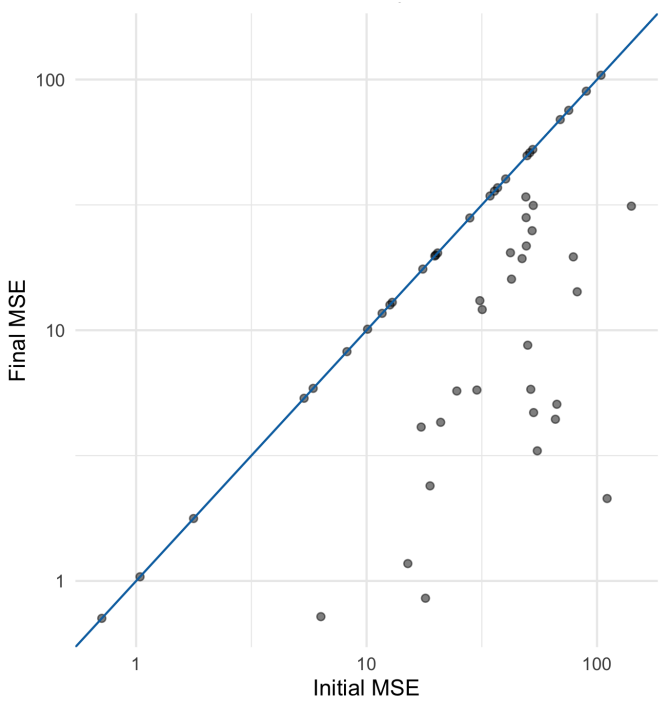
Vasimuddin,M. *et al.* (2019) Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, pp. 314–324.

Zhang,X. *et al.* (2017) Negative binomial mixed models for analyzing microbiome count data. *BMC Bioinformatics*, **18**, 4.

Supplementary Data for Identification of representative species-specific genes for abundance measurements

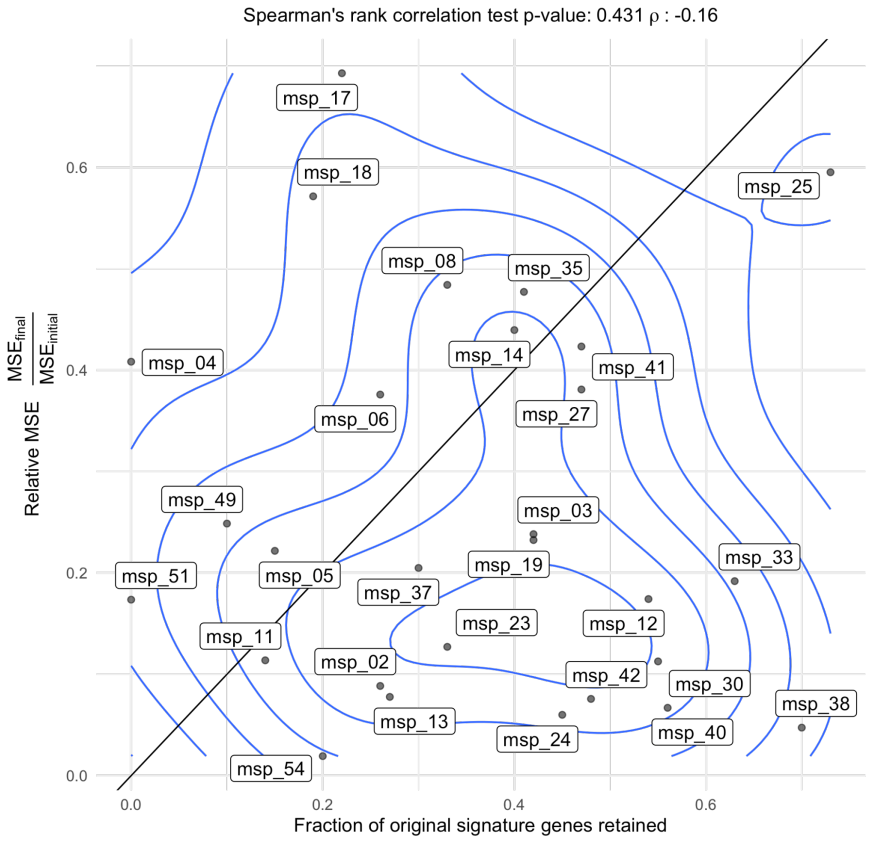
Trine Zachariasen, Anders Østergaard Petersen, Asker Brejnrod, Aron Eklund, Gisle Alberg Vestergaard and Henrik Bjørn Nielsen

Table of Contents



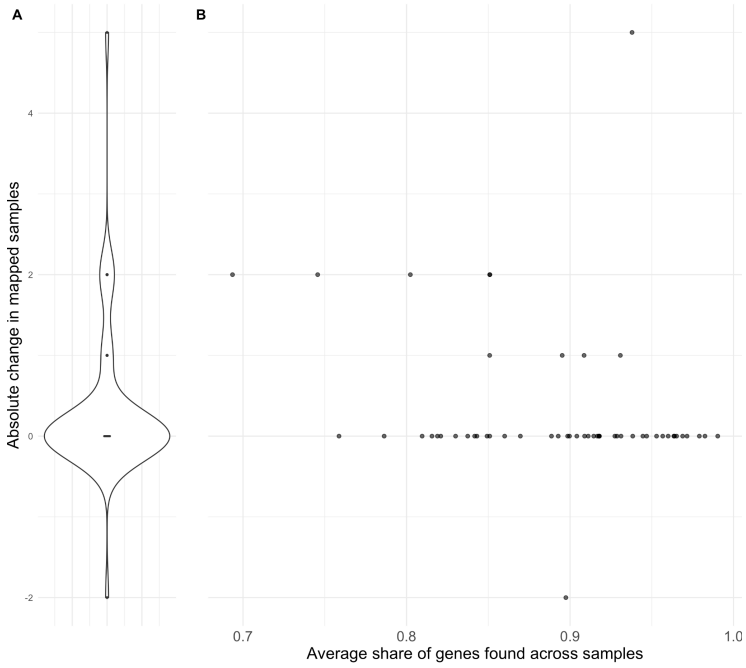
Supplementary Figure 1

Improvement in MSE from initial to refined SG sets. MSEs of the number of identified SGs, *d*, for a given number of reads assigned to SGs, *k*, before and after a two-step refinement of all 54 signature gene sets.



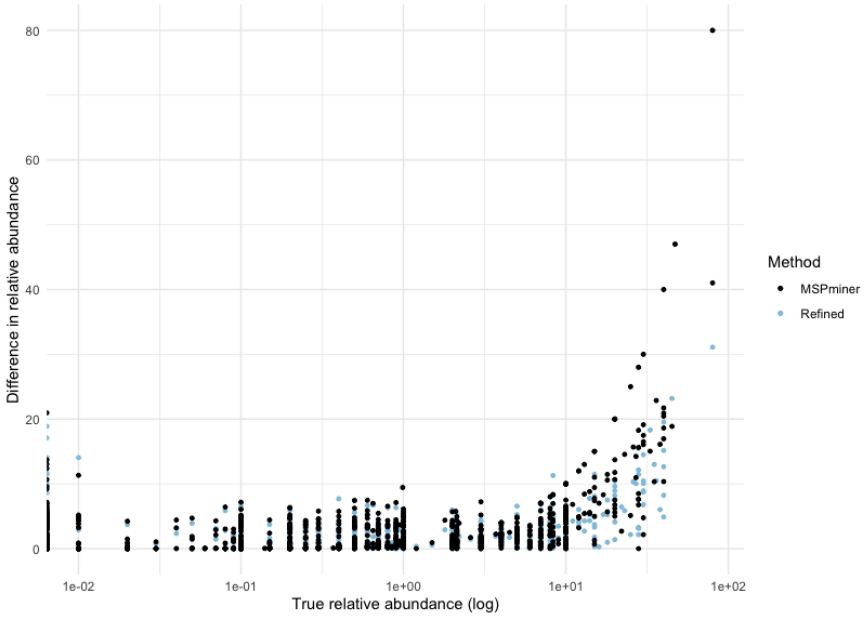
Supplementary Figure 2

Relative improvement in MSE as function of initial SG's retained. Illustrating relative MSEs and fraction of original signature genes kept throughout refinement. Each dot is a set of SGs that has undergone SG refinement. 26 of the MSPs are not having any SGs replaced and are thus not included in the figure.



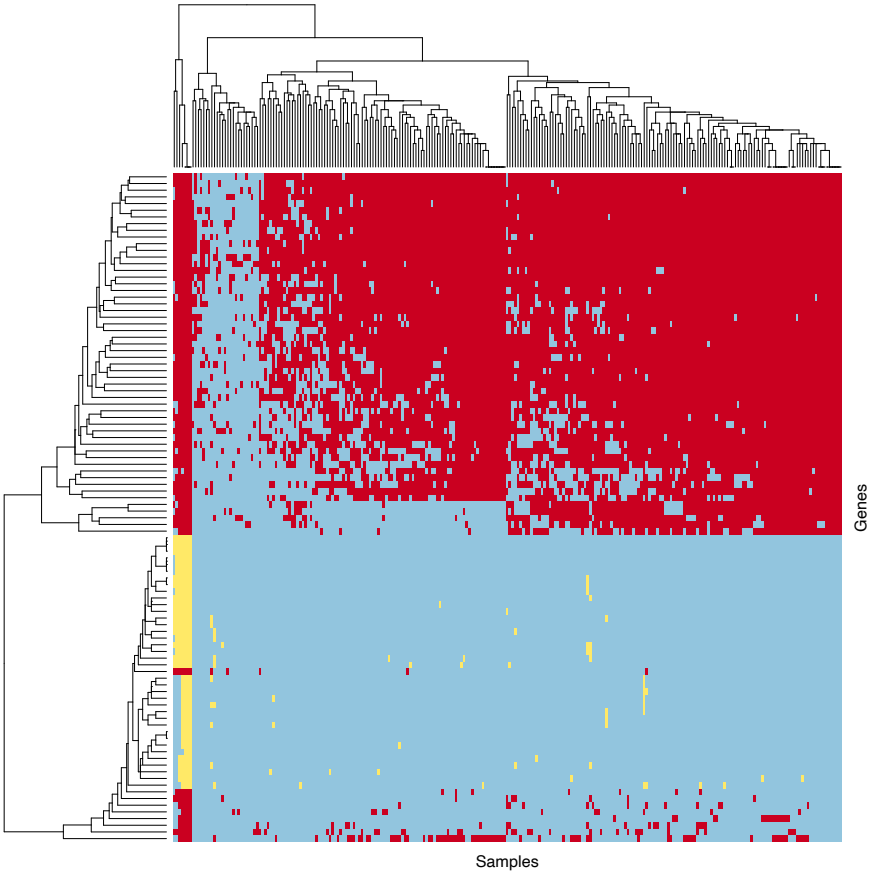
Supplementary Figure 3:

Change in number of mapped samples A) Combination violin and boxplot of relative number of mapped samples for all MSPs. B) Relative change in the number of mapped samples on the y-axis and degree of overlap between initial and final set of signature genes. A sample is considered mapped if 3 or more total read counts to SGs are observed.



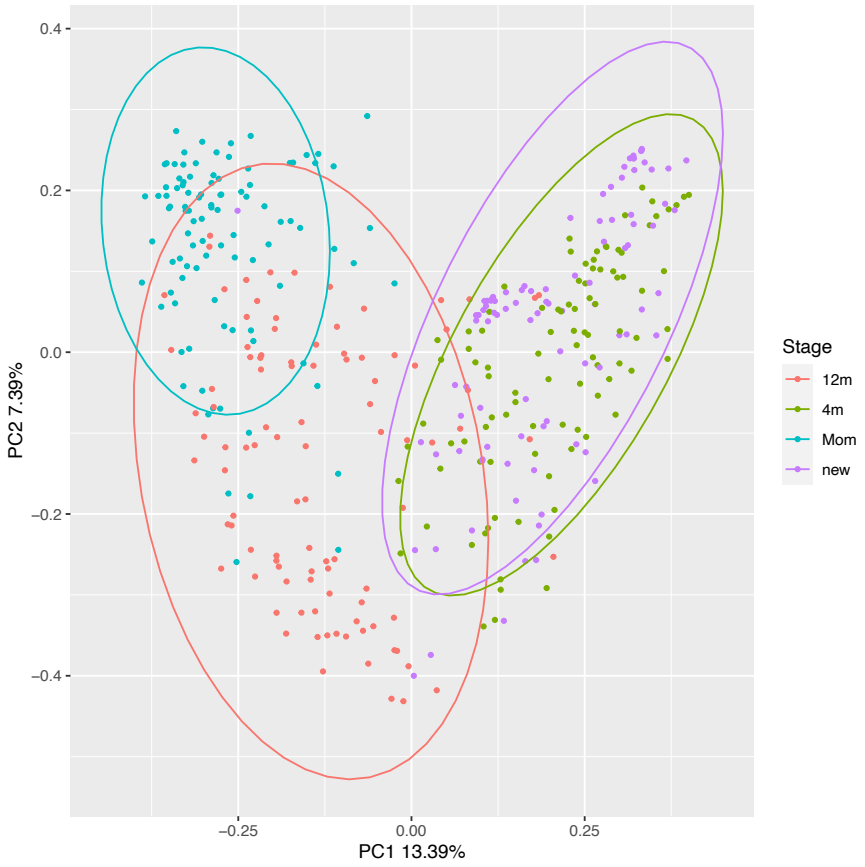
Supplementary Figure 4

Error in relative abundance given the calculated and true abundance of species-level taxa. The error of the calculated relative abundances given the true relative abundance for both the SGs from MSPminer and the refined set. The error is calculated as the absolute difference from the predicted relative abundance and the true relative abundance. Each dot represents a species-level taxa in a sample.



Supplementary Figure 5

Heatmap of the initial 100 Signature Genes from Cluster5004. Each row is a gene, and each column is a sample. The gene detection is binary (present/absent) in each sample. The samples and genes are clustered using hierarchical clustering.



Supplementary Figure 6

PCoA plots of beta diversity of the First Year of Life study using the relative abundance from the Signature Genes. PCoA was calculated with Bray-Curtis distances.

Paper II

MAGinator enables strain-level quantification of de novo MAGs

Zachariasen, T., Russel, J., Petersen, C., Vestergaard, G. A., Shah, S., Turvey, S., Sørensen, S.J., Lund, O., Stokholm, J., Brejnrod, A., Thorsen, J.

Submitted to Nature Biotechnology, 2023

The work carried out in relation to this thesis included project conceptualizing, software development and extensive and iterative testing and development. Data analysis and interpretation of several data sets has been done. As the main author I have been in charge of writing the first draft of the paper and as the corresponding author I have been in charge of the submission of the paper.

MAGinator enables strain-level quantification of *de novo* MAGs

Trine Zachariasen[1]*, Jakob Russel[2], Charisse Petersen[3], Gisle A. Vestergaard[1], Shiraz Shah[4], Stuart E. Turvey[3], Søren J. Sørensen[2], Ole Lund[1], Jakob Stokholm[2,4], Asker Brejnrod[1] and Jonathan Thorsen[4]

[1]Department of Health and Technology, Section of Bioinformatics, Technical University of Denmark, 2800 Lyngby, Denmark,

[2]Department of Biology, Section of Microbiology, University of Copenhagen, 2100 Copenhagen, Denmark

[3]Department of Pediatrics, BC Children's Hospital, University of British Columbia, 950 West 28th Avenue, Vancouver, BC V6H 3V4, Canada

[4]COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, 2820 Copenhagen, Denmark

*Corresponding author

Abstract

Motivation

Metagenomic sequencing has provided great advantages in the characterization of microbiomes, but currently available analysis tools lack the ability to combine strain-level taxonomic resolution and abundance estimation with functional profiling of assembled genomes. In order to define the microbiome and its associations with human health, improved tools are needed to enable comprehensive understanding of the microbial composition and elucidation of the phylogenetic and functional relationships between the microbes.

Results

Here, we present MAGinator, a freely available tool, tailored for the profiling of shotgun metagenomics datasets. MAGinator provides *de novo* identification of subspecies-level microbes and accurate abundance estimates of metagenome-assembled genomes (MAGs). MAGinator utilises the information from both gene- and contig-based methods yielding insight into both taxonomic profiles and the origin of genes as well as genetic content, used for inference of functional content of each sample by host organism. Additionally, MAGinator facilitates the reconstruction of phylogenetic relationships between the MAGs, providing a framework to identify clade-level differences within subspecies MAGs.

Availability and implementation: MAGinator is available as a Python module at https://github.com/Russel88/MAGinator

Contact: Trine Zachariasen, trine_zachariasen@hotmail.com

Introduction

DNA sequencing has revolutionised our ability to gain insight into microbial compositions without relying on the ability to cultivate organisms. To explore these compositions various methods have been developed that either rely on databases of marker genes of known organisms or attempt to reconstruct the chromosomes directly from the short reads by first assembling into longer contigs and then binning these based on co-occurrences or DNA composition.

Mapping reads against marker gene databases with tools such as MetaPhlAn[1], MetaPhyler[2] and mOTUs[3] is a fast and effective way of recovering the microbial composition both because the library depth required can be quite shallow and because the computational requirements are smaller, but have limitations originating from the reliance on predefined databases, limited ability to estimate abundances at higher taxonomic resolution[4,5], and the lack of information on the functional repertoire of the identified taxa. Conversely, *de novo* binning strategies require high sequencing depth but can recover high-quality metagenome assembled genomes (MAGs) from which the functional gene content can be directly linked to a specific organism. Ideally, this can recover genomes of strains that can be used in downstream analysis to generate more specific hypotheses about associations with outcomes. One example of this is the capacity of an organism to break down Human Milk Oligosaccharides (HMOs), the main source of energy for the developing infant gut microbiome while being breastfed. Especially *Bifidobacteria* have this functionality, and it is known that certain strains or subspecies have specific preferences for certain HMO types[6–9], improving the overall utilisation of HMOs and often conferring additional benefits as a probiotic. Previously, it has been established that specifically the presence of *Bifidobacterium longum* subspecies *infantis* (*B. infantis*) together with breastfeeding, plays a crucial role in providing a protective effect to mitigate the impact of antibiotics on the early-life gut microbiome[7]. This underlines the significance of being able to accurately profile the microbiome at higher resolutions than species-level.

In this work we have developed a pipeline that takes MAGs and original reads as input and generates output including accurate abundance estimates, strain phylogenies and gene synteny clusters that can improve insights into the microbiome composition (Figure 1). We do this by grouping MAGs into clusters that are phylogenetically separated at a higher resolution than species and estimate the abundances of these. This is done by identifying a set of signature genes directly from the given data and refining them according to statistical modelling to pick the ideal set suitable for abundance estimation. The fidelity of our estimated abundances are demonstrated on the Critical Assessment of Metagenome Interpretation (CAMI) strain-madness dataset, where we benchmark MAGinator against similar tools. Additionally we show the functionality of MAGinator on a public dataset of inflammatory bowel disease (IBD) patients, where we identify differentially abundant taxa between patients and controls at high phylogenetic resolution.

MAGinator also enables Single Nucleotide Variant (SNV's) resolution phylogenetic trees, which are created from the signature genes and used for additional stratification of the MAGs and can be associated with metadata to obtain subspecies/strain-level differences. We exhibit MAGinator's ability to obtain strain-level resolutions for *Bifidobacterium* from two real-world infant datasets. In this case the signature genes were found *de novo* for one dataset and were then utilised to obtain strain-level resolution in the other cohort.

By combining the information from both contigs and gene content we identify synteny clusters of genes within strains, yielding information on shared pathways for the genes. Additionally, we show how we can associate the functional content to the identified clades, to improve hypotheses-generation on the impact of organisms, illustrated using the $COPSAC_{2010}$ cohort.

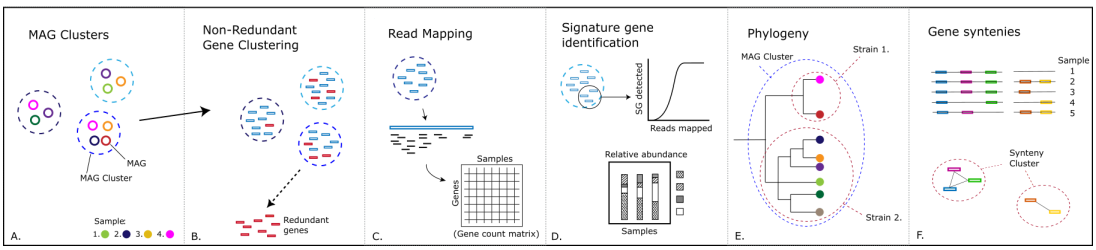


Figure 1: Schematic visualisation of the main functions of the MAGinator workflow.

Methods

Implementation

Input

The input to the MAGinator workflow comprises a set of samples with (1) shotgun metagenomic sequenced reads, (2) their sample-wise assembled contigs, and (3) sample-wise MAGs (groups of contigs from the same genome), clustered across samples, as defined by a metagenomic binning tool (see below).

Reads should be provided in a comma-separated file giving the location of the fastq files and formatted as: SampleName,PathToForwardReads,PathToReverseReads. The contigs should be nucleotide sequences in FASTA format. The MAGs should be given as a tab-separated file including the MAG identifier and contig identifier. The sample-wise MAGs should be grouped into MAG clusters representing a taxonomic entity found across the samples, which will usually be species but can also be at the subspecies level, depending on characteristics of the input data. MAGinator is flexible regarding which tool is being used for creating the MAGs, however we recommend using VAMB[10].

Dependencies

The dependencies to run MAGinator are mamba[11] and Snakemake[12] - all other dependencies are installed automatically by Snakemake through MAGinator. Additionally MAGinator needs the GTDB-tk database downloaded for taxonomic annotation of MAGs and as a reference for the phylogenetic SNV-level analysis of the signature genes.

Output generated

MAGinator generates multiple outputs and intermediate files useful for additional downstream analysis (Suppl. Table 1, Suppl. Figure 1). Importantly, MAGinator outputs the taxonomy of the MAGs, the signature genes of the MAG clusters, the sample-wise relative abundances of the MAG clusters, a non-redundant gene matrix with sample-wise mapping counts, synteny clusters and inferred phylogenies for each MAG cluster. Additionally, a folder is created containing the log information of all the jobs run by Snakemake.

Application

MAGinator is written in Python 3 and is based on a set of Snakemake[12] workflows, and easily scalable to work for both single servers and compute clusters. MAGinator is implemented as a python package and is available on GitHub at https://github.com/Russel88/MAGinator.

The MAGs are filtered based on a minimum size for inclusion, with a default size of 200,000bp. The included MAGs are taxonomically annotated using GTDB-tk (v.2.1.1)[13], by calling genes using Prodigal (v.2.6.3)[14], identifying GTDB marker genes and placing them in a reference tree. As the taxonomic annotation of the MAG clusters are found to be redundant, clusters with the same taxonomic assignment can be combined into one cluster, with the flag '--mgs_collections' which we identify as a Metagenomic Species (MGS). Redundant genes are identified by clustering with MMseqs2 (v.13.45111)[15] easy-linclust using a default clustering-coverage and sequence identity threshold of 0.8, creating a list of the representative genes along with their cluster-members. The redundant genes are filtered away, leaving a nonredundant gene catalogue. The raw reads are mapped to the gene catalogue using BWA mem2 (v.2.2.1)[16] and counted using Samtools (v.1.10)[17], leaving a gene count matrix, which is used as input for the signature gene refinement and following phylogenetic clade separation and abundance estimates.

Signature Gene Identification

We previously described the method for identifying the signature genes for the data set[18]. In brief, signature genes are selected to ensure that they 1) are unique for the MAG cluster, 2) are present in all members of the cluster, and 3) are single-copy.

To accomplish this the following steps are taken: Initially the non-redundant gene count matrix is curated to discard any genes if they have (redundant) cluster-members originating from more than one MAG cluster, as they are thus not specific for that biological entity. Subsequently, the remaining genes within each MAG cluster are sorted based on their co-abundance correlation across the samples. As the genes are unique for the species, if they are consistently detected in similar abundance across samples, it suggests that they are single-copy. This step also mitigates differences in read mappings caused by biological or technical variations. The initial set of signature genes for each biological entity are selected from the most correlated genes. Subsequently, these signature genes are further refined and optimised by fitting them to a rank-based negative binomial model that captures the characteristics of the specific microbial composition in the input data. The signature gene set

is evaluated across the samples, by calculating the probability of the detected number of signature genes given the number of reads mapping to the MAG cluster. Finally the abundance of each MAG cluster is derived from the read counts to the identified signature genes normalised according to the gene lengths.

SNV-level resolution phylogenetic trees

To elucidate the smaller biological differences within the MAG clusters, MAGinator will infer a phylogeny based on the sequences of the signature genes. Based on the read mappings to the signature genes the sample-specific SNVs are called using output from Samtools mpileup. An alignment for each signature gene is made for all samples containing the signature genes using MAFFT (v.7)[19] run with the offset value of 0.123 as no long indels are expected. MAGinator allows phylogenetic inference to be calculated with either the fast method Fast-Tree (v.2)[20] (default) or the more accurate but resource intensive method IQ-TREE (v.2)[21] (--phylo ['fasttree', 'iqtree']). In samples where no MAG was found, the phylogenies can be used to detect rare subspecies-level entities based on just a few reads mapping to the signature genes and to infer functions and genes from closely related MAGs from other samples. The criteria for inclusion in the tree can be adjusted by the user. For a sample to be included in the phylogeny the following three criteria has to be met 1) minimum fraction of non-N characters in the alignment (default –min_nonN=0.5), 2) minimum number of GTDB marker genes to be detected (default –min_marker_genes=2), 3) minimum number of signature genes to be detected (default --min_signature_genes=50). The trees can be associated with metadata to obtain clade-level differences associated with study design variables such as disease phenotype, sampling location, or environmental factors.

Gene synteny

Based on the gene clustering with MMSeqs2 a weighted graph is created, which reflects the adjacency of the genes on contigs. If genes are close enough in the graph they will be categorised as part of the same synteny cluster and it is assumed that they have related functionality and/or are part of the same functional module. Clustering is determined using mcl (v.14)[22], where the user has the options to influence the adjacency count and stringency of the clusters. Only immediate adjacency is considered. By default, genes found adjacent just once are included in the graph, but this can be tuned to make more strict clusters (default –synteny_adj_cutoff=1). The inflation parameter for mcl-clustering of the synteny graph are

important for the size of the gene clusters and are by default set high in order to small and consistent clusters (default –synteny_mcl_inflation=5).

Taxonomic scope of gene clusters

The taxonomic assignment of the sample-specific MAG is done using GTDB-tk. In some cases it will not be possible to assign a taxonomy to the MAG, which could be due to contamination, the MAG originating from a currently undescribed organism or due to too little information found in the MAG. In these cases an alternative is to assign the gene clusters, found in the MAG, a taxonomy. The taxonomic scope of the genes are described for the category they are almost all found in, given by a fraction defined by the user (default –tax_scope_threshold=0.9). E.g. if run with default options and a gene cluster has the assignment *"Bacteria Firmicutes_A Clostridia Lachnospirales Lachnospiraceae Anaerostipes NA"*, then at least 90% of the genes should be found in *Anaerostipes*. The algorithm will find the most specific taxonomic rank which has at least 90% agreement across the genes in the cluster assigned by GTDB-tk.

Workflow design

The MAGinator workflow has been constructed to make the information flow between the different modules automatically (Suppl. Figure 1).

The data goes through a series of filtering and processing steps (Figure 1), including:

A: Input MAG clusters, which are composed of one or more MAGs.

B: The genes are clustered and redundant genes are removed.

C: Reads are mapped to the genes, creating a gene count matrix.

D: Signature genes are identified for each MAG cluster, and used for abundance estimations

E: Based on the signature genes, SNV-level resolution phylogenetic trees are created and the taxonomic scope of gene clusters are identified.

F: Synteny-clusters of genes are identified, reflecting the adjacency of the genes on the contigs.

Benchmarking with OPAL on CAMI's stimulated strain-madness data set

The construction of the strain-madness benchmarking dataset was part of the second round of CAMI challenges[5]. The data consists of 100 simulated metagenomics samples consisting of paired-end short reads of 150 bp. The samples were run through a preprocessing workflow prior to the analysis. This involved the removal of adapters with BBDuk (v. 38.96

http://jgi.doe.gov/data-and-tools/bb-tools/) run with the following settings 'ktrim=r k=23 mink=11 hdist=1 hdist2=0 ptpe tbo', removal of low-quality and short reads (<75 base pairs) with Sickle (v. 1.33)[23] and removal of human contamination (reference version: UCSC hg19, GRCh37.p13) using BBmap (http://jgi.doe.gov/data-and-tools/bb-tools/) leaving an average of 6.6 million reads (SD: ±2802 reads) per sample.

To generate *de novo* assemblies, Spades (v. 3.15.5)[24] was utilised with the -meta option, with kmer sizes of 21, 33, 55 and 77, and contigs shorter than 1500 bp being discarded. Read-to-assembly mapping was carried out using BWA-mem2 (v.2.2.1)[16] and SAMTOOLS (v.1.10)[17]. Contig depths were assessed using Metabat2's jgi_summarize_bam_contig_depths (v.2.12)[25], while contigs were binned into MAGs using VAMB (v.3.0.8)[10] using default settings.

The reads, contigs and MAGs were run through the MAGinator workflow (v.0.1.16). For comparison purposes the VAMB clusters were annotated with a NCBI Taxonomy ID using CAMITAX[26]. The profile was created with Python 3 and the lineage found using NCBI's lineage taxonomy (https://ftp.ncbi.nlm.nih.gov/pub/taxonomy/new_taxdump/, accessed May 9th 2023). As the strain-identifiers from the gold-standard does not exist in the NCBI database (e.g. 1313.1), we have assigned an extra number to the Taxonomy ID for the clusters which had the same species-level annotation, starting at 1 to the number of redundantly annotated clusters.

The data for the benchmarking was obtained from CAMI second challenge evaluation of profiles. The profiles used for the benchmarking in this study were selected based on the best-performing tools found in the CAMI II paper. The top 10 profiles comprise DUDes[27] (v.0.08), LSHVec[28], MetaPhlAn2[29] (v.2.9.22), MetaPhyler[2] (v.1.25), mOTUs[3] (v.2.0.1 and v.2.5.1) and TIPP[30](v.4.3.10). The profiles were compared using OPAL, which was run with default settings.

Franzosa et al. reanalysis

Processed taxa and metadata tables were obtained from the Franzosa et al.[31] supplementary materials. Raw data were downloaded from ENA using the provided accessions, and run through the preprocessing, assembly and binning before running the entire MAGinator pipeline. Four samples failed the assembly (PRISM|7238, PRISM|7445, PRISM|7947,

PRISM|8550) and were excluded from all downstream analysis, both in the original and the MAGinator processed tables.

Statistical methods for abundance matrices

Abundance matrices were analysed in R (v.4.1.2). Sample management and beta diversity calculations were done in {phyloseq}[32], along with PCoA analysis. Differential abundance testing was done with the {DAtest} R package which uses the Wilcoxon test function (wilcox.test) from the {stats} package, with p-values adjusted by Benjamini-Hochberg false discovery rate correction. Corrected p-values less than 0.05 were considered significant.

Subspecies resolution of *Bifidobacterium longum*

COPSAC dataset - data characteristics and preparation

The $COPSAC_{2010}$ cohort consists of 700 unselected children recruited during pregnancy week 24 and followed closely throughout childhood with extensive sample collection, exposure assessments and longitudinal clinical phenotyping[33–35]. From the cohort, we used 662 deeply sequenced metagenomics samples taken at 1 year of age. The details of the study and sequencing protocol have previously been published[35]. The samples consist of 150-bp paired-end reads per with mean ± SD: 48 ± 15.5 million reads.

The data was analysed using the same approach as for the strain-madness data set, with the exception of filtering away reads shorter than 50 bp in the preprocessing step. This workflow yielded 880 MAG clusters for the samples.

MAGinator was run using the reads, contigs and MAGs from VAMB as input. Thus creating a set of signature genes for each MAG cluster which has been found *de novo* for this particular dataset.

CHILD dataset - data characteristics and preparation

The Canadian Healthy Infant Longitudinal Development (CHILD) study comprises a large longitudinal birth cohort with stool collection in infancy for microbiome analysis[36]. Stool samples used in this analysis were sequenced to an average depth of 4.85 million reads (SD: 1.79 million), and samples which included >1 million reads after preprocessing were kept for the current analysis[7].

We analysed a subset of the CHILD cohort, consisting of 2846 metagenomic sequenced faecal samples from infants. To overcome the shallow sequencing, the signature genes of the $COPSAC_{2010}$ cohort were used to profile the samples instead of running MAGinator. To ensure that the process of the read mappings was identical to COPSAC, the read mapping was carried out using the full gene catalogue. Next the read counts for the signature genes were extracted and used to derive sample-wise abundances for each MAG cluster.

Examining Bifidobacterium MAG clusters

The detection of signature genes for *B. infantis* for the $COPSAC_{2010}$ and CHILD cohorts was carried out by creating a binary detection matrix and using the standard function (heatmap) with default values in R. Furthermore, we compared the abundances of all the *Bifidobacterium* MAG clusters derived from MAGinator with abundance estimates from Metaphlan 3 (v.3.0.7) and strain phylogenies from Strainphlan 3 (v.3.0.7) for the species *Bifidobacterium longum*. The phylogenetic tree output by Strainphlan was converted into a distance matrix and clustered using partitioning around medoids into two clusters. The two clusters were annotated as *B. longum* subsp. *longum* (*B. longum*) and *B. infantis* based on the placement of *Bifidobacterium longum* reference genomes in the phylogenetic tree.

SNV-level phylogenetic trees for COPSAC dataset

For each MAG cluster the sequences of the signature genes were used as a reference to create an SNV-level phylogenetic tree. The trees for $COPSAC_{2010}$ were constructed with the default values of MAGinator, producing a tree in Newick file format and creating statistics for the alignment. The tree for *Faecalibacterium* sp900758465 was visualised in R using {ggtree}[37].

Gene syntenies and functional annotation for COPSAC dataset

The non-redundant genes were annotated using eggNOG mapper (v.2.0.2)[38–40] . Of the 14.7 million non-redundant genes 9.2 million were annotated. The visualisation of the synteny clusters was done with {igraph}[41].

Results

MAGinator can accurately detect strains in simulated data

The performance of MAGinator was evaluated against the top 10 taxonomic profiles found in the second round of CAMI[5] challenges using the simulated short-read ‘strain-madness’

dataset. This dataset has been selected as it represents a heterogeneous strain environment, making strain and species detection highly relevant.
Running the MAGinator pipeline on the strain-madness data, 73 MAG clusters were identified, of these 22 clusters were present with less than 3 reads in 3 samples, so the abundance was set to 0. Of these 51 remaining entities, 30 were assigned with strain-level annotation by CAMITAX.

The profiles have been compared with the Open-community Profiling Assessment tooL (OPAL)[42] (Figure 2). For the majority of the tools, the performance decreased as the taxonomic categories became less inclusive (Figure 2B & Suppl. Figure 2). The L1 norm measures the total error from the predicted and true abundance at each rank. From genus to species-level we observed drops in the average completeness 82.7-45.6% and the average purity 73.6-36.5%. MAGinator had the best average completeness at genus (99.8%) and species-levels (89.6%) (Suppl. Table 2). At the genus-level MAGinator ranked number 5 for purity at 92.4% and the best-performing tool for the species-level at 90.1%. The LSHVec gsa had the best performance for purity at genus-level with 100% however at species-level it has a purity of 37.5%, ranking number 5 in this group (Suppl. Table 3).

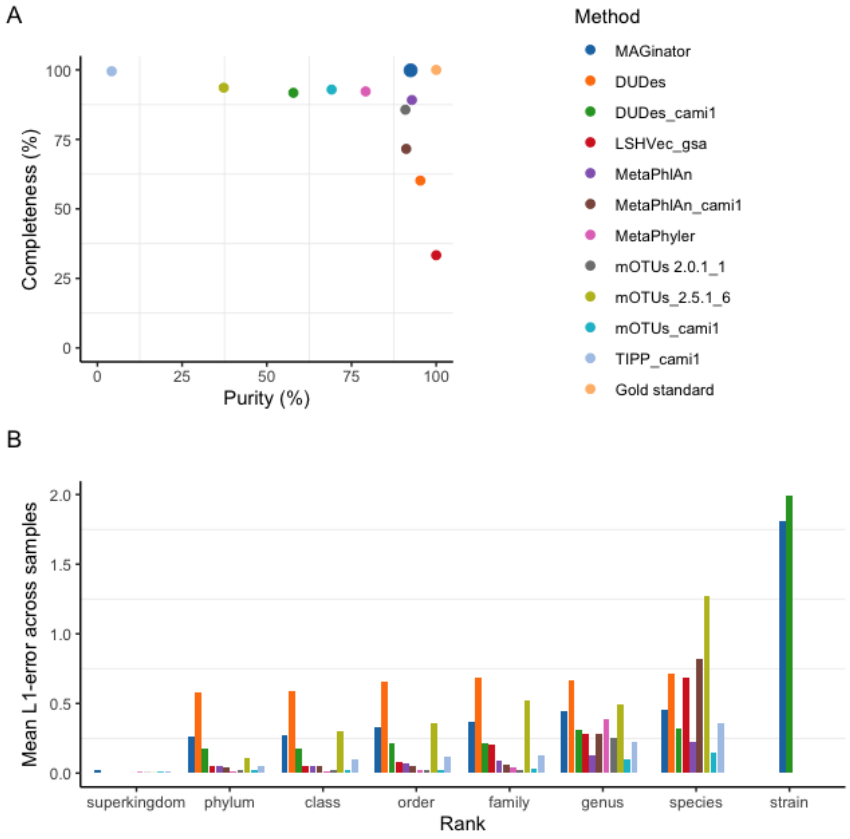


Figure 2: Benchmark using OPAL for comparing taxonomic profiling results for the CAMI strain-madness data set*. (A) Purity and completeness of the profiles are shown at genus-level (B) Mean of L1 norm error across samples for all ranks.*

MAGinator improves detection of relevant differentially abundant organisms

To demonstrate the advantages of quantifying bacterial taxa at high resolutions we have re-analysed a well-designed metagenomics study from Franzosa et al[31]. We chose this because it has deep sequencing well-suited for *de novo* MAG construction and a discovery/replication design with two distinct cohorts. In the absence of ground truth, replicating discoveries is a compelling strategy for making sure that findings are not false discoveries.

Beta diversity analysis of the two abundance matrices (MAGinator vs. their matrix created using MetaPhlAn2) revealed a similar separation for IBD patients vs healthy controls. For this study MAGinator produces abundance matrices of much higher dimensionality (2140 vs 201 taxa) because of the higher resolution in taxa identifications, therefore prevalence and/or abundance filtering might be relevant in MAGinator produced tables for noise reduction (Figure 3A-C).

To illustrate the improved ability of MAGinator to identify differentially abundant taxa we performed a regular differential abundance (DA) hypothesis test with Wilcoxon's test (Figure 3D-F). We looked for differentially abundant taxa defined as significant in the discovery cohort and replicated in the independent validation cohort. In the original analysis, 18 taxa were successfully validated in the independent cohort. With MAGinator, this increased to 213 taxa (Figure3 D-F).

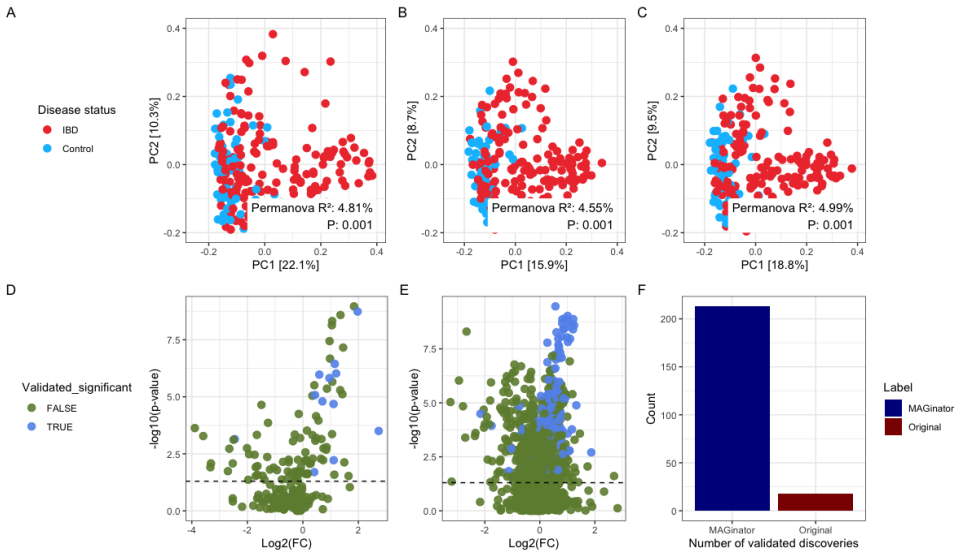


Figure 3: IBD case study shows similar performance of MAGinator with beta diversity and improvements in DA analysis. *PCoA and PERMANOVA (999 permutations) for beta diversity analysis with jsd distances and wilcoxon's test for differential abundance analysis. (A) PCoA of the original Franzosa et al. data (B) MAGinator abundances (C) filtered MAGinator abundances showing similar separation of IBD and control samples. (D) DA analysis of Franzosa et al. data, green points are taxa not significant in both cohorts (E)*

similar analysis on MAGinator abundances (F) Summary of validated discoveries using the two methods.

MAGinator enables tracking of strains across datasets at a high resolution

B. infantis is a gut microbe particularly adapted to the infant gut due to its ability to metabolise HMOs, which are complex sugars that infants cannot metabolise themselves[43]. These capabilities are different from other major subspecies including *B. longum* . Early-life colonisation with *B. infantis* has been linked to beneficial health outcomes which has sparked interest in its potential as a health-promoting infant probiotic which may even contribute to protection from asthma[7,44]. To demonstrate the utility of subspecies abundance estimation in MAGinator, we identified the signature gene set from one deeply sequenced infant cohort ($COPSAC_{2010}$) and used it to track subspecies abundances on another infant cohort (CHILD) with shallower sequencing but more samples. In the MAGinator pipeline, we identified two MAG clusters; one annotated as *B. infantis* and one as *B. longum* with GTDB-tk. In MetaPhlAn output we identified only an overall abundance for the species *Bifidobacterium longum*. Correlation analysis of these abundances shows that summed abundances of the *B. infantis* and *B. longum* MAG clusters explain 87% of the variance in the MetaPhlAn *B. longum* species (Suppl. Figure 3). In addition, we analysed the samples from both cohorts with StrainPhlAn[45] which detects strains in samples using prespecified species-level marker genes. Here, clustering of the sample-wise consensus sequences of the *B. longum* marker genes identified two clusters, one which clustered with reference strains of *B. longum* and one which clustered with reference strains of *B. infantis*. This result was previously shown for the CHILD cohort[7] and here we found similar results for $COPSAC_{2010}$ (Suppl. Figure 4). We hypothesised that this apparent duality may actually represent the underlying balance of these two subspecies in each sample. We confirmed this by comparing the StrainPhlAn-clusters with the MAGinator relative abundances of all *Bifidobacterium* species, where we saw that the StrainPhlAn clusters depended on the ratio of *B. infantis* to *B. longum* (Figure 4), but that more detailed information was accessible using the MAGinator derived relative abundances of each subspecies. This is an example of how *de novo* identification of subspecies-level MAG clusters and subsequent refinement of signature genes allows a higher resolution depiction of taxa for which the sequence coverage is sufficient in a given set of samples.

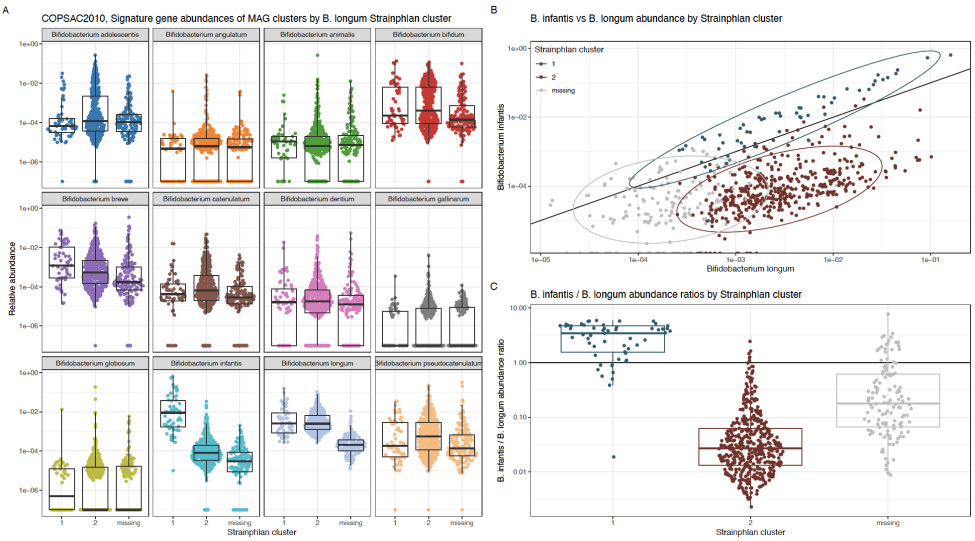


Figure 4: Stratification of StrainPhlAn clusters using the relative abundances of Bifidobacterium longum subspecies from MAGinator *Cluster 1 indicates B. infantis and Cluster 2 indicates B. longum.*
(A) Relative abundance of StrainPhlAn clusters stratified by all Bifidobacterium clusters identified by MAGinator (B) Relative abundance of B. infantis and B. longum identified with MAGinator coloured by StrainPhlAn cluster. (C) The ratio of B. infantis to B. longum is displayed for the StrainPhlAn clusters.

Additionally we used the signature genes identified from the COPSAC cohort to track the two subspecies in the CHILD cohort. The relative abundances of the MAGinator clusters and the StrainPhlAn clusters was likewise examined (Suppl. Figure 5). When using the signature genes as a reference for the CHILD cohort MAGinator was still able to resolve the two subspecies into more well-defined clusters yielding detailed profiling of the samples.

In order to estimate the fit of the signature genes for the two cohorts we compared the read mappings and presence of signature genes (Suppl. Figure 6A). As previously described by us[18] the expected number of detected signature genes within a sample can be calculated from the number of reads that map to those genes using a negative binomial distribution. We find that the $COPSAC_{2010}$ cohort deviates with a mean squared error (MSE) of 103.95, whereas the CHILD cohort deviates with a MSE of 878.09, indicating that the signature genes are

better suited for profiling of the specific strains found in the COPSAC cohort. To examine the cause of this large deviation for CHILD we created a heatmap of the read mappings to the signature genes (Suppl. Figure 6B). In accordance with Suppl. Figure 6A the samples cluster into two groups, which could be due to strain-differences. Additionally the genes are seen to cluster into multiple groups, wherefrom a group is seen to be absent in a large proportion of the samples, indicating that these genes have not been adequately selected for this strain for this dataset.

MAGinator provides SNV-level phylogenetic trees for each MAG cluster

By using the sequences of the signature genes as a reference it is possible to create a SNV-level phylogenetic tree of the samples, thus even being able to include samples in the tree, which do not contain enough reads to contain a MAG. For the MAG cluster *Faecalibacterium* sp900758465 we identified MAGs in 85 samples. For the tree 13 additional samples were included (Suppl. Figure 7), since these samples met the inclusion criteria as described in methods.

MAGinator identifies synteny clusters used for inference of functions

Genes can be grouped into synteny clusters based on their genomic adjacency. Genes close to each other in the genome will be grouped into a synteny cluster, and they are usually part of the same pathway or have a related function. Part of the MAGinator workflow creates these synteny clusters. For the $COPSAC_{2010}$ cohort 746,251 synteny clusters were identified with an average of 3 genes per cluster (Suppl. Figure 8A+B). In order to evaluate the accuracy of the synteny clusters, functional gene annotations were performed using eggNOG mapper. Subsequently, the predominant KEGG module within each synteny cluster was determined, and the proportion of genes sharing this annotation within the cluster was calculated (see Suppl. Figure 8C). Only synteny clusters with 5 or more genes and at least two annotated genes were included, leaving 35,798 clusters. For 28,341 clusters all genes in the synteny cluster were assigned the same KEGG module, and 80.5% of the modules had more than 80% agreement.

Discussion

MAGinator is a novel pipeline for quantifying the abundances of *de novo* generated MAG clusters. In contrast to reference-based abundance estimations, this allows extensive

integration of abundance and functional properties for individual members of the microbial community. Furthermore, it features generation of signature gene derived phylogenies for MAG clusters and discovery of gene synteny clusters. It is implemented in Snakemake to take advantage of the integrated work distribution capabilities necessary for processing large scale metagenomics data. It features logging for ease of monitoring progress and visualisation for diagnostic purposes. We have demonstrated the functionality and utility of MAGinator via several avenues, both simulated and real datasets.

The performance of MAGinator was evaluated in comparison to existing profiling tools. We benchmarked MAGinator using the simulated strain-madness dataset produced by CAMI II. We found that MAGinator is capable of profiling samples at a comparable level to the already established tools. Notably, while many tools performed well at the genus-level, a decline in performance was observed when focusing on the species-level classification. This drop in performance is expected from reference-based methods, as they are limited to identify only what already exists in their database and are thus unable to annotate novel species. MAGinator demonstrated a notable advantage in this regard, exhibiting the highest average completeness and purity when classifying samples at the species-level. This indicates that MAGinator has the ability to achieve a more accurate and precise characterization of microbial species present in the samples. It should be noted that the high completeness by MAGinator implies a greater sensitivity in detecting and including less abundant or rare taxa in the analysis. However, it may also introduce a certain level of noise or misclassification, which influences the estimation of beta diversity.

When examining the performance of MAGinator on a real dataset the beta diversity was comparable to the analysis carried out by Franzosa et al. Reanalysing their data demonstrates how MAGinator can be used for a metagenomic association study. With the higher resolution of MAGinator when quantifying MAG clusters investigators have the possibility of discovering differentially abundant taxa in much richer detail without compromising other parts of a traditional analysis such as PCoA. Depending on the intention of the study, and the taxonomic composition of the studied microbiomes, the high resolution can also be utilised to gain deeper insights into the subspecies taxonomies. This is for instance relevant when analysing the *Bifidobacterium longum* subspecies.

B. infantis is highly relevant to investigate, as it is known for its greater capacity to metabolise HMOs compared with its closely related subspecies, such as *B. longum*. As their genomes are very similar, distinguishing them by database-dependent approaches is challenging. With StrainPhlAn we are able to identify 2 mutually exclusive clusters, each representing a subspecies, however we see that the two MAG clusters identified with MAGinator for *B. infantis* and *B. longum* yield higher resolution in the form of individual abundance estimates for each. MAGinator is able to successfully classify samples containing the subspecies in samples with low abundance and even when a MAG is not produced in that sample.

These results were reproduced in the CHILD cohort using the signature genes identified in $COPSAC_{2010}$ for the two subspecies. As samples from the CHILD cohort used in this study had lower sequencing depth, still being able to separate the subspecies is valuable. Importantly, it is worth noticing that the separation would most likely have been stronger if the signature genes had been found *de novo* for the specific cohort. This is supported by the read mappings to the signature genes showing a subset of the signature genes defined in $COPSAC_{2010}$ missing in the CHILD cohort, which presumably resulted in underestimation of the abundance for a subset of the samples. This phenomenon highlights the importance of *de novo* dataset-specific discovery of signature genes to yield the best possible abundance estimates of closely related taxonomic entities. A similar phenomenon would be expected when using database-derived strain marker genes.

From the $COPSAC_{2010}$ cohort we demonstrated MAGinators ability to create SNV-level trees based on the sequences from the signature genes of a MAG cluster, used for more fine grained stratification of the MAGs. Even in samples where no MAG was found, they are placed on the tree if they have enough reads that map to the signature genes. By placing these samples in the tree, information from the closely related MAGs can be utilised and allows detection of subspecies-level entities even for samples with very low abundance. From the clusters of the tree it is possible to associate the samples with the gene content of the related MAGs yielding information about clade-specific genes, leaving us with the ability to pair the metadata of the study with the clades and their functions.

Additionally the $COPSAC_{2010}$ cohort was used to illustrate MAGinators ability to group genes co-localised on the chromosome into synteny clusters, further combining the strengths of using both genes and contigs. As genes found close together are often part of the same

genetic pathway or share the same function, this is a valuable insight for associating organisms with the outcomes of a study. This has been validated by functionally annotating the genes of the predicted synteny clusters, confirming that the genes found in synteny are often annotated to be part of the same metabolic pathway.

Conclusion

In conclusion, we have described the development of MAGinator - a pipeline for quantifying MAG clusters and demonstrated the benefits of this approach to commonly generated data types in the metagenomics field. Through reanalysis of publicly available data we have illustrated how new insights can be gained from MAGinator at a higher taxonomic resolution than available from commonly used tools. We believe that this higher resolution is key to unlocking the potential of metagenomics to identify critical strains for human health and environmental investigations. MAG cluster resolution metagenomics allows for accurate integration of abundance, taxonomic and functional annotation in microbiome studies, which is needed to empower investigations in the microbiome field.

Data availability

CAMI II strain-madness benchmarking dataset is available at https://frl.publisso.de/data/frl:6425521/strain/short_read/. The gold standard and benchmark profiles are found at https://github.com/CAMI-challenge/second_challenge_evaluation/tree/master/profiling.

The dataset from Franzosa et al. used for benchmarking is available as supplementary from their paper and the raw data is available at ENA accession SAMN08049618.

The raw COPSAC fastq files are available at NCBI BioProject PRJNA715601.

CHILD shotgun metagenomics sequencing data is available at NCBI BioProject PRJNA838575.

ACKNOWLEDGEMENTS

We express our deepest gratitude to the children and families of the COPSAC cohort studies for all their support and commitment. We acknowledge and appreciate the unique efforts of the COPSAC research team. All funding received by COPSAC is listed on www.copsac.com. The Lundbeck Foundation (Grant no R16-A1694); The Ministry of Health (Grant no

903516); Danish Council for Strategic Research (Grant no 0603-00280B) and The Capital Region Research Foundation have provided core support to the COPSAC research center. JS has received funding from the Danish Council for Independent Research (Grant no. 8045-00081B).

We thank the CHILD Cohort Study (CHILD) participant families for their dedication and commitment to advancing health research. CHILD was initially funded by CIHR and AllerGen NCE,and the metagenomic data reported here were generated with support from Genome Canada and Genome BC (274CHI).

References

1. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01688-w.
2. Liu, B., Gibbons, T., Ghodsi, M. & Pop, M. MetaPhyler: Taxonomic profiling for metagenomic sequences. in *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 95–100 (IEEE, 2010). doi:10.1109/BIBM.2010.5706544.
3. Milanese, A. *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**, 1014 (2019).
4. Liu, Y. *et al.* CSMD: a computational subtraction-based microbiome discovery pipeline for species-level characterization of clinical metagenomic samples. *Bioinformatics* btz790 (2019) doi:10.1093/bioinformatics/btz790.
5. Meyer, F. *et al.* Critical Assessment of Metagenome Interpretation: the second round of challenges. *Nat. Methods* **19**, 429–440 (2022).
6. Underwood, M. A., German, J. B., Lebrilla, C. B. & Mills, D. A. Bifidobacterium longum subspecies infantis: champion colonizer of the infant gut. *Pediatr. Res.* **77**, 229–235 (2015).
7. Dai, D. L. Y. *et al.* Breastfeeding enrichment of B. longum subsp. infantis mitigates the effect of antibiotics on the microbiota and childhood asthma risk. *Med* **4**, 92-112.e5 (2023).
8. Asakuma, S. *et al.* Physiology of Consumption of Human Milk Oligosaccharides by Infant Gut-associated Bifidobacteria. *J. Biol. Chem.* **286**, 34583–34592 (2011).

9. Ojima, M. N. *et al.* Priority effects shape the structure of infant-type Bifidobacterium communities on human milk oligosaccharides. *ISME J.* **16**, 2265–2279 (2022).

10. Nissen, J. N. *et al. Binning microbial genomes using deep learning*. http://biorxiv.org/lookup/doi/10.1101/490078 (2018) doi:10.1101/490078.

11. Mamba, https://github.com/mamba-org/mamba, QuantStack & mamba contributors, 2020

12. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Research* **10**, 33 (2021).

13. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).

14. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

15. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

16. Vasimuddin, Md., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 314–324 (IEEE, 2019). doi:10.1109/IPDPS.2019.00041.

17. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

18. Zachariasen, T. *et al.* Identification of representative species-specific genes for abundance measurements. *Bioinforma. Adv.* **3**, vbad060 (2023).

19. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

20. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5**, e9490 (2010).

21. 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).

22. Van Dongen, S. Graph Clustering Via a Discrete Uncoupling Process. *SIAM J. Matrix Anal. Appl.* **30**, 121–141 (2008).

23. Joshi NA, Fass JN. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ

files. (2011).

24. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

25. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

26. Bremges, A., Fritz, A. & McHardy, A. C. CAMITAX: Taxon labels for microbial genomes. *GigaScience* **9**, giz154 (2020).

27. Piro, V. C., Lindner, M. S. & Renard, B. Y. DUDes: a top-down taxonomic profiler for metagenomics. *Bioinformatics* **32**, 2272–2280 (2016).

28. Shi, L. & Chen, B. LSHvec: a vector representation of DNA sequences using locality sensitive hashing and FastText word embeddings. in *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics* 1–10 (ACM, 2021). doi:10.1145/3459930.3469521.

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

30. Nguyen, N., Mirarab, S., Liu, B., Pop, M. & Warnow, T. TIPP: taxonomic identification and phylogenetic profiling. *Bioinformatics* **30**, 3548–3555 (2014).

31. Franzosa, E. A. *et al.* Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* **4**, 293–305 (2018).

32. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **8**, e61217 (2013).

33. Bisgaard, H. *et al.* Deep phenotyping of the unselected COPSAC $_{2010}$ birth cohort study. *Clin. Exp. Allergy* **43**, 1384–1394 (2013).

34. Stokholm, J. *et al.* Maturation of the gut microbiome and risk of asthma in childhood. *Nat. Commun.* **9**, 141 (2018).

35. Li, X. *et al.* The infant gut resistome associates with E. coli, environmental exposures, gut microbiome maturity, and asthma-associated bacterial composition. *Cell Host Microbe* **29**, 975-987.e4 (2021).

36. Moraes, T. J. *et al.* The Canadian Healthy Infant Longitudinal Development Birth Cohort Study: Biological Samples and Biobanking: The CHILD study: biological samples. *Paediatr. Perinat. Epidemiol.* **29**, 84–92 (2015).

37. Xu, S. *et al. Ggtree* : A serialized data object for visualization of a phylogenetic tree and annotation data. *iMeta* **1**, (2022).

38. 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).

39. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).

40. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

41. Csardi, G. & Nepusz, T. The igraph software package for complex network research.

42. Meyer, F. *et al.* Assessing taxonomic metagenome profilers with OPAL. *Genome Biol.* **20**, 51 (2019).

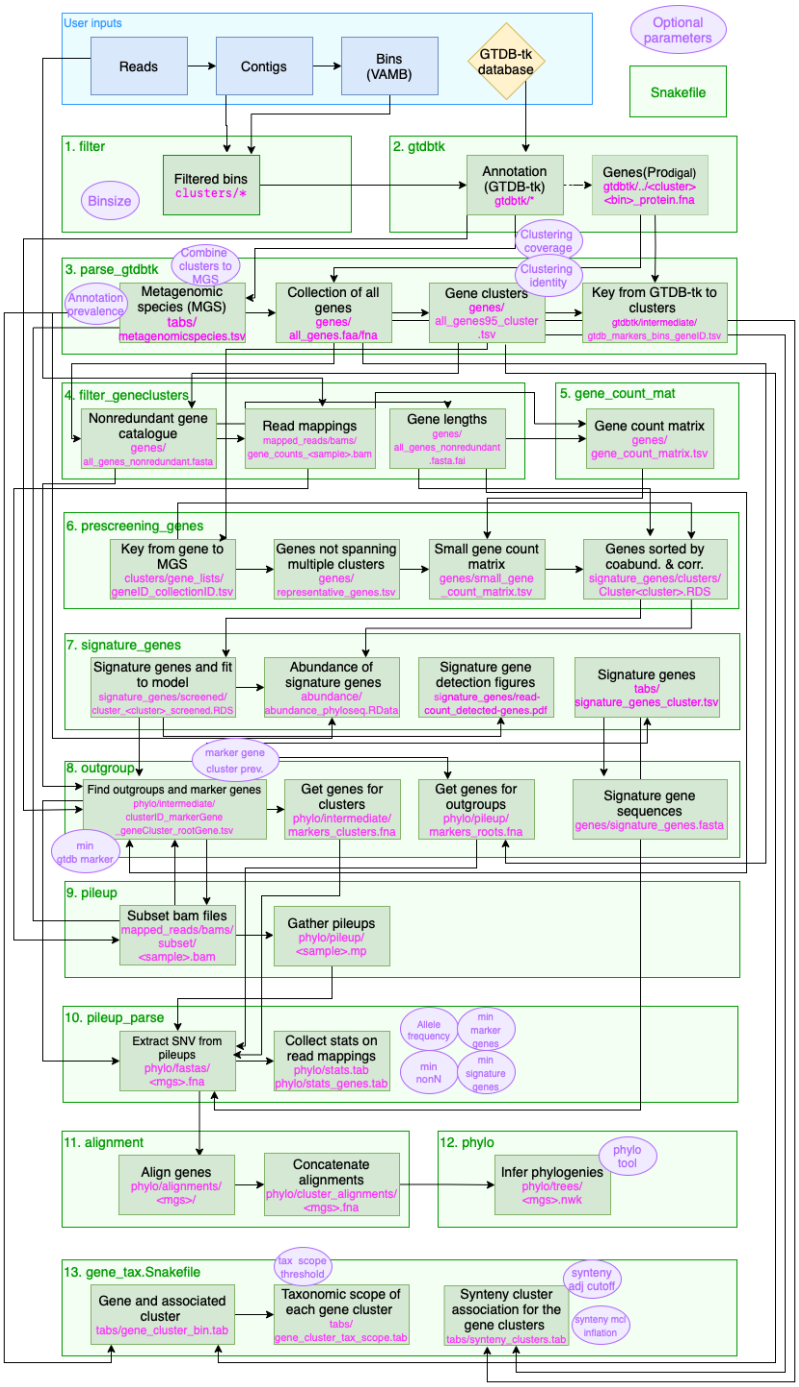
43. LoCascio, R. G., Desai, P., Sela, D. A., Weimer, B. & Mills, D. A. Broad Conservation of Milk Utilization Genes in *Bifidobacterium longum* subsp. *infantis* as Revealed by Comparative Genomic Hybridization. *Appl. Environ. Microbiol.* **76**, 7373–7381 (2010).

44. Alessandri, G., Ossiprandi, M. C., MacSharry, J., Van Sinderen, D. & Ventura, M. Bifidobacterial Dialogue With Its Human Host and Consequent Modulation of the Immune System. *Front. Immunol.* **10**, 2348 (2019).

45. Beghini, F. *et al.* Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife* **10**, e65088 (2021).

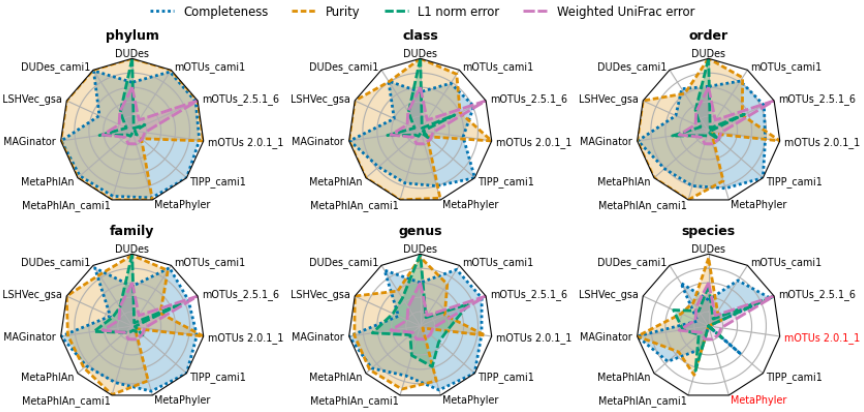
Supplementary Data: MAGinator enables strain-level quantification of *de novo* MAGs

Table of contents


User inputs
Reads
Contigs
Bins (VAMB)
GTDB-tk database
Optional parameters
Snakefile
1. filter
Binsize
Filtered bins
clusters/*
2. gtdbtk
Annotation (GTDB-tk)
gtdbtk/*
Genes(Prodigal)
gtdbtk/../<cluster><bin>_protein.fna
Clustering coverage
Clustering identity
3. parse_gtdbtk
Combine clusters to MGS
Annotation prevalence
Metagenomic species (MGS)
tabs/
metagenomicspecies.tsv
Collection of all genes
genes/
all_genes.faa/fna
Gene clusters
genes/
all_genes95_cluster.tsv
Key from GTDB-tk to clusters
gtdbtk/intermediate/
gtdb_markers_bins_geneID.tsv
4. filter_geneclusters
Nonredundant gene catalogue
genes/
all_genes_nonredundant.fasta
Read mappings
mapped_reads/bams/
gene_counts_<sample>.bam
Gene lengths
genes/
all_genes_nonredundant.fasta.fai
5. gene_count_mat
Gene count matrix
genes/
gene_count_matrix.tsv
6. prescreening_genes
Key from gene to MGS
clusters/gene_lists/
geneID_collectionID.tsv
Genes not spanning multiple clusters
genes/
representative_genes.tsv
Small gene count matrix
genes/small_gene_count_matrix.tsv
Genes sorted by coabund. & corr.
signature_genes/clusters/
Cluster<cluster>.RDS
7. signature_genes
Signature genes and fit to model
signature_genes/screened/
cluster_<cluster>_screened.RDS
Abundance of signature genes
abundance/
abundance_phyloseq.RData
Signature gene detection figures
signature_genes/read-count_detected-genes.pdf
Signature genes
tabs/
signature_genes_cluster.tsv
marker gene cluster prev.
8. outgroup
Find outgroups and marker genes
phylo/intermediate/
clusterID_markerGene_geneCluster_rootGene.tsv
Get genes for clusters
phylo/intermediate/
markers_clusters.fna
Get genes for outgroups
phylo/pileup/
markers_roots.fna
Signature gene sequences
genes/signature_genes.fasta
min gtdb marker
9. pileup
Subset bam files
mapped_reads/bams/
subset/
<sample>.bam
Gather pileups
phylo/pileup/
<sample>.mp
10. pileup_parse
Extract SNV from pileups
phylo/fastas/
<mgs>.fna
Collect stats on read mappings
phylo/stats.tab
phylo/stats_genes.tab
Allele frequency
min marker genes
min nonN
min signature genes
11. alignment
Align genes
phylo/alignments/
<mgs>/
Concatenate alignments
phylo/cluster_alignments/
<mgs>.fna
12. phylo
Infer phylogenies
phylo/trees/
<mgs>.nwk
phylo tool
13. gene_tax.Snakefile
tax_scope threshold
Gene and associated cluster
tabs/gene_cluster_bin.tab
Taxonomic scope of each gene cluster
tabs/
gene_cluster_tax_scope.tab
Synteny cluster association for the gene clusters
tabs/synteny_clusters.tab
synteny adj cutoff
synteny mcl inflation

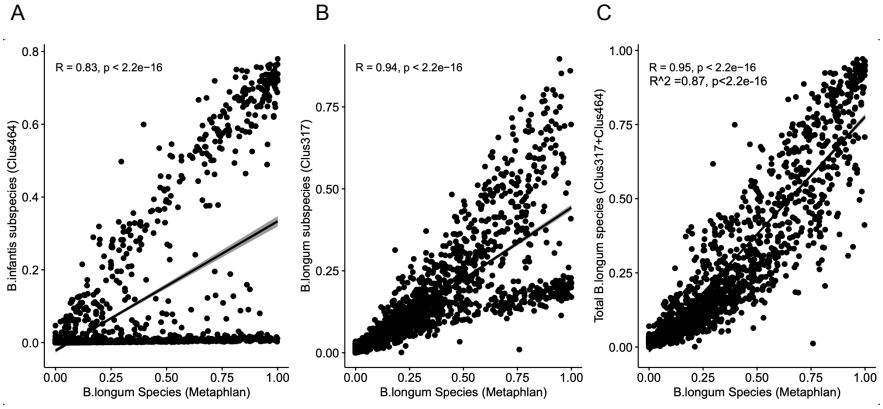
Supplementary Figure 1: *MAGinator workflow.*

A light green box indicates a snakefile and the darker green box indicates a deliverable (directory or file). The purple circles indicate user configurable parameters. The arrow indicates data dependencies, where the flow of information from one file is used to create the file it points towards.



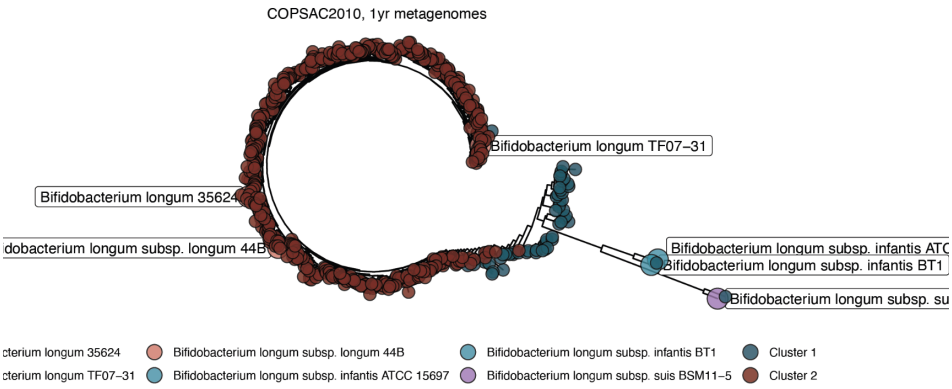
Supplementary Figure 2*: Benchmark using OPAL*

Comparing taxonomic profiling results for the CAMI strain-madness data set. Metrics for the relative abundance of the profiles are calculated across samples for Completeness, Purity, L1 norm error and Weighted UniFrac error and are shown in a Spider-plot for the taxonomic ranks between phylum and species-level. The tools indicated with red means no data was available for that rank.



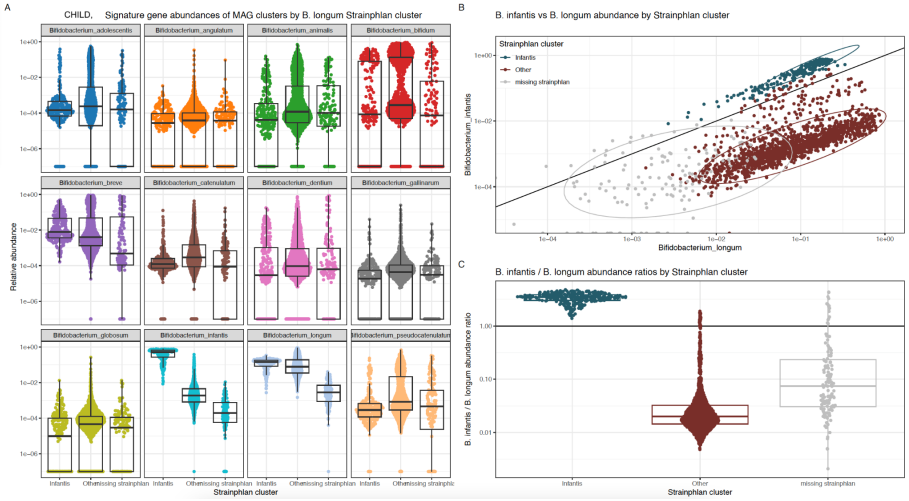
Supplementary Figure 3: *Strain-resolved tracking of B. infantis metagenomes*

The relationship between the relative abundance for the *B. longum* species in the CHILD dataset found with MetaPhlAn and MAGinator's representative clusters for (A) *B. longum* (B) *B. infantis* (C) Both abundances added together. Each dot indicates a sample.



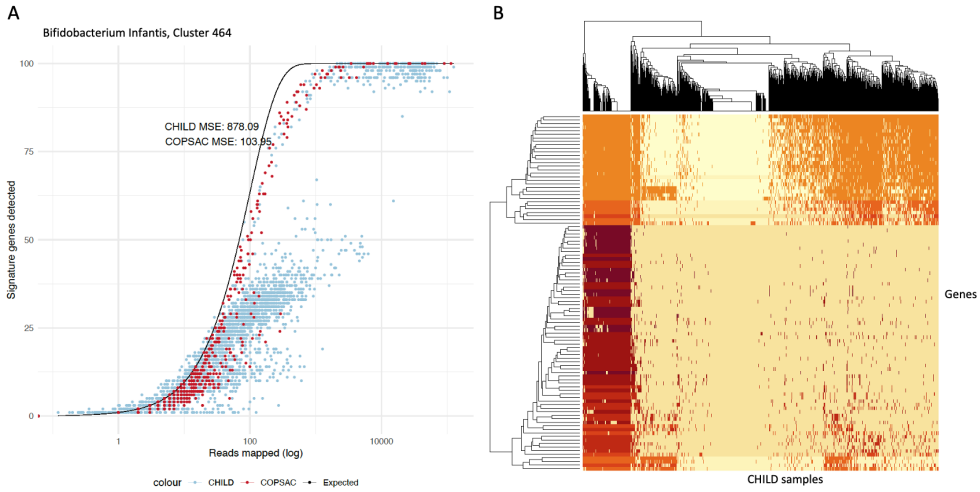
Supplementary Figure 4*: Strain-level analysis of B. longum* in $COPSAC_{2010}$

StrainPhlAn phylogenetic tree of samples based on SNVs of B. longum markers, resulting in 2 clades. Dots represent samples with sufficient marker coverage as well as the 6 references. Cluster 1 indicates *B. infantis* and Cluster 2 indicates *B. longum.*



Supplementary Figure 5: *Stratification of StrainPhlAn clusters using relative abundance of MAGinator clusters - CHILD cohort*

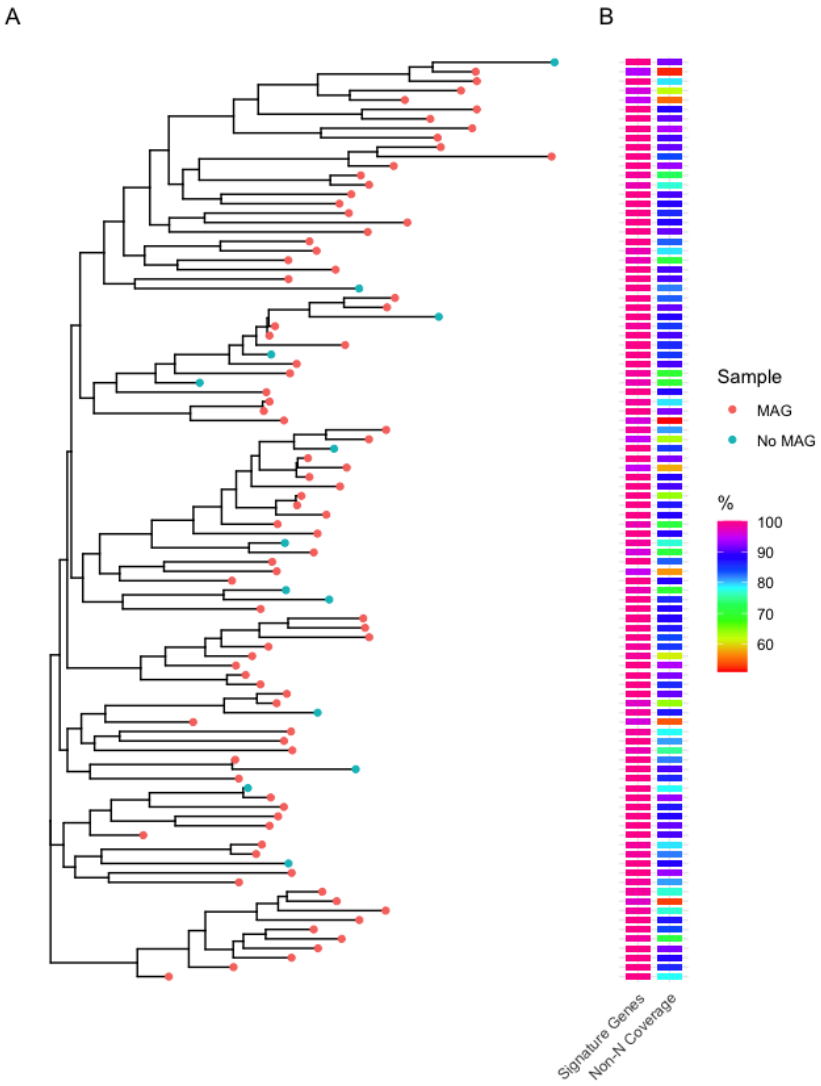
Cluster 1 indicates B. infantis and Cluster 2 indicates B. longum.
(A) Relative abundance of StrainPhlAn clusters stratified by all Bifidobacterium clusters identified by MAGinator (B) Relative abundance of B. infantis and B. longum identified with MAGinator coloured by StrainPhlAn cluster. (C) The ratio of B. infantis to B. longum is displayed for the StrainPhlAn clusters.



Supplementary Figure 6: *Read mappings of B. infantis signature genes*

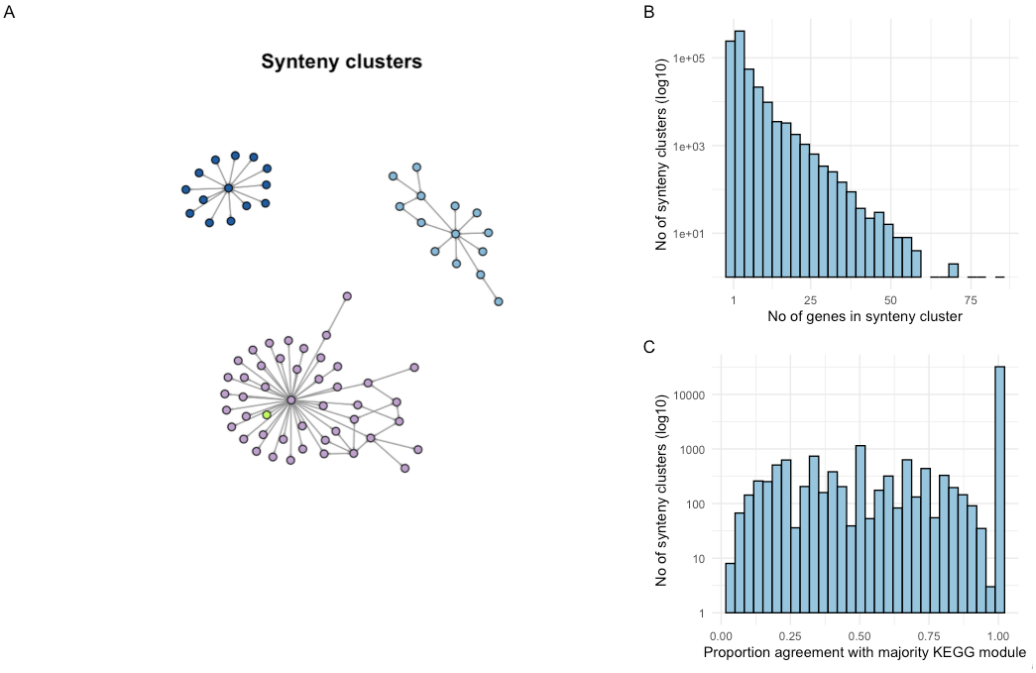
(A) The number of reads mapped to the signature genes (defined in $COPSAC_{2010}$) of the *B. infantis* cluster is presented with the number of signature genes detected. Each dot is a sample. The red colour indicates $COPSAC_{2010}$ samples, the blue colour indicates CHILD samples. The black line indicates the expected distribution[1]. (B) Heatmap of the read mappings of the *B. infantis* signature genes for the CHILD samples.

[1] Zachariasen, T. *et al.* Identification of representative species-specific genes for abundance measurements. *Bioinforma. Adv.* **3**, vbad060 (2023).



***Supplementary Figure 7**: SNV-level phylogenetic tree of a MAG cluster based on signature genes*

A) Phylogenetic tree constructed from readmappings to the signature genes of a MAG cluster annotated as *Faecalibacterium* sp900758465 from $COPSAC_{2010}$. Tip colour indicates if the sample has a MAG. B) Heatmap showing how many of the 100 signature genes that are detected within the sample and the fraction of bases that are Non-N in the alignment of the reads to the signature gene sequence in each sample (%).



Supplementary Figure 8: *Synteny clusters and functional annotation of* $COPSAC_{2010}$

A) The graph network of 3 synteny clusters are shown. The colours represent KEGG modules (green indicates no KEGG module annotation). B) The distribution of synteny cluster size. C) The proportion of genes in the synteny cluster in agreement with the most common KEGG module in the cluster. Only synteny clusters with 5 or more genes are included.

***Supplementary Table 1**: Output generated by MAGinator*

Directory	Content
abundance	abundance_phyloseq.RData - Phyloseq object for R, with abundance and taxonomic data
clusters	.fa - Fasta files with nucleotide sequence of bins
genes	all_genes.faa - Amino acid sequences of all ORFs all_genes.fna - Nucleotide sequences of all ORFs all_genes_nonredundant.fasta - Nucleotide sequences of gene cluster representatives all_genes_cluster.tsv - Gene clusters matrix/gene_count_matrix.tsv - Read count for each gene cluster for each sample synteny/ - Intermediate files for synteny clustering of gene clusters
gtdbtk	GTDB-tk taxonomic annotation for each MAG cluster
logs	Log files
mapped_reads	bams/ - Bam files for mapping reads to gene clusters
phylo	alignments/ - Alignments for each signature gene cluster_alignments/ - Concatenated alignments for each MAG cluster pileup/ - SNV information for each MAG cluster and each sample trees/ - Phylogenetic trees for each MAG cluster stats.tab - Mapping information such as non-N fraction, number of signature genes and marker genes, read depth, and number of bases not reaching allele frequency cutoff stats_genes.tab - Same as above but the information is split per gene
signature_genes	R data files with signature gene optimization read-count_detected-genes.pdf - Figure for each MAG cluster displaying number of identified SG's in each sample along with the number of reads mapped.
tabs	gene_cluster_bins.tab - Table listing which bins each gene cluster was found in gene_cluster_tax_scope.tab - Table listing the taxonomic scope of each gene cluster metagenomicspecies.tab - Table listing which, if any, clusters where merged in MAG cluster and the taxonomy of those signature_genes_cluster.tsv - Table with the signature genes for each MAG cluster synteny_clusters.tab - Table listing the synteny cluster association for the gene clusters. Gene clusters from the same synteny cluster are genomically adjacent. tax_matrix.tsv - Table with taxonomy information for MAG cluster

Supplementary Table 2: *OPAL benchmark. Average completeness (%) across taxonomic ranks.*

The mean of the tools are indicated with bold.

	superkingdom	phylum	class	order	family	genus	species	strain
Mean across tools	**100**	**90.1**	**75.1**	**79.0**	**83.3**	**82.7**	**45.6**	**0.7**
DUDes	100	67	53.6	61.3	55.1	60.2	32.4	0
DUDes cami1	100	98.8	81.6	68.8	95.7	91.8	63.4	0
LSHVec gsa	100	50	40	50	33.3	33.3	15	0
MAGinator	100	99.5	99	99.2	99.6	99.8	89.6	7.4
MetaPhlAn	100	96	76.8	80.7	84.3	89.2	67.1	0
MetaPhlAn_cami1	100	94.8	75.8	79.8	81.8	71.6	29.4	0
MetaPhyler	100	97	80	83.3	94.2	92.2	0	0
mOTUs 2.0.1_1	100	94.2	60	79.5	90.8	85.7	0	0
mOTUs 2.5.1_6	100	97	81.6	84.7	85.9	93.6	84	0
mOTUs cami1	100	97	78	81.7	96	92.9	67.4	0
TIPP cami1	100	100	100	100	100	99.5	53.8	0

Supplementary Table 3*: OPAL benchmark. Average purity (%) across taxonomic ranks.*

The mean of the tools are indicated with bold.

	superkingdom	phylum	class	order	family	genus	species	strain
Mean across tools	**95.2**	**92.4**	**85.9**	**79.9**	**80.0**	**73.6**	**36.5**	**8.8**
DUDes	100	100	100	100	98.1	95.3	84.9	0
DUDes cami1	100	100	80.6	61.1	84.3	57.8	34.9	0
LSHVec gsa	100	100	100	100	100	100	37.5	0
MAGinator	100	100	100	100	90.3	92.4	90.1	96.7
MetaPhlAn	100	100	100	100	88.5	92.8	49.5	0
MetaPhlAn cami1	97	100	100	100	99.3	91.1	63.8	0
MetaPhyler	100	100	97.4	72.7	78.9	79.2	0	0
mOTUs 2.0.1_1	100	100	100	100	100	90.9	0	0
mOTUs 2.5.1_6	100	100	60.6	51.5	43	37.3	17.8	0
mOTUs cami1	100	100	93.2	87.4	91.7	69.1	21	0
TIPP cami1	50	16.9	12.6	6.1	5.43	4.21	2.43	0

Paper III

Differential responses of the gut microbiome and resistome to antibiotic exposures in infants and adults

Li, X., Brejnrod, A., Thorsen, J., **Zachariasen, T.**, Russel, J., Trivedi, U., Vestergaard, G. A., Stokholm, J., Rasmussen, M. A., Sørensen, S. J.

Submitted and in second review at Nature Communications, 2023

The results from the work carried out in relation to this thesis involved the preprocessing, assembly and MAG construction of the sample of the two cohorts. The work also included a phylogenetic analysis, which is presented on page 8-9 of the manuscript as well as in Suppl. figure S3+S4.

Differential responses of the gut microbiome and resistome to antibiotic exposures in infants and adults

Xuanji Li[1], Asker Brejnrod[2], Jonathan Thorsen[3], Trine Zachariasen[2], Jakob Russel[1], Urvish Trivedi[1], Gisle Alberg Vestergaard[2], Jakob Stokholm[3,4], Morten Arendt Rasmussen[3,4*], Søren Johannes Sørensen[1*]

Affiliation: [1]Department of Biology, Section of Microbiology, University of Copenhagen, 2100 Copenhagen, Denmark; [2]Technical University of Denmark, Section of Bioinformatics, Department of Health Technology, 2800 Kgs. Lyngby, Denmark; [3]COPSAC, Copenhagen Prospective Studies on Asthma in Childhood, Herlev and Gentofte Hospital, University of Copenhagen, Copenhagen, Denmark; [4]Department of Food Science, Section of Microbiology and Fermentation, University of Copenhagen, 1958 Frederiksberg C, Denmark.

***Corresponding author**: Søren J. Sørensen and Morten A. Rasmussen, Universitetsparken 15, bldg. 1, DK2100 Copenhagen, telephone: +45 51 82 70 07, Fax: +45 35 32 20 40, e-mail: sjs@bio.ku.dk and mortenr@food.ku.dk

Summary

Despite their crucial importance for human health, little is known about how the gut resistome changes with age or in response to antibiotic treatment across ages. Here, we used fecal metagenomic data from Danish infants and young adults to fill this gap. The gut resistomes were characterized by a bimodal distribution driven by *E. coli* composition. The typical profile of the gut resistome differed significantly between adults and infants, with the latter distinguished by higher gene and plasmid abundances. However, the predominant antibiotic resistance genes (ARGs) were the same. Antibiotic treatment reduced bacterial diversity and increased ARG and plasmid abundances in both cohorts, especially core ARGs. The effects of antibiotic treatments on the gut microbiome lasted longer in adults than in infants, and different antibiotics were associated with distinct impacts. Overall, this study broadens our current understanding of gut resistome dynamics and the impact of antibiotic treatment across age groups.

Keywords

Gut microbiome·antibiotics·different ages·duration·antibiotics resistance genes·E. coli·distribution diversity·infants·adults

Introduction

The rampant use of antibiotics has escalated the spread of antibiotic resistance among bacteria to the point where multi-drug resistant infections have become untreatable, posing a major challenge for modern medicine[1,2]. The indigenous bacteria residing in the human gut[3] constitute a large reservoir of antibiotic resistance genes (ARGs) which they exchange among themselves and with pathogens through horizontal gene transfer[4,5]. A comprehensive understanding of antibiotic resistance profiles and the ARG-carrying bacteria in the human gut is essential for developing novel intervention strategies to minimize the spread of antibiotic resistance. Metagenomic sequencing has provided initial characterizations of ARGs in the human gut microbiome[6–9], yet the links between antibiotic use, age, bacterial hosts, and ARGs remain poorly explored, particularly in large human cohorts.

Antibiotic resistance emerges in the infant gut through early colonization by bacteria, mainly acquired from the mother[10,11] and environmental exposures[12–15]. Previous work by our group described how the infant gut serves as a reservoir of ARGs, with *E. coli* being the largest single contributor[16]. Through the first years of life, the gut microbiome gradually comes to resemble that of adults, after which it is believed to be relatively stable[17]. Many studies have described the compositional differences in the gut microbiome between infants and adults[18–21], but to date, little is known about the differences in their ARG profiles. However, this information is necessary to understand the spread and succession of ARGs and to improve antibiotic stewardship in infants and adults.

More generally, the problem of antibiotic resistance can only be addressed through an improved understanding of the effects of antibiotics on the body, and how these might differ at different ages and life stages. It is well known that antibiotic treatments can have negative effects on the gut microbiome[22–25]. Given the differences in community composition, stability, and resilience between infant and adult gut assemblages, it is possible that the manner and extent to which the microbiome responds and recovers from antibiotic treatment may vary with age. For example, an animal study showed that the recovery of gut microbes from antibiotic treatment was affected by host diet, bacterial community structure, and host living environment[26]. However, with respect to differences between the infant and adult gut microbiome in humans, the response variance to conventional antibiotic therapy has not been fully explored, although such information is critical for understanding how antibiotics remodel the gut.

In this study, we sequenced fecal metagenomes from a Danish cohort of 217 young adults, aged 18 years, and used metagenomic bins to associate ARGs with their bacterial hosts, thus gaining insight into the distribution of ARGs across bacterial species. Moreover, we comprehensively compared the abundance and community composition of ARGs (in bacteria as well as plasmids) and ARG-carrying hosts between these adults and a cohort of 662 one-year-old Danish infants, and explored the underlying drivers for the differences in resistance gene profiles. Finally, we investigated and compared the influence of conventional antibiotic treatment on the infant and adult gut microbiomes, as assessed by changes in microbial composition, antibiotic resistance, and mobile genetic elements, including plasmids.

Results

The distribution of ARG profiles in the adult gut is bimodal and reflects the role of *E. coli* as an ARG reservoir

First, we characterized ARGs in the gut microbiome of 217 young adults, aged 18 years, who were members of the COPSAC2000 cohort. A total of 293 ARGs were detected, which conferred resistance to 33 drug classes. In this assemblage, genes associated with resistance to tetracycline and fluoroquinolone were the most abundant (Fig. 1A), followed by those targeting penam, cephalosporin, macrolide, and rifamycin. The main mechanism of resistance encoded by ARGs was antibiotic efflux pumps (Fig. S1). Almost half of all ARGs (42.7%) encoded resistance to at least two different drug classes, and are referred to hereafter as multiple-drug resistance genes (MDR ARGs) (Fig. S1). The most common type of MDR ARG conferred resistance to fluoroquinolone and tetracycline. The majority of ARGs (53% in abundance) in the adult gut originated from Proteobacteria (Fig. S1), specifically from *E. coli* (≈ 40%). The next-largest contribution came from Bacteroidetes, with 31%. Within Proteobacteria, ARG richness was high in several taxa, such as *Escherichia* species, *Pseudomonas aeruginosa*, *Citrobacter braakii*, *Klebsiella pneumoniae*, and *Enterobacter hormaechei* (Fig. 1H). The detailed distribution of ARGs in different bacteria species is shown in Table S1. Different bacterial phyla exhibited distinct patterns both in terms of the number and type of ARGs present (Fig. S1 and Table S2). For example, Proteobacteria contained the highest number of unique ARGs (163), and these were mainly β-lactam resistance genes.

Based on their abundance patterns, ARGs were divided into four non-overlapping groups (Fig. 1B). Notably, the distribution of ARG richness among samples was bimodal, with one peak with low richness and another peak with high richness (Fig. 1C). Likewise, clustering based on ARG abundance revealed two distinct groups of samples (Fig. 1B): cluster 1 *high ARG richness* (n = 87) and cluster 2 *low ARG richness* (n = 130), which was supported by a 'partitioning around medoids' (PAM) clustering analysis (Fig. 1D-F). Compared to cluster 2, ARGs in cluster 1 were not only more abundant but also more diverse (Fig. 1E).

To investigate the factors underlying the bimodal ARG distribution, we compared the bacterial composition of the two clusters. We first determined that there were no differences in sequencing coverage between the samples in the two clusters (Wilcoxon test; P = 0.21, Fig. S2), ruling out the influence of sequencing depth. We then characterized bacterial composition using MetaPhlAn[27]. A significant correlation was detected between the composition of bacterial communities and that of ARGs through Procrustes analysis (permutational test; r = 0.77, P < 0.001, Fig. 2B). Furthermore, the two clusters differed significantly in their bacterial composition (PERMANOVA; P < 0.001). To identify which bacteria were critical to this difference, we analyzed differentially abundant bacteria between the two clusters and ranked them according to their importance in shaping the clustering pattern. Among the 542 bacterial species detected, 16 species were differentially abundant between the two clusters (Fig. 1G), and the most important of these was clearly *E. coli*. Indeed, the mean relative abundance of *E. coli* in cluster 1 was 66 times higher than that in cluster 2 (mean relative abundance; cluster 1 vs. cluster 2, 4.55% vs. 0.069%). In addition, random forest analysis demonstrated that *E. coli* content was a determining factor in grouping ARGs, and that it was far more important than any other taxon (Fig. S1).

To investigate this further, we assessed the bacterial origin of ARGs using metagenome-assembled genomes (MAGs). In total, we detected *E. coli* MAGs with ARGs in 112 samples, 86 of which were from cluster 1 and 26 from cluster 2. When we removed these *E. coli*–associated ARGs from all samples, we observed an eight-fold reduction in the proportion of variance explained by the two ARG clusters, from 16.4% to 2.1% (Fig. 1F). Without *E. coli*, ARG abundance and diversity in cluster 1 were significantly lower, to the point that values in the two clusters became comparable (Fig. 1E). This provided clear evidence of the abundance of ARGs in *E. coli* and the effect this has on the overall gut microbiome. Although the mean relative abundance of *E. coli* was only around 1.86% in the adult gut, the mean relative abundance of ARGs in this bacteria accounted for about 32% of the total, with the majority in cluster 1 (Fig. 1H).

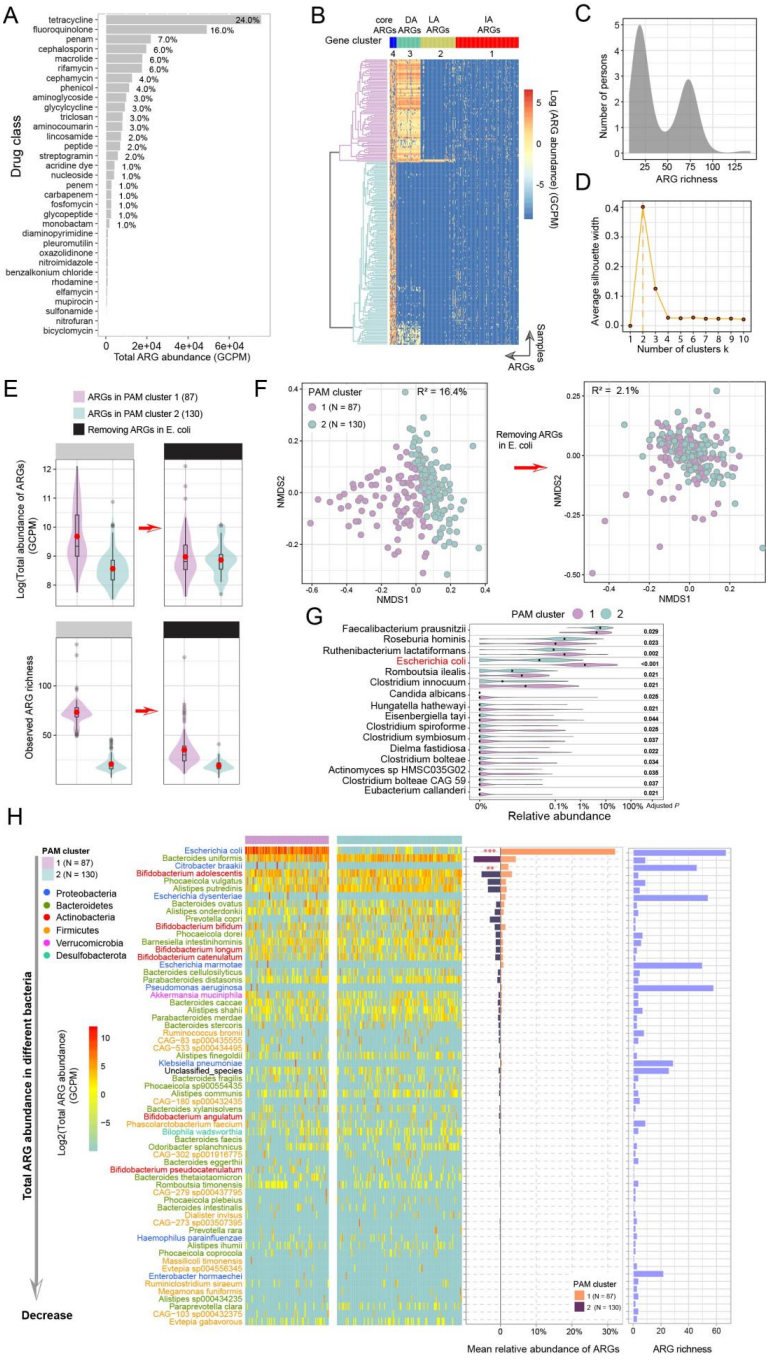

A
Drug class
tetracycline 24.0%
fluoroquinolone 16.0%
penam 7.0%
cephalosporin 6.0%
macrolide 6.0%
rifamycin 6.0%
cephamycin 4.0%
phenicol 4.0%
aminoglycoside 3.0%
glycylcycline 3.0%
triclosan 3.0%
aminocoumarin 3.0%
lincosamide 2.0%
peptide 2.0%
streptogramin 2.0%
acridine dye 1.0%
nucleoside 1.0%
penem 1.0%
carbapenem 1.0%
fosfomycin 1.0%
glycopeptide 1.0%
monobactam 1.0%
diaminopyrimidine
pleuromutilin
oxazolidinone
nitroimidazole
benzalkonium chloride
rhodamine
elfamycin
mupirocin
sulfonamide
nitrofuran
bicyclomycin
0 2e+04 4e+04 6e+04
Total ARG abundance (GCPM)
B
core ARGs
DA ARGs
LA ARGs
IA ARGs
Gene cluster
4 3 2 1
Log (ARG abundance) (GCPM)
Samples
ARGs
C
Number of persons
ARG richness
D
Average silhouette width
Number of clusters k
E
ARGs in PAM cluster 1 (87)
ARGs in PAM cluster 2 (130)
Removing ARGs in E. coli
Log(Total abundance of ARGs) (GCPM)
Observed ARG richness
F
PAM cluster
1 (N = 87)
2 (N = 130)
R² = 16.4%
NMDS1
NMDS2
Removing ARGs in E. coli
R² = 2.1%
G
PAM cluster 1 2
Faecalibacterium prausnitzii 0.029
Roseburia hominis 0.023
Ruthenibacterium lactatiformans 0.002
Escherichia coli <0.001
Romboutsia ilealis 0.021
Clostridium innocuum 0.021
Candida albicans 0.025
Hungatella hathewayi 0.021
Eisenbergiella tayi 0.044
Clostridium spiroforme 0.029
Clostridium symbiosum 0.037
Dielma fastidiosa 0.022
Clostridium bolteae 0.034
Actinomyces sp HMSC035G02 0.035
Clostridium bolteae CAG 59 0.037
Eubacterium callanderi 0.021
0% 0.1% 1% 10% 100% Adjusted P
Relative abundance
H
PAM cluster
1 (N = 87)
2 (N = 130)
Proteobacteria
Bacteroidetes
Actinobacteria
Firmicutes
Verrucomicrobia
Desulfobacterota
Log2(Total ARG abundance) (GCPM)
Total ARG abundance in different bacteria
Decrease
Escherichia coli
Bacteroides uniformis
Citrobacter braakii
Bifidobacterium adolescentis
Phocaeicola vulgatus
Alistipes putredinis
Escherichia dysenteriae
Bacteroides ovatus
Alistipes onderdonkii
Prevotella copri
Bifidobacterium bifidum
Phocaeicola dorei
Barnesiella intestinihominis
Bifidobacterium longum
Bifidobacterium catenulatum
Escherichia marmotae
Bacteroides cellulosilyticus
Parabacteroides distasonis
Pseudomonas aeruginosa
Akkermansia muciniphila
Bacteroides caccae
Alistipes shahii
Parabacteroides merdae
Bacteroides stercoris
Ruminococcus bromii
CAG-83 sp000435555
CAG-533 sp000434495
Alistipes finegoldii
Klebsiella pneumoniae
Unclassified_species
Bacteroides fragilis
Phocaeicola sp900554435
Alistipes communis
CAG-180 sp000432435
Bacteroides xylanisolvens
Bifidobacterium angulatum
Phascolarctobacterium faecium
Bilophila wadsworthia
Bacteroides faecis
Odoribacter splanchnicus
CAG-302 sp001918775
Bacteroides eggerthii
Bifidobacterium pseudocatenulatum
Bacteroides thetaiotaomicron
Romboutsia timonensis
CAG-279 sp000437795
Phocaeicola plebeius
Bacteroides intestinalis
Dialister invisus
CAG-273 sp000307395
Prevotella rara
Haemophilus parainfluenzae
Alistipes ihumii
Phocaeicola coprocola
Massiliicoli timonensis
Eubacterium sp004556345
Enterobacter hormaechei
Ruminiclostridium siraeum
Megamonas funiformis
Alistipes sp000434235
Paraprevotella clara
CAG-103 sp000432375
Evtepia gabavorous
***
**
PAM cluster
1 (N = 87)
2 (N = 130)
0 10% 20% 30%
Mean relative abundance of ARGs
0 20 40 60
ARG richness

Fig. 1 ARG characteristics of different bacteria in the adult gut and bimodal distribution of ARGs in the adult gut, driven by *E. coli*. A) The total abundance of ARGs resistant to 33 drug classes. **B)** Heatmap depicting the abundance of 293 ARGs across the samples. Samples were clustered with complete linkage hierarchical clustering based on Euclidean distance. ARGs were clustered into four categories by PAM clustering based on Euclidean distance; Cluster 4 (core ARGs, N = 15) contained ARGs that were highly abundant and prevalent across all samples. Cluster 3 (differentially abundant (DA) ARGs, N = 55) contained ARGs that differed significantly in their abundance among samples. Cluster 2 (low-abundance (LA) ARGs, N = 80) contained ARGs present at a low abundance in samples. Cluster 1 (intermediate-abundance (IA) ARGs, N = 143) contained ARGs whose abundance in the samples fell between those in cluster 4 and those in cluster 2. **C)** Density plot of ARG richness in the cohort. **D)** Average silhouette width of PAM clustering for k = 1 to 10 clusters. The higher the silhouette width, the stronger the clustering effect. **E)** Log-transformed total ARG abundance and observed ARG richness (α-diversity) before (left) and after (right) the removal of *E. coli* ARGs from the two ARG PAM clusters. **F)** NMDS ordination plot of Bray-Curtis dissimilarity matrix of ARG abundances before (left) and after (right) the removal of *E. coli* ARGs from the two ARG PAM clusters. The percent of explained variance (R^2) generated with the PERMANOVA test is shown in the figure. **G)** Relative abundances of 16 species (of 542 in total) that differed in abundance between the two clusters. Relative abundance on the x-axis is shown on a logarithmic scale; black dots indicate median value; *P*-values were generated by the Wilcoxon rank-sum test, and FDR adjustments are represented as adjusted *P*-values. **H)** Total ARG abundance in the bacterial species in each sample in two clusters (left), mean relative abundance of ARGs in bacterial species in two clusters (middle), and ARG richness in different bacterial species (right). For ease of viewing, only the 63 species with the highest ARG abundance are listed. ***P*-value < 0.01 and ****P*-value < 0.001, obtained from the Wilcoxon test with FDR adjustment.

ARGs are more abundant in the infant gut than in adults, with *E. coli* as the largest single contributor

The distribution of ARGs in the gut has never been systematically compared between adults and infants. Therefore, we performed a comprehensive comparison of the ARGs described above and those identified, using the same workflow, in a cohort of 662 one-year-old Danish infants.

Overall, ARG profiles were significantly different between adults and infants (β-diversity (Bray-Curtis), PERMANOVA; R^2 = 8.5%, P < 0.001, Fig. 2A). Procrustes analysis revealed a significant correlation between the composition of bacterial communities and that of ARGs in both the adult and infant gut (permutational test; r_adults = 0.77, r_infants=0.78, both P < 0.001, Fig. 2B), suggesting that ARG distribution was strongly tied to overall bacterial composition regardless of host age.

β-diversity analysis also highlighted a significant difference in gut microbial composition between adults and infants (β-diversity (Bray-Curtis), PERMANOVA; R^2 = 10%, P < 0.001, Fig. 2C). Furthermore, of the 896 bacterial species detected, 482 (54%) were differentially abundant between the two cohorts (Wilcoxon test; FDR adjusted P < 0.05), indicating that the differences between adults and infants were influenced by the overall bacterial composition. However, considering that *E. coli* contains a large proportion of ARGs in both adults and infants[16] and that the relative abundance of *E. coli* differed between adults and infants (mean relative abundance, infants vs. adults, 5.4% vs. 1.86%, Fig. S3), we wanted to determine whether these age-related differences remained even in the absence of *E. coli*. We thus removed all *E. coli*–associated ARGs from the two groups and re-evaluated the overall differences in ARG composition (Fig. S3). We found that the percentage of variance in ARG profiles that was explained by the two age groups did not decrease in the absence of *E. coli*, indicating that this species is not the only factor shaping age-related differences (Fig. S3).

ARGs were more abundant in the infant gut than in the adult gut, as reflected in both the number of ARGs per million genes and the relative abundance of ARGs (Wilcoxon test; P < 0.001, Fig. 2D and Fig. 2E). When we removed *E. coli*–associated ARGs from the analysis, the difference between adults and infants in the mean number of ARGs per million genes and the mean relative abundance of ARGs decreased by 53% and 51%, respectively (Fig. 2D and 2E). These results suggest that, although it is not the only factor at work, *E. coli* still plays an important role in the differences in ARG load between the adult and infant gut.

Plasmids are important mobile genetic elements that can transfer ARGs between cells. We therefore specifically investigated mobile ARGs carried on plasmids in the adult and infant gut. As in the overall analysis, the abundances of plasmids and mobile ARGs were higher in the infant gut than in the adult gut (Wilcoxon test; P < 0.001, Fig. 2F). However, the proportion of mobile ARGs on plasmids relative to total ARGs did not differ between cohorts (Wilcoxon test; P = 0.19, Fig. 2F).

To gain more insight into the ARGs carried by *Escherichia* species in the two cohorts, we plotted phylogenetic trees of *Escherichia* MAGs and clustered MAGs based on

their ARG profiles; for the sake of comparison, we also carried out the same procedure for *Bifidobacterium* MAGs. From a phylogenetic perspective, *Escherichia* MAGs differed between the two cohorts (PERMANOVA; $P = 0.02$, Fig. S4). In addition, *Escherichia* MAGs belonging to four main species correlated with ARG profiles (PERMANOVA; $P = 0.01$, Fig. S4). However, we did not find an ARG profile in *Escherichia* that was exclusive to the adult or infant gut. Instead, in *Bifidobacterium* we found one ARG profile almost exclusively in infants that was also predominantly distributed in one specific MAG cluster (Fig. S5). In addition, many *Bifidobacterium* MAGs did not carry ARGs.

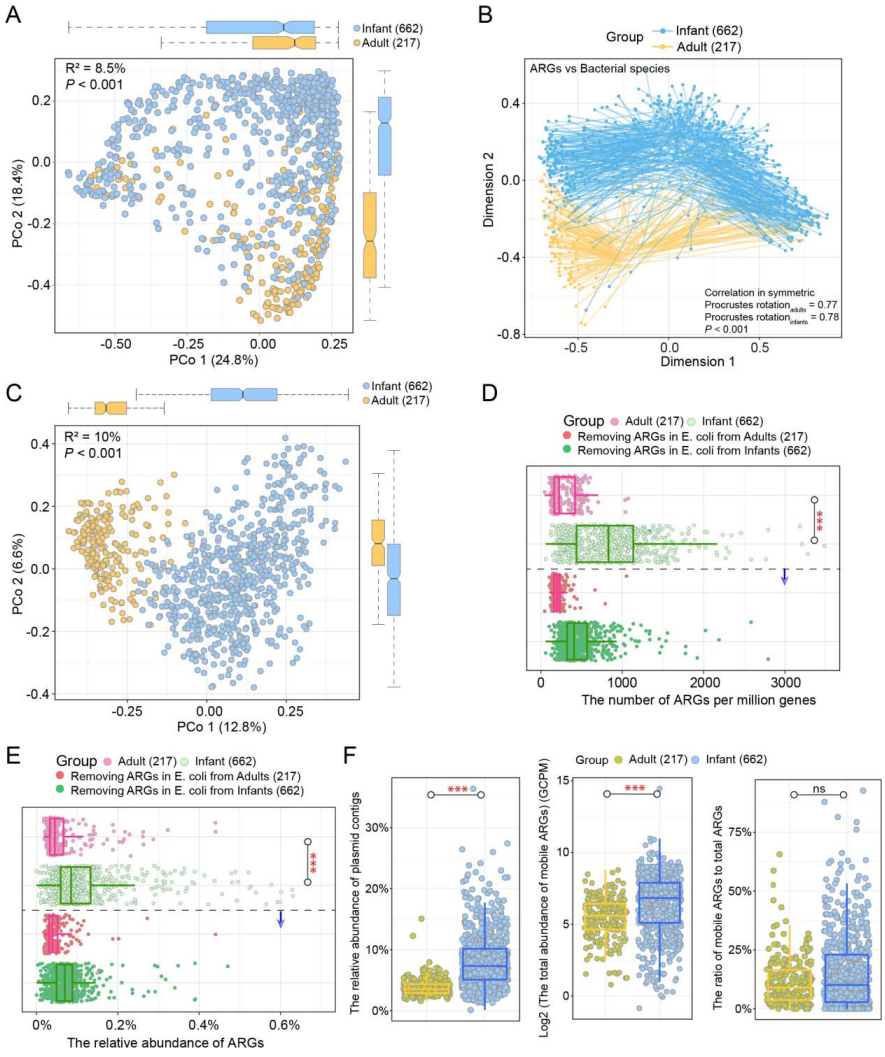


Fig. 2 ARG profiles differed significantly between the infant and adult gut, with infants containing a higher abundance of ARGs. A) PCoA plot based on Bray-Curtis dissimilarity matrices of ARG abundance in the adult and infant gut (values in brackets represent the percentage of variance explained by the principal coordinates). *P*-value and R^2 were generated with a PERMANOVA test. Box plots along each axis show the value of each point at the respective coordinates. **B)** Procrustes analysis of the association between the composition of ARGs and that of bacterial communities as characterized by MetaPhlAn in the gut of adults and infants. **C)** PCoA plot based on Bray-Curtis dissimilarity matrices of bacterial community composition characterized by MetaPhlAn in the adult and infant gut. *P*-value and R^2 were generated with a PERMANOVA test. Box plots along each axis show the value of each point

at the respective coordinates. **D&E)** Boxplot with jitter points showing the number of ARGs per million genes (D) and the relative abundance of ARGs out of all genes (E) before and after removing *E. coli* ARGs in the adult and infant gut. ***P-value < 0.001, obtained from the Wilcoxon test. **F)** Boxplot with jitter points showing the relative abundance of plasmid contigs out of all contigs, the log-transformed total abundance of ARGs in plasmids, and the ratio of mobile ARGs to total ARGs in the adult and infant gut. ARGs carried on plasmids are defined as mobile ARGs. ***P-value < 0.001, ns: P-value > 0.05, obtained from the Wilcoxon test.

Infants and adults share dominant ARGs and bacterial species carrying them in the gut microbiome

Although the overall ARG profiles differed between the infant and adult gut, we wanted to investigate if certain aspects of these assemblages might be shared across age groups. To evaluate this, we explored commonalities between the infant and adult gut in terms of six aspects. First, we compared the alpha diversity (observed richness) of these groups, and found that the number of ARG-carrying bacterial species and the number of mobile ARGs on plasmids were significantly higher in the adult gut than in the infant gut (Wilcoxon test; $P < 0.001$, Fig. S6). When we identified the ARGs and ARG-carrying bacteria that were shared by both infants and adults, we found that they included some of the most abundant representatives in both cohorts. Specifically, infants and adults shared 106 ARG-carrying bacterial species, which contributed 68% and 53% of the total ARGs in each group (relative abundance), respectively, while unique species contributed only about 6% of ARGs (relative abundance) (Fig. 3A). Likewise, 191 ARGs were shared between the two cohorts, accounting for over 98% of the total ARG abundance in each (Fig. 3B). For the other ARG-related aspects investigated, the results were similar. ARGs and drug-resistance classes that were unique to only one cohort tended to be present in lower abundance (Fig. 3C, 3D, 3E and 3F). Details on the comparison of shared and unique features with respect to these six ARG-related groups are listed in Table S3.

Next, we investigated the top ten drug classes to which these ARGs conferred resistance. For most of these drug classes, infants had a significantly higher abundance of associated ARGs than adults did (Wilcoxon test; adjusted $P < 0.05$, Fig. 3G, 3H). In both cohorts, tetracycline and fluoroquinolone ARGs were the most abundant. Tetracycline and aminoglycoside were the drug classes most commonly

targeted by mobile ARGs in the infant gut, while mobile ARGs in the adult gut more often targeted tetracycline and macrolide.

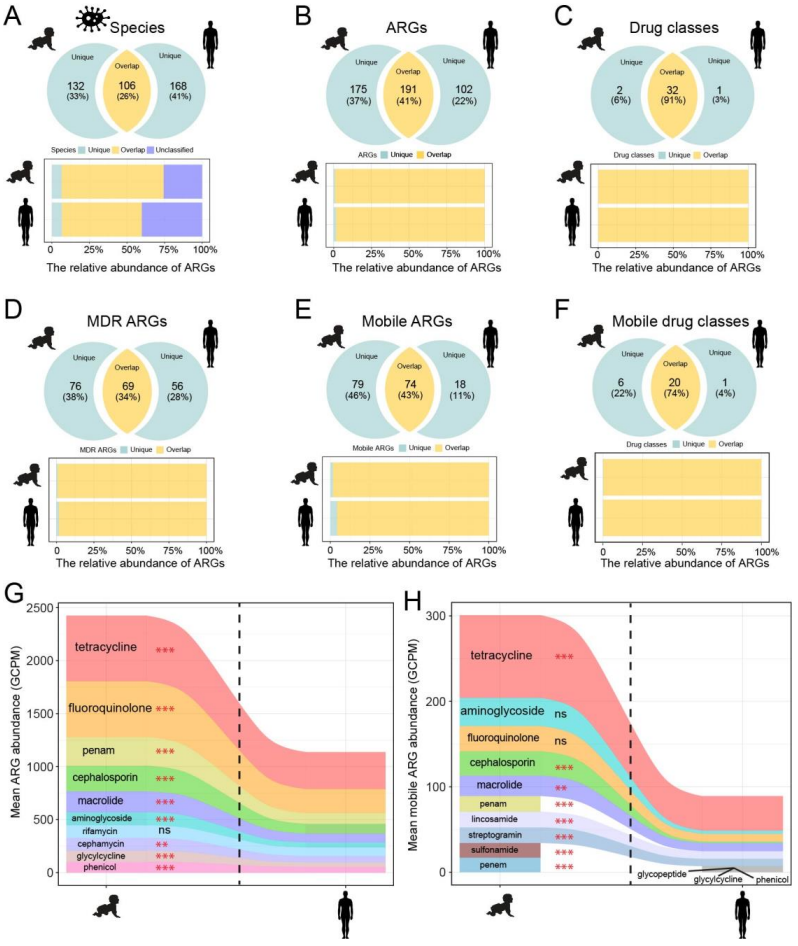


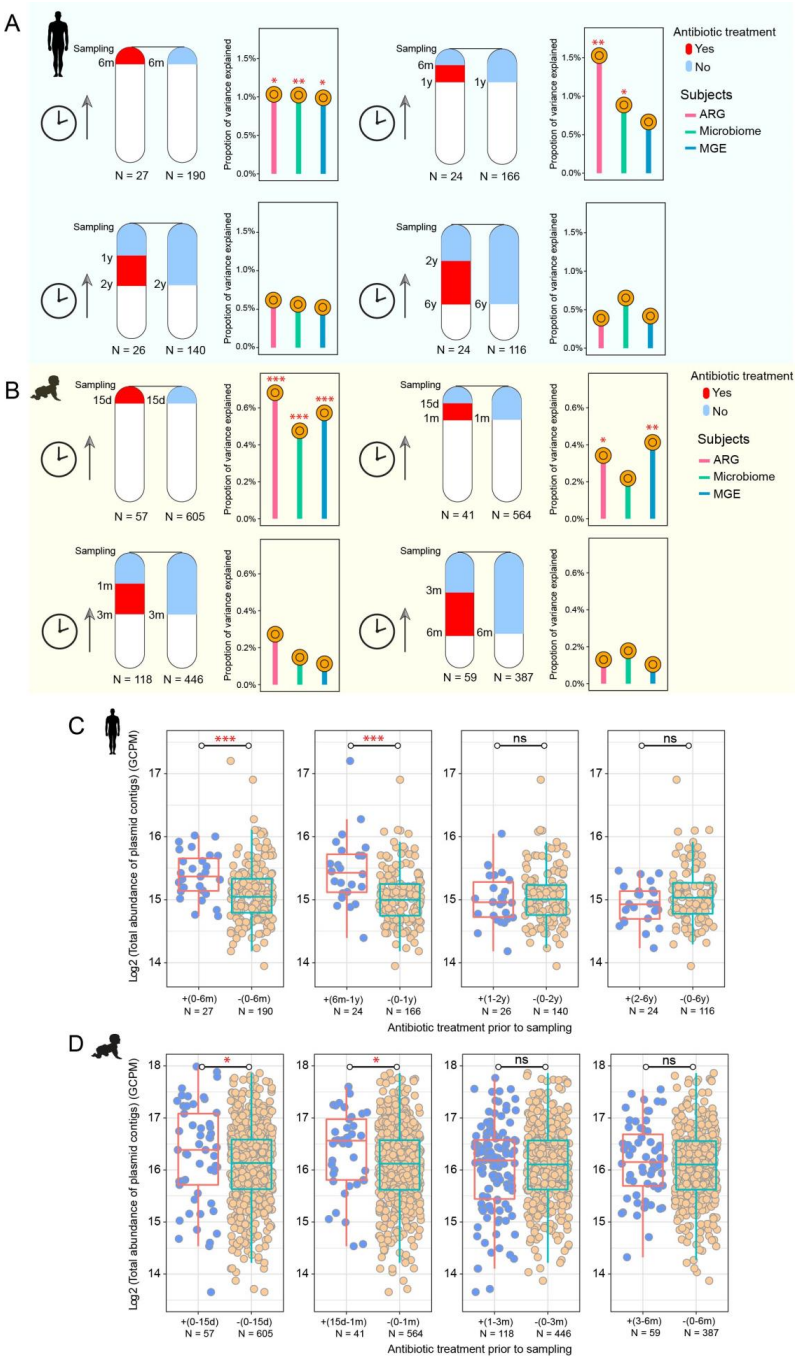
Fig. 3 ARGs shared by the adult and infant gut accounted for the vast majority of ARG abundance in each cohort. Analyses of the unique and shared **(A)** ARG-carrying bacterial species, **(B)** ARGs, **(C)** drug classes targeted by ARGs, **(D)** MDR ARGs, **(E)** mobile ARGs, **(F)** and drug classes targeted by mobile ARGs in both gut, with respect to the number of individual species/genes/drug classes (top panel) and their relative abundance in the total population of ARGs (bottom panel). **G&H)** Mean abundance of the 10 most commonly targeted drug classes by ARGs (G) and by mobile ARGs (H) in the adult and infant gut. ***P*-value < 0.01, ****P*-value < 0.001, and ns: *P*-value > 0.05, from the Wilcoxon test with Bonferroni adjustment. Seven of the 10 mobile drug classes were shared between cohorts.

Compared to infants, antibiotic treatment in adults had a longer-lasting effect on microbial composition, ARG and MGE profiles, and plasmid abundance

It is well known that antibiotic therapy changes the gut microbiome[23,28], but the extent to which this effect may differ according to age has not yet been characterized. Here, we compared the association between antibiotic treatment and alterations in the gut microbiome in adults and infants. In particular, we examined differences in microbial composition, ARGs, and mobile genetic elements (MGEs), which here included the genetic elements related to mobility, such as integrases, transposases, and insertion sequences. In the adult cohort, the effects of antibiotic treatment persisted up to about one year. Instead, for infants, the effects of antibiotic treatment were detectable for about one month. Specifically, ARG profiles and microbial community composition were significantly different in the gut of adults who had taken antibiotics within 6 months or between 6 months and 1 year before sampling compared to those who had not (β-diversity, PERMANOVA; < 6m, $P = 0.02$, 0.002, respectively; 6m – 1y, $P = 0.005$, 0.03, respectively; Fig. 4A). Instead, MGE profiles differed only in the group that had taken antibiotics within 6 months of sampling (< 6m, $P = 0.03$, Fig. 4A). No effects were detectable for any of these three indicators when the antibiotic use had occurred more than 1 year prior to sampling ($P > 0.05$, Fig. 4A). In the infant cohort, ARG and MGE profiles were different in individuals who had received antibiotic treatment within 15 days of sampling or between 15 days and 1 month before sampling compared to those who had not (< 15d, $P < 0.001$; 15d – 1m, $P = 0.03$, 0.009, respectively, Fig. 4B). Infants who had taken antibiotics more recently also demonstrated alterations in microbial community composition (< 15d, $P < 0.001$, Fig. 4B). None of these effects were apparent if the antibiotic use had occurred more than 1 month before sampling ($P > 0.05$, Fig. 4B).

The duration of the effect of antibiotics in adults and infants was also reflected in plasmid abundance. Plasmids can horizontally transfer resistance and virulence genes between bacterial cells. In the adult gut, the effect of antibiotics on plasmids lasted up to about 1 year: the total abundance of plasmids was higher in the gut of adults who had taken antibiotics within 6 months of sampling or between 6 months and 1 year before sampling than those in the corresponding control groups (Wilcoxon test; $P < 0.001$, Fig. 4C). In contrast, there were no differences in plasmid abundance

between adults who had taken antibiotics more than 1 year before sampling and those who had not (Wilcoxon test; $P > 0.05$, Fig. 4C). Similarly, plasmid abundance in the gut of infants who had taken antibiotics more than 1 month before sampling did not differ from those who had not (Wilcoxon test; $P > 0.05$, Fig. 4D). However, plasmids were more abundant in the gut of infants who had received antibiotics between 15 days and 1 month before sampling or within 15 days of sampling than in individuals in the corresponding control groups (Wilcoxon test; $P = 0.03$ (0 – 15d), $P = 0.01$ (15d – 1m), Fig. 4D).


A
Sampling
6m
6m
N = 27
N = 190
Propotion of variance explained
1.5%
1.0%
0.5%
0.0%
*
**
*
Sampling
6m
1y
1y
N = 24
N = 166
**
*
Antibiotic treatment
Yes
No
Subjects
ARG
Microbiome
MGE
Sampling
1y
2y
2y
N = 26
N = 140
Sampling
2y
6y
6y
N = 24
N = 116
B
Sampling
15d
15d
N = 57
N = 605
***
***
***
Sampling
15d
1m
1m
N = 41
N = 564
*
**
Antibiotic treatment
Yes
No
Subjects
ARG
Microbiome
MGE
Sampling
1m
3m
3m
N = 118
N = 446
Sampling
3m
6m
6m
N = 59
N = 387
C
Log2 (Total abundance of plasmid contigs) (GCPM)
17
16
15
14
***
***
ns
ns
+(0-6m)
N = 27
-(0-6m)
N = 190
+(6m-1y)
N = 24
-(0-1y)
N = 166
+(1-2y)
N = 26
-(0-2y)
N = 140
+(2-6y)
N = 24
-(0-6y)
N = 116
Antibiotic treatment prior to sampling
D
Log2 (Total abundance of plasmid contigs) (GCPM)
18
17
16
15
14
*
*
ns
ns
+(0-15d)
N = 57
-(0-15d)
N = 605
+(15d-1m)
N = 41
-(0-1m)
N = 564
+(1-3m)
N = 118
-(0-3m)
N = 446
+(3-6m)
N = 59
-(0-6m)
N = 387
Antibiotic treatment prior to sampling

336

Fig. 4 Antibiotic treatment had longer-lasting effects on the adult gut microbiome than on the infant gut microbiome, as reflected in microbial composition, ARG and MGE profiles, and plasmid abundance. A) Duration of the effect of antibiotic administration on the β-diversity of microbiome, ARG and MGE compositions in the adult gut. Bray-Curtis distance was used as the measure of β-diversity. Adult subjects were divided into four groups depending on when they had taken antibiotics: within 6 months of sampling, 6 to 12 months prior, 1 to 2 years prior, or 2 to 6 years prior to sampling; the corresponding control groups had not received antibiotics in those periods. **P*-value < 0.05 and ***P*-value < 0.01, obtained from the PERMANOVA test. **B)** Duration of the effect of antibiotic administration on the β-diversity (Bray-Curtis distance) of microbiome, ARG and MGE compositions in the infant gut. Infant subjects were divided into four groups depending on when they had taken antibiotics: within 15 days of sampling, 15 to 30 days prior, 1 to 3 months prior, and 3 to 6 months prior; the corresponding control groups had not received antibiotics in those periods. **P*-value < 0.05, ***P*-value < 0.01, and ****P*-value < 0.001, obtained from the PERMANOVA test. **C&D)** Duration of the effect of antibiotic administration on total plasmid abundance in the adult gut (C) and in the infant gut (D). The four studied periods are the same as in panel A or in panel B. "+" represents antibiotics administered in a given period, and "-" represents antibiotics not administered in a given period. ****P*-value < 0.001, ns: *P*-value > 0.05, from the Wilcoxon test.

Antibiotic treatment enhances ARG and MGE abundance and reduces bacterial richness

In addition to the overall alterations, we also observed differences in total ARG and MGE abundance, and bacterial richness as a result of antibiotic treatment. Specifically, ARGs were significantly more abundant in the gut of adults who had taken antibiotics within 1 year of sampling compared with those who had not (Wilcoxon test; $P < 0.001$, Fig. 5A), and the bacterial richness was lower (Wilcoxon test; $P = 0.02$, Fig. 5A). With respect to MGEs, total abundance was higher in adults who had taken antibiotics within 6 months of sampling than in those who had not (Wilcoxon test; $P = 0.036$, Fig. 5A). For infants, the same phenomenon was observed: compared to the corresponding control groups, total ARG abundance was higher in the gut of infants who had taken antibiotics within 1 month of sampling, and gut bacterial diversity was lower in infants who had taken antibiotics within 15 days of sampling (Wilcoxon test; $P < 0.001$, 0.005, respectively, Fig. 5B).

We then explored the effects of antibiotics on the abundance of different types of ARGs: specifically, the four groups of ARGs in the adult gut, clustered using the PAM algorithm (core, DA, IA, and LA; Fig. 1E) and three clusters in the infant gut, obtained

using the same methodology (Fig. S7). We found that antibiotic treatment enhanced the total abundance of low-abundance ARGs in adults and intermediate-abundance ARGs in infants (Wilcoxon test; adjusted $P = 0.044$, $P < 0.001$, respectively, Fig. 5C, 5D). Interestingly, the total abundance of core ARGs—resistance genes that are highly abundant and prevalent overall—also increased in the gut of both adults and infants after antibiotic treatment (Wilcoxon test; adjusted $P < 0.001$, 0.015, respectively, Fig. 5C, 5D). The mean abundance of most individual core ARGs was higher in individuals who had taken antibiotics than in those who had not, although this was not statistically significant (Wilcoxon test; adjusted $P > 0.05$, Fig. S8).

Fifteen core ARGs, mostly associated with tetracycline and MLS resistance (Fig. S8), were detected in the adult gut and were found in between 54% and 100% of samples (mean 76.2%). For several of these ARGs—specifically, *ErmB/H/G*, *tet(40)/O/Q/W*, and *vanI*—more than 20% of these genes were retrieved from plasmids. Two core ARGs (*adeF* and *tetQ*) were detected in 97.7% and 85.8% of the infant gut samples, respectively, and 36% of the latter appeared on plasmids (Fig. S8).

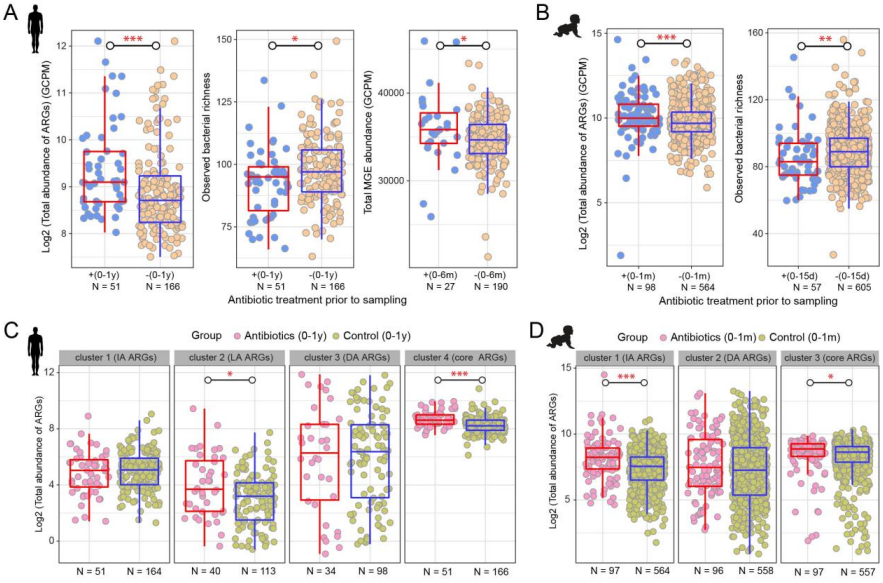


Fig. 5 Antibiotic treatment resulted in an elevated abundance of ARGs and MGEs, and a decrease in observed bacterial richness. A) Changes in ARG abundance and bacterial diversity in the gut of adults who had taken antibiotics within one year of sampling and changes in MGE abundance

in the gut of adults who had taken antibiotics within 6 months of sampling. Individuals who had not taken antibiotics during those periods were used as controls. *P-value < 0.05, ***P-value < 0.001, obtained from the Wilcoxon test. **B)** Changes in ARG abundance in the gut of infants who had taken antibiotics within one month of sampling and changes in bacterial diversity in the gut of infants who had taken antibiotics within 15 days of sampling. Individuals who had not taken antibiotics during those periods were used as controls. **P-value < 0.01, ***P-value < 0.001, from the Wilcoxon test**. C&D)** Changes in the abundance of ARG clusters in the gut of adults (C) who had taken antibiotics within one year of sampling and in the gut of infants (D) who had taken antibiotics within one month of sampling. Individuals who had not taken antibiotics during those periods were used as controls. For adults, the definitions of these four groups and the methodological basis for clustering are described in the legend of Fig. 1B. For infants, ARGs were clustered into three categories by PAM clustering based on Euclidean distance (Fig. S7); Cluster 3 (core ARGs, N = 2) contains highly abundant and prevalent ARGs. Cluster 2 (differentially abundant (DA) ARGs, N = 55) contains ARGs with significant differences in abundance between samples. Cluster 1 (intermediate-abundance (IA) ARGs, N = 311) contains ARGs whose abundance in the samples falls between the ARGs in cluster 3 and those in cluster 2. *P-value < 0.05, ***P-value < 0.001, obtained from the Wilcoxon test with FDR adjustment.

The influence of different antibiotics on the gut microbiome of adults and infants

In the group of adults who had received antibiotic treatment in the year before sampling, we examined whether the type of antibiotic taken had a detectable influence on characteristics of the gut microbiome. With the exception of β-lactam plus sulfonamide, each type of antibiotic was associated with an increase in the mean abundance of ARGs, with tetracycline and β-lactam plus macrolide having a statistically significant effect (Wilcoxon test; adjusted P = 0.036, 0.029, respectively, Fig. 6A). Each antibiotic type was also associated with an increase in mean plasmid abundance, with β-lactam, tetracycline, and β-lactam plus macrolide having statistically significant effects (Wilcoxon test; adjusted P = 0.049, 0.038, 0.0005, respectively, Fig. 6B). Four of the five antibiotic types were also associated with a reduction in mean bacterial richness (exception was β-lactam plus sulfonamide, Fig. S9), and all five antibiotics were associated with increases in mean MGE abundance (Fig. S9). Finally, treatment with tetracycline or macrolide resulted in a significant reduction in the relative abundance of *Bifidobacterium adolescentis* and *Bifidobacterium longum*, two of the 20 most abundant species (Wilcoxon test; adjusted P < 0.05, Fig. 6E).

In the infant cohort, we evaluated whether treatment with one of three major antibiotics—macrolide, penicillin, and ampicillin—in the 15 days before sampling had

distinguishable effects on the infant gut microbiome. All antibiotics were associated with an increase in mean ARG abundance, with macrolide and penicillin having a statistically significant relationship (Wilcoxon test; adjusted P = 0.028, 0.028, respectively, Fig. 6C). Furthermore, all antibiotics were associated with non-significant increases in mean plasmid abundance (Wilcoxon test; adjusted $P > 0.05$, Fig. 6D) and reductions in mean bacterial richness (Fig. S9). Macrolide and penicillin were linked with increases in mean MGE abundance (Fig. S9). None of the antibiotics had a statistically significant influence on the abundance of the 20 most abundant bacterial species. When we investigated the effect on the broader bacterial community, we found that antibiotics were associated with a significant decrease in the relative abundance of *Faecalibacterium prausnitzii* and *Haemophilus parainfluenzae* (Wilcoxon test; adjusted $P < 0.05$, Fig. 6E). Additionally, we observed an increase in the abundance of *E. coli*, although the adjusted P value was not significant.

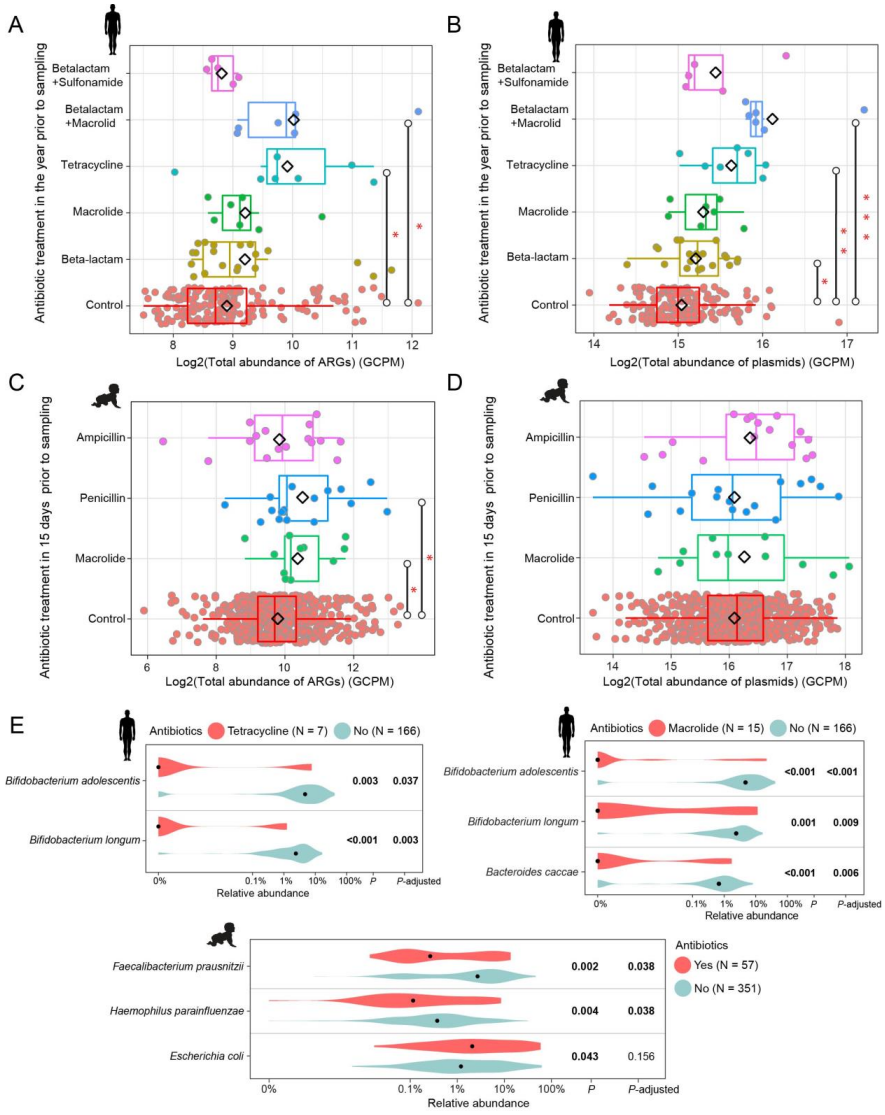


Fig. 6 The effects of different antibiotics on ARG and plasmid abundance, and on the relative abundance of bacterial species. A&B) Changes in ARG abundance (A) and plasmid abundance (B) in the gut of adults who had taken one of five major antibiotics or antibiotic combinations in the year before sampling. Individuals who had not taken antibiotics in that period were used as controls. **P*-value < 0.05, obtained from the Wilcoxon test with FDR adjustment. The black diamond indicates the mean value. **C&D)** Changes in ARG abundance (C) and plasmid abundance (D) in the gut of infants who had taken one of three major antibiotics in the 15 days before sampling. Infants who had not taken

antibiotics within 15 days of sampling were used as controls. *P-value < 0.05, obtained from the Wilcoxon test with FDR adjustment. The black diamond indicates the mean value. **E)** Members of the 20 most abundant bacterial species whose abundance in the gut differed significantly between (top) adults who had taken tetracycline or macrolide in the year before sampling and those who had not received antibiotic treatment, and (bottom) infants who had taken antibiotics in the 15 days before sampling and those who had not within the first year. Relative abundance on the x-axis is shown on a logarithmic scale; black dots indicate median value; P-values were generated by the Wilcoxon rank-sum test and adjusted using FDR.

Discussion

Metagenomic sequencing offers the possibility to gain deeper insight into the distribution and function of ARGs in gut microbes at the species or strain level. Using this approach, we examined the distribution of ARGs in the gut bacteria of 217 young Danish adults, aged 18 years. By combining this information with similar data from 662 one-year-old Danish infants, we were able to describe age-related patterns in the abundance and distribution of ARGs in the gut, as well as associations between antibiotic use and alterations in the gut microbiome, ARGs, and MGEs, including plasmids, across age groups.

In the adult cohort, we obtained evidence that ARGs follow a bimodal distribution that is driven by the abundance of *E. coli*. A similar bimodal distribution had been found for ARGs in the infant gut[16], which suggests that this phenomenon is independent of age. Numerous genomic/molecular studies and *in vitro* resistance assays have shown that members of family Enterobacteriaceae possess an extremely broad array of antibiotic resistance[29–33], particularly to beta-lactam, which has largely been attributed to gene flow under sustained selective pressure resulting from the increase in antibiotic use in recent decades[34,35]. In both the adult and infant gut, the ARG profiles on *Escherichia* MAGs were quite similar, providing additional evidence for the frequent influx of genes into the *Escherichia* genome. Moreover, many studies have shown that this gene transfer is not unidirectional: the rich pool of resistance elements in Enterobacteriaceae genomes also flows to other bacteria[36–38], thereby exacerbating the spread of resistance genes.

Although our study is not longitudinal, it does provide a cross-sectional view of the differences in gut ARGs between early life and adulthood in the Danish population.

We discovered that the dominant ARGs, and the bacterial species on which they were found, were the same in both infants and young adults, which could indicate a prolonged selective advantage or a shared community reservoir. Such a selective advantage, i.e., the persistence of certain genes or gene-carrying bacteria throughout childhood, would likely be due to ongoing selection from external factors such as repeated antibiotic therapy[39–41] and/or a competitive advantage over their bacterial neighbors.

Compared to infants, the proportion, number, and abundance of ARGs was lower in the adult gut, and this was associated with decreased levels of clinically relevant bacteria that contain abundant resistance genes, such as *E. coli*. This mirrors previous findings that infants have a higher load of resistance genes in their gut compared to their mothers[42,43]. Similar results have even been reported from cattle and pigs, in which the abundances of ARGs and resistance-carrying Enterobacteriaceae in the gut are also high early in life and decline with age[44]. Importantly, this early-life peak in Enterobacteriaceae does not seem to be driven by any external factors such as antibiotic use; instead, its trajectory in the gut may be related to favorable environmental conditions and host regulation. Facultative anaerobes such as *E. coli* can consume oxygen and produce an anaerobic environment, thus favoring subsequent colonization by and growth of strictly anaerobic bacteria[45,46]. Previous studies have highlighted various mechanisms by which a host can manage the development of the gut microbiome, such as the immune system response[47], the production of nitrogen-rich mucins, and the creation of a more suitable habitat[48,49]. It is possible that the natural processes of gut maturation may be altered by the presence or abundance of ARGs or ARG-carrying bacteria. Indeed, previous work by our group demonstrated that high ARG abundance was associated with a low degree of maturation of the infant gut microbial community[16]. Obviously, such enrichment poses a threat to infant health by reducing the effectiveness of antibiotic therapy for bacterial infections[50]. Our observation that plasmids were abundant in the infant gut also implies a high frequency of HGT[51,52] which can provide an advantage for the dissemination and persistence of ARGs even in the absence of antibiotics[53,54].

Compared to adults, though, the gut microbiome in infants recovered more quickly from antibiotic therapy. The infant gut microbiome is very dynamic[55] and less diverse

than that of adults, which may indicate that the ecological processes at play are simpler and can more easily recover from perturbations. However, this effect is also mediated by the types and doses of antibiotics used[56–58]. In Denmark, the type and dose of common antibiotics vary according to age[59]. Moreover, the length of the recovery period after antibiotic treatment has also been found to depend on the disease targeted. The present study examined the effects of routine antibiotic treatment of common infections. Instead, in a group of neonates with sepsis who were treated with broad-spectrum antibiotics, the overall gut microbiome took 12 months to return to normal[24]. It is important to note that our analysis examined the mixed effect of all antibiotics taken, where the effects of additional antibiotics may confound the results. Furthermore, although our results indicated that the infant gut microbiome typically returned to baseline levels after about 30 days, we cannot rule out some potential long-term effects that were not addressed in our analysis, such as alterations in specific resistance genes and bacteria[60], immune maturation[61], or metabolic changes[62]. In addition, we cannot rule out confounding by indication—that the antibiotic-treated vs. non-treated infants and adults differed due to factors that contributed to the condition their treatment was prescribed for.

The total abundance of core ARGs was significantly elevated in both the infant and adult gut following antibiotic exposure, implying that they are the primary weapons of bacteria against antibiotics and thus possess the potential for widespread dissemination. This was also supported by the patterns we identified in high ARG prevalence and abundance, as well as plasmid presence. However, different antibiotics had different effects on the abundance of both ARGs and plasmids. Of the five major antibiotics used in adults, tetracycline and beta-lactam plus macrolide had the strongest impact on ARG and plasmid abundance. The effect of the former may be related to the extreme abundance of tetracycline resistance genes in bacteria and plasmids in the adult gut. Although the medical use of tetracycline has declined over the past 20 years and it is no longer used to treat pregnant women and children under 8 years of age[63], it remains one of the most widely used classes of antibiotics worldwide[64]. With respect to the latter, there may be a synergistic effect of taking separate courses of beta-lactam and macrolide within the span of a year which simultaneously calls into action resistance genes against both beta-lactam and macrolide as well as plasmids carrying relevant genes in the gut. In infants, the

administration of penicillin or macrolide in the 15 days prior to sampling was significantly associated with high ARG abundance. In previous work, we found that the influence of macrolide treatment on macrolide resistance genes in the infant gut could last for approximately 2 months, whereas the effect of penicillin was much shorter[16]. A study on Finnish children (2–7 years, median age 5 years) also confirmed that macrolide treatment had a stronger impact on the gut microbiome than penicillin did[57]. In the adult gut, both tetracycline and macrolide were associated with dramatically reduced levels of the beneficial bacteria *Bifidobacterium adolescentis* and *Bifidobacterium longum*, which are the most prevalent species in the adult gut[65–67] and are effective degraders of plant-derived fructooligosaccharides[68]. Similarly, antibiotic administration in infants was found to reduce gut levels of *Haemophilus parainfluenzae*, a conditionally pathogenic bacterium that can cause multiple infections[69–71], but simultaneously reduced levels of *Faecalibacterium prausnitzii*, which is widely considered to be beneficial to host health[72–74]. This reflects the double-edged nature of antibiotic treatment, which kills pathogenic bacteria to cure disease but can also kill sensitive beneficial bacteria. Therefore, the type of antibiotic used, and its potential double-edged effects, should be fully considered in the choice of antibiotic treatment.

STAR★Methods

Key Resources Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Adult feces samples	COPSAC2000 cohort (This study)	http://copsac.com/home/copsac-cohorts/copsac2000cohort/
Infant feces samples	COPSAC2010 cohort (This study)	http://copsac.com/home/copsac-cohorts/copsac2010-cohort/
Software and Algorithms		
GNU Parallel version 20180722	Tange, 2018[75]	https://www.gnu.org/software/parallel/
BBTools v38.19	sourceforge.net/projects/bbmap/	https://jgi.doe.gov/data-and-tools/bbtools/
Sickle v1.33	Joshi and Fass, 2011[76]	https://github.com/najoshi/sickle/releases
Nonpareil v3.30	Rodriguez-R et al., 2018[77]	https://nonpareil.readthedocs.io/en/latest/
MetaPhlAn v2.7.5	Segata et al., 2012[27]	https://pypi.org/project/MetaPhlAn/

SPAdes v3.12.0	Nurk et al., 2017[78]	https://cab.spbu.ru/files/release3.12.0/manual.html
VAMB	Nissen et al., 2021[79]	https://github.com/RasmussenLab/vamb
GTDB-Tk v1.7.0 toolkit	Chaumeil et al., 2018[80]	https://github.com/Ecogenomics/GTDBTk
Prodigal v2.6.3	Hyatt et al., 2010[81]	https://github.com/hyattpd/Prodigal
RGI	Jia et al., 2017[82]	https://card.mcmaster.ca/analyze/rgi
HMMER3 v3.1b2	Mistry et al., 2013[87]	http://hmmer.org/
Bowtie2 v2.3.5	Langmead et al., 2012[91]	https://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Samtools v1.12	Li et al., 2009[93]	http://www.htslib.org/
Platon v5.3	Schwengers et al., 2020[94]	https://github.com/ideasrule/platon/releases
FastANI v1.33103	Jain et al., 2018[98]	https://github.com/ParBLiSS/FastANI/releases
R	core Team, 2018[105]	https://www.r-project.org/
Deposited Data		
Metagenomics data (adults)	This paper	Sequence Read Archive (SRA) under the accession number PRJNA916259
Metagenomics data (infants)	This paper	Sequence Read Archive (SRA) under the accession number PRJNA715601
Other		
NucleoSpin® 96 Soil DNA Isolation Kit optimized for epMotion®	Macherey-Nagel, Düren, DE	https://www.mn-net.com/nucleospin-96-soil-96-well-kit-for-dna-from-soil-740787.2
NovaSeq	Illumina	N/A

Resource Availability

Lead contact

Further information and requests for reagents and resources should be directed to and will be fulfilled by the Lead Contact, Søren J. Sørensen, University of Copenhagen, (sjs@bio.ku.dk).

Experimental Model and Subject Details

Human samples

The $COPSAC_{2000}$ cohort is a mother-child cohort assembled for the primary purpose of studying asthma. It consists of 411 mothers and their children[83]. The 217 fecal samples used for this study were collected as part of the 18-year follow-up visit at the

research clinic or at home following detailed instructions. All samples were stored at -80°C prior to DNA extraction. The 662 fecal samples were obtained from one-year-old infants in the $COPSAC_{2010}$ cohort[84,85].

Ethics

The study was designed with the guiding principles of the Declaration of Helsinki in mind and was approved by the Local Ethics Committee of the Danish Capital Region (COPSAC2000: KF 01-289/96, COPSAC2010: (H-B-2008-093)) and the Danish Data Protection Authority (both cohorts: 2015-41-3696).

Covariates

During scheduled visits to COPSAC clinics, information was collected from participants on the use of antibiotics (including any treatment prior to sampling), the use and duration of other medications, pet ownership, siblings, living area, income, alcohol consumption, smoking, and experiences with disease. This information was verified against registration records.

Method details

Metagenomic sequencing of fecal samples and data processing

Genomic DNA was extracted from fecal samples (~200–250 mg) using the NucleoSpin® 96 Soil DNA Isolation Kit optimized for epMotion® (Macherey-Nagel, Düren, DE) using the epMotion® robotic platform model (Eppendorf) following the manufacturer's protocol. DNA library preparation and data processing was carried out for adult samples following the same protocol used for infant samples[16]. In brief, the DNA library was prepared for Illumina sequencing with the Kapa HyperPrep kit (KAPA Biosystems, Wilmington, MA, USA). Paired-end (150 bp) sequencing of the samples in the DNA library was performed with the Illumina NovaSeq platform by Novogene (China). Bioinformatics analyses were executed in parallel using GUN parallel v20180722[75]. Adapters were removed using BBDuk of BBTools v38.19 (sourceforge.net/projects/bbmap/). Sickle v1.33 was used for the removal of low-quality reads[76]. Human DNA was filtered out using BBMap of BBTools v38.19. In total, 217 gut samples were successfully sequenced, generating between 52.9 and 103 million clean reads per sample (mean ± SD: 58.9 ± 4.5 million reads). The average metagenomic coverage and sequence diversity for each sample was estimated using Nonpareil v3.30 in kmer mode[77]. The mean coverage of adult and infant metagenomic data was 96.42% and 98.23%, respectively (Fig. S10), which represented 'almost

complete coverage' (≥95% of mean coverage). The species-level composition of microbial communities was described using MetaPhlAn v2.7.5[27]. Sequence assembly was performed with SPAdes v3.12.0 using default metagenomic settings[86]. Bins were created using Variational Autoencoders for Metagenomics Binning (VAMB)[79], a method that uses deep learning to bin microbial genomes. All metagenome-assembled genomes (MAGs) at least 200 kbp in length were submitted for taxonomic assignment with the GTDB-Tk v1.7.0 toolkit, based on the GTDB database (release 202)[80]. Among them, the taxonomy of 84.4% big MAGs in 1250 clusters was assigned, which can cover 70% contigs in MAGs. Genes were predicted with Prodigal v2.6.3 in META mode[81]. The reads assigned to *E. coli* by MetaPhlAn were subdivided into two main MAGs, one for *E. coli* and the other for *E. flexneri*. In the presentation of this analysis in Fig. 1 and Fig. 2, however, we classified ARGs from both MAGs as coming from *E. coli*.

ARG and MGE prediction and gene abundance calculation

Resistance gene identifiers (RGI) were used to annotate ARGs based on the Comprehensive Antibiotic Resistance Database (CARD v3.0.7)[82]. ARGs with the strict and perfect thresholds of the RGIs were kept for further analysis. MGE homologs were characterized by HMM search in HMMER3 v3.1b2[87] in combination with the PFAM[88] and TnpPred[89] databases, with "cut_ga" as a threshold criterion[90,92]. If multiple MGE alignments were detected for one gene, only the one with the lowest E value was kept.

Reference genes were indexed using bowtie2-build of Bowtie2 v2.3.5 before aligning reads[91]. Clean reads were aligned against the predicted genes with Bowtie2 aligner. The number of mapped reads in bam files was calculated with Samtools idxstats of Samtools v1.12[93]. Values of gene coverage per million (GCPM)[16], which normalize sequencing depth and gene length, were used to quantify gene abundance. The sum of the GCPM values for all predicted genes in each sample was one million, making it comparable across samples. The formula for calculating GCPM for each gene is $\frac{(\text{counts}/\text{gene length})\times 10^6}{\sum_1^n \text{counts}/\text{gene length}}$, where counts is the number of mapped reads, gene length is the length of the gene, and n is the total number of the predicted gene in each sample.

Plasmid prediction and calculation of contig abundance

Plasmid contigs were identified and characterized with Platon v5.3 using the default settings[94]. Reference contigs were indexed using bowtie2-build before aligning reads.

Clean reads were aligned against the contigs with Bowtie2 aligner. The number of mapped reads in bam files was calculated with Samtools idxstats. GCPM values were used to quantify contig abundance as described above. The sum of the GCPM values for all contigs in each sample was one million, and the formula for calculating GCPM for each contig is $\frac{(\text{counts /contig length}) \times 10^6}{\sum_1^n \text{counts /contig length}}$, where counts is the number of mapped reads, contig length is the length of the contig, and n is the total number of the contigs in each sample.

Relative importance of bacterial species as evaluated by Random Forest

The relative importance of bacterial species in shaping ARG clusters was evaluated by Random Forest analysis[95] using the R-package 'randomForest' v4.7.1.1[96]. The number of trees (ntree) and the number of variables per split (mtry) in the random forest model were set to 500 and 50, respectively, resulting in a stable classifier and a low error rate of 5.99%. The mean decrease in Gini value associated with a predictor was used to estimate the importance of a bacterial species; a higher value indicates a higher importance for that variable.

Comparing ARG and bacterial distributions using Procrustes analysis

Procrustes analysis was used to evaluate the association between the distribution of microbial species and the distribution of ARGs in each sample[97]. A Hellinger transformation was first performed on the ARG matrix and the species abundance matrix, respectively. Bray-Curtis dissimilarity values were calculated between all samples in the two matrices using the R function 'vegdist' in the 'vegan' package, v2.6.2. PCoA ('phyloseq' package v1.38.0) was used to ordinate each dissimilarity matrix. The two ordinated dissimilarity matrics were rotated with the R function 'procrustes' in the 'vegan' package. The R function 'protest' in the 'vegan' package was used to calculate the symmetric Procrustes correlation coefficient r, the sum of squared distance, and a *P*-value with 9999 permutations. The association between the distribution of microbial species and ARGs was visualized with ggplot2.

Construction of phylogenetic tree of metagenome-assembled genomes (MAGs)

The nucleotide-level similarity between MAGs assigned to *Escherichia* or *Bifidobacterium* was assessed with average nucleotide identity (ANI) values using FastANI v1.33[98]. We then used the neighbor-joining method to construct phylogenetic trees[99]. Based on the presence or absence of ARGs in the contigs, the PAM clustering method was used to group *Escherichia* and *Bifidobacterium* MAGs into four categories

each, represented by different colored branches. MAGs assigned to *Escherichia* and *Bifidobacterium* belonged to a total of seven and eight metagenomic species, respectively. The dissimilarity between MAGs was quantified using the cophenetic distance. Permutational multivariate analysis of variance (PERMANOVA) was used to investigate differences in cophenetic distances between MAG clusters based on ARG profiles or between MAGs (R-package 'vegan' v2.6.2)[100]. With respect to genus *Escherichia*, MAGs from the four main species—*E. coli*, *E. coli_D*, *E. flexneri*, and *E. dysenteriae*—were included in the statistical analysis.

α-diversity and β-diversity

All data processing and statistical analyses were carried out using the open-source statistical program R. The observed richness of ARGs and bacterial species was used to assess within-individual diversity (α-diversity), while the Bray-Curtis index served as a measure of between-individual diversity (β-diversity). The ordination of β-diversity matrices was performed with NMDS or PCoA (R-package 'phyloseq' v1.38.0)[101]. The Wilcoxon rank-sum test was used to test for differences in α-diversity among groups (R package 'stats' v4.1.2). PERMANOVA was used to investigate differences in β-diversity. Adjustments were made for multiple comparisons using the Benjamini-Hochberg correction.

Partitioning Clustering for samples or ARGs based on ARG composition

Cluster analyses of samples or ARGs based on ARG composition were performed with Partitioning Around Medoids (PAM) clustering[102] using the R function 'pam' in package 'cluster' v2.1.3[103]. The average silhouette width, which serves as an estimate of the average distance between clusters, was used to assess the quality of PAM clustering; a larger value means better clustering. Euclidean distance was applied to the PAM clustering analysis. The R function 'fviz_nbclust' in package 'factoextra' v1.0.7[104] was used to determine and visualize the optimal number of PAM clusters.

Differential abundance analysis

Wilcoxon rank-sum tests were used to identify the bacterial taxa that were differentially abundant between two groups, with multiple tests corrected by FDR. Likewise, ARG, MGE, and plasmid abundances were compared between two groups using the Wilcoxon rank-sum test with FDR correction.

Linear regression analysis

A linear model (R function 'lm') was fitted to investigate the extent to which the abundance of *E. coli* explained the variance in the number of ARGs per million genes

and the relative ARG abundance. The normality assumption of residuals was checked using the QQ plot.

All statistical analyses were conducted in R version 4.1.2[105]**.**

Data and code accessibility

The COPSAC2010 metagenomics datasets are available in the Sequence Read Archive (SRA) under the accession number PRJNA715601. The COPSAC2000 metagenomics data have been deposited in the SRA under the accession number PRJNA916259 and will be publicly accessible with the publication of the paper. According to Danish and European law, data involving the personal privacy of project participants cannot be publicly available without a cooperation agreement and data transfer agreement. All other data that support the results of this study are available from the corresponding author upon request. The R code used for the data analyses is available from the authors upon request.

Acknowledgments and funding

We appreciate the commitment and assistance provided by the children and families who participated in the COPSAC cohort study. We also recognize and value the special contributions made by each member of the COPSAC research team. This research has been funded by Novo Nordisk Foundation Grant no. NNF19OC0057934598, Novo Nordisk Foundation Grant no. NNF17OC0025014 and Research Council of Norway project no. 300489. Metagenomics analysis was performed by Computerome.

Author contributions

X.L., S.J.S., and M.A.R. conceived the project. M.A.R., J.S., and J.T. collected the samples and information about various environmental exposures. X.L., A.B., T.Z, J.R., and G.A.V. performed metagenomics and statistical data analysis. X.L. wrote the paper. M.A.R., J.T., J.S., A.B., J.R., and U.T. helped interpret the data. All authors read, revised, and approved the final manuscript.

Declaration of Interests

The authors declare no competing interests

References

1. Laxminarayan, R., Duse, A., Wattal, C., Zaidi, A.K.M., Wertheim, H.F.L., Sumpradit, N., Vlieghe, E., Hara, G.L., Gould, I.M., Goossens, H., et al. (2013). Antibiotic resistance-the need for global solutions. Lancet Infect. Dis. 10.1016/S1473-3099(13)70318-9.
2. Cars, O., Högberg, L.D., Murray, M., Nordberg, O., Sivaraman, S., Lundborg, C.S., So, A.D., and Tomson, G. (2008). Meeting the challenge of antibiotic resistance. BMJ.
3. Ley, R.E., Peterson, D.A., and Gordon, J.I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell. 10.1016/j.cell.2006.02.017.
4. Stecher, B., Denzler, R., Maier, L., Bernet, F., Sanders, M.J., Pickard, D.J., Barthel, M., Westendorf, A.M., Krogfelt, K.A., Walker, A.W., et al. (2012). Gut inflammation can boost horizontal gene transfer between pathogenic and commensal Enterobacteriaceae. Proc. Natl. Acad. Sci. *109*, 1269–1274. 10.1073/pnas.1113246109.
5. Forster, S.C., Liu, J., Kumar, N., Gulliver, E.L., Gould, J.A., Escobar-Zepeda, A., Mkandawire, T., Pike, L.J., Shao, Y., Stares, M.D., et al. (2022). Strain-level characterization of broad host range mobile genetic elements transferring antibiotic resistance from the human microbiome. Nat. Commun. *13*. 10.1038/s41467-022-29096-9.
6. Hu, Y., Yang, X., Qin, J., Lu, N., Cheng, G., Wu, N., Pan, Y., Li, J., Zhu, L., Wang, X., et al. (2013). Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. Nat. Commun. *4*. 10.1038/ncomms3151.
7. Feng, J., Li, B., Jiang, X., Yang, Y., Wells, G.F., Zhang, T., and Li, X. (2018). Antibiotic resistome in a large-scale healthy human gut microbiota deciphered by metagenomic and network analyses. Environ. Microbiol. 10.1111/1462-2920.14009.
8. Bengtsson-Palme, J., Angelin, M., Huss, M., Kjellqvist, S., Kristiansson, E., Palmgren, H., Joakim Larsson, D.G., and Johansson, A. (2015). The human gut microbiome as a transporter of antibiotic resistance genes between continents. Antimicrob. Agents Chemother. 10.1128/AAC.00933-15.
9. Pehrsson, E.C., Tsukayama, P., Patel, S., Mejía-Bautista, M., Sosa-Soto, G.,

Navarrete, K.M., Calderon, M., Cabrera, L., Hoyos-Arango, W., Bertoli, M.T., et al. (2016). Interconnected microbiomes and resistomes in low-income human habitats. Nature. 10.1038/nature17672.

10. Yassour, M., Jason, E., Hogstrom, L.J., Arthur, T.D., Tripathi, S., Siljander, H., Selvenius, J., Oikarinen, S., Hyöty, H., Virtanen, S.M., et al. (2018). Strain-Level Analysis of Mother-to-Child Bacterial Transmission during the First Few Months of Life. Cell Host Microbe. 10.1016/j.chom.2018.06.007.
11. Ferretti, P., Pasolli, E., Tett, A., Asnicar, F., Gorfer, V., Fedi, S., Armanini, F., Truong, D.T., Manara, S., Zolfo, M., et al. (2018). Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. Cell Host Microbe. 10.1016/j.chom.2018.06.005.
12. Yassour, M., Vatanen, T., Siljander, H., Hämäläinen, A.M., Härkönen, T., Ryhänen, S.J., Franzosa, E.A., Vlamakis, H., Huttenhower, C., Gevers, D., et al. (2016). Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. Sci. Transl. Med. 10.1126/scitranslmed.aad0917.
13. Thorsen, J., McCauley, K., Fadrosh, D., Lynch, K., Barnes, K.L., Bendixsen, C.G., Seroogy, C.M., Lynch, S. V., and Gern, J.E. (2019). Evaluating the Effects of Farm Exposure on Infant Gut Microbiome. J. Allergy Clin. Immunol. 10.1016/j.jaci.2018.12.911.
14. Fehr, K., Moossavi, S., Sbihi, H., Boutin, R.C.T., Bode, L., Robertson, B., Yonemitsu, C., Field, C.J., Becker, A.B., Mandhane, P.J., et al. (2020). Breastmilk Feeding Practices Are Associated with the Co-Occurrence of Bacteria in Mothers' Milk and the Infant Gut: the CHILD Cohort Study. Cell Host Microbe. 10.1016/j.chom.2020.06.009.
15. Stokholm, J., Thorsen, J., Blaser, M.J., Rasmussen, M.A., Hjelmsø, M., Shah, S., Christensen, E.D., Chawes, B.L., Bønnelykke, K., Brix, S., et al. (2020). Delivery mode and gut microbial changes correlate with an increased risk of childhood asthma. Sci. Transl. Med. 10.1126/scitranslmed.aax9929.
16. Li, X., Stokholm, J., Brejnrod, A., Alberg Vestergaard, G., Russel, J., Trivedi, U., Thorsen, J., Gupta, S., Hjort Hjelmsø, M., A Shah, S., et al. (2021). The infant gut resistome associates with E. coli, environmental exposures, gut microbiome maturity, and asthma-associated bacterial composition. Cell Host Microbe, 1–13. 10.1016/j.chom.2021.03.017.

17. Kundu, P., Blacher, E., Elinav, E., and Pettersson, S. (2017). Our Gut Microbiome: The Evolving Inner Self. Cell. 10.1016/j.cell.2017.11.024.
18. Costello, E.K., Stagaman, K., Dethlefsen, L., Bohannan, B.J.M., and Relman, D.A. (2012). The application of ecological theory toward an understanding of the human microbiome. Science (80-.). 10.1126/science.1224203.
19. Yatsunenko, T., Rey, F.E., Manary, M.J., Trehan, I., Dominguez-Bello, M.G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R.N., Anokhin, A.P., et al. (2012). Human gut microbiome viewed across age and geography. Nature *486*, 222–227. 10.1038/nature11053.
20. Lozupone, C.A., Stombaugh, J.I., Gordon, J.I., Jansson, J.K., and Knight, R. (2012). Diversity , stability and resilience of the human gut microbiota. 10.1038/nature11550.
21. Hildebrand, F., Gossmann, T.I., Frioux, C., Özkurt, E., Myers, P.N., Ferretti, P., Kuhn, M., Bahram, M., Nielsen, H.B., and Bork, P. (2021). Dispersal strategies shape persistence and evolution of human gut bacteria. Cell Host Microbe. 10.1016/j.chom.2021.05.008.
22. Reijnders, D., Goossens, G.H., Hermes, G.D.A., Neis, E.P.J.G., van der Beek, C.M., Most, J., Holst, J.J., Lenaerts, K., Kootte, R.S., Nieuwdorp, M., et al. (2016). Effects of Gut Microbiota Manipulation by Antibiotics on Host Metabolism in Obese Humans: A Randomized Double-Blind Placebo-Controlled Trial. Cell Metab. 10.1016/j.cmet.2016.06.016.
23. Palleja, A., Mikkelsen, K.H., Forslund, S.K., Kashani, A., Allin, K.H., Nielsen, T., Hansen, T.H., Liang, S., Feng, Q., Zhang, C., et al. (2018). Recovery of gut microbiota of healthy adults following antibiotic exposure. Nat. Microbiol. 10.1038/s41564-018-0257-9.
24. Reyman, M., van Houten, M.A., Watson, R.L., Chu, M.L.J.N., Arp, K., de Waal, W.J., Schiering, I., Plötz, F.B., Willems, R.J.L., van Schaik, W., et al. (2022). Effects of early-life antibiotics on the developing infant gut microbiome and resistome: a randomized trial. Nat. Commun. *13*, 1–12. 10.1038/s41467-022-28525-z.
25. Anthony, W.E., Wang, B., Sukhum, K. V., D'Souza, A.W., Hink, T., Cass, C., Seiler, S., Reske, K.A., Coon, C., Dubberke, E.R., et al. (2022). Acute and persistent effects of commonly used antibiotics on the gut microbiome and resistome in healthy adults. Cell Rep. *39*, 110649.

10.1016/j.celrep.2022.110649.

26. Ng, K.M., Aranda-Díaz, A., Tropini, C., Frankel, M.R., Van Treuren, W., O'Laughlin, C.T., Merrill, B.D., Yu, F.B., Pruss, K.M., Oliveira, R.A., et al. (2019). Recovery of the Gut Microbiota after Antibiotics Depends on Host Diet, Community Context, and Environmental Reservoirs. Cell Host Microbe. 10.1016/j.chom.2019.10.011.

27. Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. Nat. Methods. 10.1038/nmeth.2066.

28. Bokulich, N.A., Chung, J., Battaglia, T., Henderson, N., Jay, M., Li, H., Lieber, A.D., Wu, F., Perez-Perez, G.I., Chen, Y., et al. (2016). Antibiotics, birth mode, and diet shape microbiome maturation during early life. Sci. Transl. Med. 10.1126/scitranslmed.aad7121.

29. Iredell, J., Brown, J., and Tagg, K. (2016). Antibiotic resistance in Enterobacteriaceae: Mechanisms and clinical implications. BMJ. 10.1136/bmj.h6420.

30. Moradigaravand, D., Palm, M., Farewell, A., Mustonen, V., Warringer, J., and Parts, L. (2018). Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data. PLoS Comput. Biol. 10.1371/journal.pcbi.1006258.

31. Kumar, V., Sun, P., Vamathevan, J., Li, Y., Ingraham, K., Palmer, L., Huang, J., and Brown, J.R. (2011). Comparative genomics of Klebsiella pneumoniae strains with different antibiotic resistance profiles. Antimicrob. Agents Chemother. 10.1128/AAC.00052-11.

32. De Oliveira, D.M.P., Forde, B.M., Kidd, T.J., Harris, P.N.A., Schembri, M.A., Beatson, S.A., Paterson, D.L., and Walker, M.J. (2020). Antimicrobial resistance in ESKAPE pathogens. Clin. Microbiol. Rev. 10.1128/CMR.00181-19.

33. Gibson, M.K., Wang, B., Ahmadi, S., Burnham, C.A.D., Tarr, P.I., Warner, B.B., and Dantas, G. (2016). Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. Nat. Microbiol. 10.1038/nmicrobiol.2016.24.

34. Wellington, E.M.H., Boxall, A.B.A., Cross, P., Feil, E.J., Gaze, W.H., Hawkey, P.M., Johnson-Rollings, A.S., Jones, D.L., Lee, N.M., Otten, W., et al. (2013). The role of the natural environment in the emergence of antibiotic resistance in Gram-negative bacteria. Lancet Infect. Dis. 10.1016/S1473-3099(12)70317-1.

35. Stecher, B., Maier, L., and Hardt, W.D. (2013). "Blooming" in the gut: How

dysbiosis might contribute to pathogen evolution. Nat. Rev. Microbiol. *11*, 277–284. 10.1038/nrmicro2989.

36. Doucet-Populaire, F., Trieu-Cuot, P., Dosbaa, I., Andremont, A., and Courvalin, P. (1991). Inducible transfer of conjugative transposon Tn1545 from Enterococcus faecalis to Listeria monocytogenes in the digestive tracts of gnotobiotic mice. Antimicrob. Agents Chemother. 10.1128/AAC.35.1.185.
37. Jones, B. V., Sun, F., and Marchesi, J.R. (2010). Comparative metagenomic analysis of plasmid encoded functions in the human gut microbiome. BMC Genomics. 10.1186/1471-2164-11-46.
38. Dagan, T., Artzy-Randrup, Y., and Martin, W. (2008). Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc. Natl. Acad. Sci. U. S. A. 10.1073/pnas.0800679105.
39. Raymond, F., Ouameur, A.A., Déraspe, M., Iqbal, N., Gingras, H., Dridi, B., Leprohon, P., Plante, P.L., Giroux, R., Bérubé, È., et al. (2016). The initial state of the human gut microbiome determines its reshaping by antibiotics. ISME J. 10.1038/ismej.2015.148.
40. Zhu, Y.-G., Johnson, T.A., Su, J.-Q., Qiao, M., Guo, G.-X., Stedtfeld, R.D., Hashsham, S.A., and Tiedje, J.M. (2013). Diverse and abundant antibiotic resistance genes in Chinese swine farms. Proc. Natl. Acad. Sci. *110*, 3435–3440. 10.1073/pnas.1222743110.
41. Maier, L., Pruteanu, M., Kuhn, M., Zeller, G., Telzerow, A., Anderson, E.E., Brochado, A.R., Fernandez, K.C., Dose, H., Mori, H., et al. (2018). Extensive impact of non-antibiotic drugs on human gut bacteria. Nature. 10.1038/nature25979.
42. Pärnänen, K., Karkman, A., Hultman, J., Lyra, C., Bengtsson-Palme, J., Larsson, D.G.J., Rautava, S., Isolauri, E., Salminen, S., Kumar, H., et al. (2018). Maternal gut and breast milk microbiota affect infant gut antibiotic resistome and mobile genetic elements. Nat. Commun. 10.1038/s41467-018-06393-w.
43. Sosa-Moreno, A., Comstock, S.S., Sugino, K.Y., Ma, T.F., Paneth, N., Davis, Y., Olivero, R., Schein, R., Maurer, J., and Zhang, L. (2020). Perinatal risk factors for fecal antibiotic resistance gene patterns in pregnant women and their infants. PLoS One. 10.1371/journal.pone.0234751.
44. Gaire, T.N., Scott, H.M., Sellers, L., Nagaraja, T.G., and Volkova, V. V. (2021). Age Dependence of Antimicrobial Resistance Among Fecal Bacteria in Animals:

A Scoping Review. Front. Vet. Sci. 10.3389/fvets.2020.622495.

45. Mueller, N.T., Bakacs, E., Combellick, J., Grigoryan, Z., and Dominguez-Bello, M.G. (2015). The infant microbiome development: Mom matters. Trends Mol. Med. 10.1016/j.molmed.2014.12.002.

46. Pantoja-Feliciano, I.G., Clemente, J.C., Costello, E.K., Perez, M.E., Blaser, M.J., Knight, R., and Dominguez-Bello, M.G. (2013). Biphasic assembly of the murine intestinal microbiota during early development. ISME J. 10.1038/ismej.2013.15.

47. Guittar, J., Shade, A., and Litchman, E. (2019). Trait-based community assembly and succession of the infant gut microbiome. Nat. Commun. 10.1038/s41467-019-08377-w.

48. Li, H., Limenitakis, J.P., Fuhrer, T., Geuking, M.B., Lawson, M.A., Wyss, M., Brugiroux, S., Keller, I., Macpherson, J.A., Rupp, S., et al. (2015). The outer mucus layer hosts a distinct intestinal microbial niche. Nat. Commun. 10.1038/ncomms9292.

49. Round, J.L., and Mazmanian, S.K. (2009). The gut microbiota shapes intestinal immune responses during health and disease. Nat. Rev. Immunol. 10.1038/nri2515.

50. Patangia, D. V., Ryan, C.A., Dempsey, E., Stanton, C., and Ross, R.P. (2022). Vertical transfer of antibiotics and antibiotic resistant strains across the mother/baby axis. Trends Microbiol. 10.1016/j.tim.2021.05.006.

51. Bottery, M.J. (2022). Ecological dynamics of plasmid transfer and persistence in microbial communities. Curr. Opin. Microbiol. *68*, 102152. 10.1016/j.mib.2022.102152.

52. Lerminiaux, N.A., and Cameron, A.D.S. (2019). Horizontal transfer of antibiotic resistance genes in clinical environments. Can. J. Microbiol. 10.1139/cjm-2018-0275.

53. Gumpert, H., Kubicek-Sutherland, J.Z., Porse, A., Karami, N., Munck, C., Linkevicius, M., Adlerberth, I., Wold, A.E., Andersson, D.I., and Sommer, M.O.A. (2017). Transfer and persistence of a multi-drug resistance plasmid in situ of the infant gut microbiota in the absence of antibiotic treatment. Front. Microbiol. 10.3389/fmicb.2017.01852.

54. Lopatkin, A.J., Meredith, H.R., Srimani, J.K., Pfeiffer, C., Durrett, R., and You, L. (2017). Persistence and reversal of plasmid-mediated antibiotic resistance. Nat. Commun. 10.1038/s41467-017-01532-1.

55. Kostic, A.D., Gevers, D., Siljander, H., Vatanen, T., Hyötyläinen, T., Hämäläinen, A.M., Peet, A., Tillmann, V., Pöhö, P., Mattila, I., et al. (2015). The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. Cell Host Microbe. 10.1016/j.chom.2015.01.001.
56. Shekhar, S., and Petersen, F.C. (2020). The Dark Side of Antibiotics: Adverse Effects on the Infant Immune Defense Against Infection. Front. Pediatr. 10.3389/fped.2020.544460.
57. Korpela, K., Salonen, A., Virta, L.J., Kekkonen, R.A., Forslund, K., Bork, P., and De Vos, W.M. (2016). Intestinal microbiome is related to lifetime antibiotic use in Finnish pre-school children. Nat. Commun. 10.1038/ncomms10410.
58. Ribeiro, C.F.A., Silveira, G.G.D.O.S., Cândido, E.D.S., Cardoso, M.H., Espínola Carvalho, C.M., and Franco, O.L. (2020). Effects of Antibiotic Treatment on Gut Microbiota and How to Overcome Its Negative Impacts on Human Health. ACS Infect. Dis. 10.1021/acsinfecdis.0c00036.
59. Aabenhus, R., Siersma, V., Hansen, M.P., and Bjerrum, L. (2016). Antibiotic prescribing in Danish general practice 2004-13. J. Antimicrob. Chemother. 10.1093/jac/dkw117.
60. Jernberg, C., Löfmark, S., Edlund, C., and Jansson, J.K. (2010). Long-term impacts of antibiotic exposure on the human intestinal microbiota. Microbiology. 10.1099/mic.0.040618-0.
61. Simon, A.K., Hollander, G.A., and McMichael, A. (2015). Evolution of the immune system in humans from infancy to old age. Proc. R. Soc. B Biol. Sci. 10.1098/rspb.2014.3085.
62. Cox, L.M., Yamanishi, S., Sohn, J., Alekseyenko, A. V., Leung, J.M., Cho, I., Kim, S.G., Li, H., Gao, Z., Mahana, D., et al. (2014). Altering the intestinal microbiota during a critical developmental window has lasting metabolic consequences. Cell. 10.1016/j.cell.2014.05.052.
63. Cross, R., Ling, C., Day, N.P.J., Mcgready, R., and Paris, D.H. (2016). Revisiting doxycycline in pregnancy and early childhood - Time to rebuild its reputation? Expert Opin. Drug Saf. 10.1517/14740338.2016.1133584.
64. de Vries, L.E., Vallès, Y., Agersø, Y., Vaishampayan, P.A., García-Montaner, A., Kuehl, J. V., Christensen, H., Barlow, M., and Francino, M.P. (2011). The gut as reservoir of antibiotic resistance: Microbial diversity of tetracycline resistance in mother and infant. PLoS One. 10.1371/journal.pone.0021644.

65. Arboleya, S., Watkins, C., Stanton, C., and Ross, R.P. (2016). Gut bifidobacteria populations in human health and aging. Front. Microbiol. 10.3389/fmicb.2016.01204.
66. Derrien, M., Turroni, F., Ventura, M., and Sinderen, D. Van (2022). Insights into endogenous Bi fi dobacterium species in the human gut microbiota during adulthood. Trends Microbiol. *xx*, 1–8. 10.1016/j.tim.2022.04.004.
67. Schmidt, V., Enav, H., Spector, T.D., Youngblut, N.D., and Ley, R.E. (2020). Strain-Level Analysis of Bifidobacterium spp. from Gut Microbiomes of Adults with Differing Lactase Persistence Genotypes . mSystems. 10.1128/msystems.00911-20.
68. Oliver, A., Chase, A.B., Weihe, C., Orchanian, S.B., Riedel, S.F., Hendrickson, C.L., Lay, M., Sewall, J.M., Martiny, J.B.H., and Whiteson, K. (2021). High-Fiber, Whole-Food Dietary Intervention Alters the Human Gut Microbiome but Not Fecal Short-Chain Fatty Acids. mSystems. 10.1128/msystems.00115-21.
69. Hansson, S., Svedhem, Å., Wennerström, M., and Jodal, U. (2007). Urinary tract infection caused by Haemophilus influenzae and Haemophilus parainfluenzae in children. Pediatr. Nephrol. 10.1007/s00467-007-0531-1.
70. Cardines, R., Giufrè, M., Atti, M.L.C.D., Accogli, M., Mastrantonio, P., and Cerquetti, M. (2009). Haemophilus parainfluenzae meningitis in an adult associated with acute otitis media. New Microbiol.
71. Middleton, A.M., Dowling, R.B., Mitchell, J.L., Watanabe, S., Rutman, A., Pritchard, K., Tillotson, G., Hill, S.L., and Wilson, R. (2003). Haemophilus parainfluenzae infection of respiratory mucosa. Respir. Med. 10.1053/rmed.2002.1454.
72. Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermúdez-Humarán, L.G., Gratadoux, J.J., Blugeon, S., Bridonneau, C., Furet, J.P., Corthier, G., et al. (2008). Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. Proc. Natl. Acad. Sci. U. S. A. 10.1073/pnas.0804812105.
73. Miquel, S., Martín, R., Rossi, O., Bermúdez-Humarán, L.G., Chatel, J.M., Sokol, H., Thomas, M., Wells, J.M., and Langella, P. (2013). Faecalibacterium prausnitzii and human intestinal health. Curr. Opin. Microbiol. *16*, 255–261. 10.1016/j.mib.2013.06.003.
74. Rios-Covian, D., Gueimonde, M., Duncan, S.H., Flint, H.J., and De Los Reyes-

Gavilan, C.G. (2015). Enhanced butyrate formation by cross-feeding between Faecalibacterium prausnitzii and Bifidobacterium adolescentis. FEMS Microbiol. Lett. 10.1093/femsle/fnv176.

75. Tange, O. (2018). GNU Parallel 2018 dx.doi.org/10.5281/zenodo.1146014.
76. Joshi, N., and Fass, J. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle.
77. Rodriguez-R, L.M., Gunturu, S., Tiedje, J.M., Cole, J.R., and Konstantinidis, K.T. (2018). Nonpareil 3: Fast Estimation of Metagenomic Coverage and Sequence Diversity. mSystems. 10.1128/msystems.00039-18.
78. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 10.1089/cmb.2012.0021.
79. Nissen, J.N., Johansen, J., Allesøe, R.L., Sønderby, C.K., Armenteros, J.J.A., Grønbech, C.H., Jensen, L.J., Nielsen, H.B., Petersen, T.N., Winther, O., et al. (2021). Improved metagenome binning and assembly using deep variational autoencoders. Nat. Biotechnol. 10.1038/s41587-020-00777-4.
80. Chaumeil, P.A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2020). GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. Bioinformatics. 10.1093/bioinformatics/btz848.
81. Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 10.1186/1471-2105-11-119.
82. Jia, B., Raphenya, A.R., Alcock, B., Waglechner, N., Guo, P., Tsang, K.K., Lago, B.A., Dave, B.M., Pereira, S., Sharma, A.N., et al. (2017). CARD 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Res. 10.1093/nar/gkw1004.
83. Bisgaard, H. (2004). The Copenhagen Prospective Study on Asthma in Childhood (COPSAC): Design, rationale, and baseline data from a longitudinal birth cohort study. Ann. Allergy, Asthma Immunol. 10.1016/S1081-1206(10)61398-1.
84. Bisgaard, H., Stokholm, J., Chawes, B.L., Vissing, N.H., Bjarnadóttir, E., Schoos, A.M.M., Wolsk, H.M., Pedersen, T.M., Vinding, R.K., Thorsteinsdóttir, S., et al.

(2016). Fish oil-derived fatty acids in pregnancy and wheeze and asthma in offspring. N. Engl. J. Med. 10.1056/NEJMoa1503734.

85. Bisgaard, H., Vissing, N.H., Carson, C.G., Bischoff, A.L., Følsgaard, N. V., Kreiner-Møller, E., Chawes, B.L.K., Stokholm, J., Pedersen, L., Bjarnadóttir, E., et al. (2013). Deep phenotyping of the unselected COPSAC2010 birth cohort study. Clin. Exp. Allergy. 10.1111/cea.12213.

86. Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017). MetaSPAdes: A new versatile metagenomic assembler. Genome Res. 10.1101/gr.213959.116.

87. Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 10.1093/nar/gkt263.

88. Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: Towards a more sustainable future. Nucleic Acids Res. 10.1093/nar/gkv1344.

89. Riadi, G., Medina-Moenne, C., and Holmes, D.S. (2012). TnpPred: A web service for the robust prediction of prokaryotic transposases. Comp. Funct. Genomics *2012*. 10.1155/2012/678761.

90. Sáenz, J.S., Marques, T.V., Barone, R.S.C., Cyrino, J.E.P., Kublik, S., Nesme, J., Schloter, M., Rath, S., and Vestergaard, G. (2019). Oral administration of antibiotics increased the potential mobility of bacterial resistance genes in the gut of the fish Piaractus mesopotamicus. Microbiome. 10.1186/s40168-019-0632-7.

91. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods. 10.1038/nmeth.1923.

92. Li, X., Rensing, C., Vestergaard, G., Arumugam, M., Nesme, J., Gupta, S., Brejnrod, A.D., and Sørensen, S.J. (2022). Metagenomic evidence for co-occurrence of antibiotic, biocide and metal resistance genes in pigs. Environ. Int. 10.1016/j.envint.2021.106899.

93. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. 10.1093/bioinformatics/btp352.

94. Schwengers, O., Barth, P., Falgenhauer, L., Hain, T., Chakraborty, T., and

Goesmann, A. (2020). Platon: Identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. Microb. Genomics. 10.1099/mgen.0.000398.

95. Svetnik, V., Liaw, A., Tong, C., Christopher Culberson, J., Sheridan, R.P., and Feuston, B.P. (2003). Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. J. Chem. Inf. Comput. Sci. 10.1021/ci034160g.

96. Liaw, a, and Wiener, M. (2002). Classification and Regression by randomForest. R news *2*, 18–22. 10.1177/154405910408300516.

97. Gower, J.C. (2015). Procrustes Analysis. In International Encyclopedia of the Social & Behavioral Sciences: Second Edition 10.1016/B978-0-08-097086-8.43078-3.

98. Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat. Commun. 10.1038/s41467-018-07641-9.

99. Criscuolo, A., and Gascuel, O. (2008). Fast NJ-like algorithms to deal with incomplete distance matrices. BMC Bioinformatics. 10.1186/1471-2105-9-166.

100. Dixon, P. (2003). VEGAN, a package of R functions for community ecology. J. Veg. Sci. *14*, 927–930. 10.1111/j.1654-1103.2003.tb02228.x.

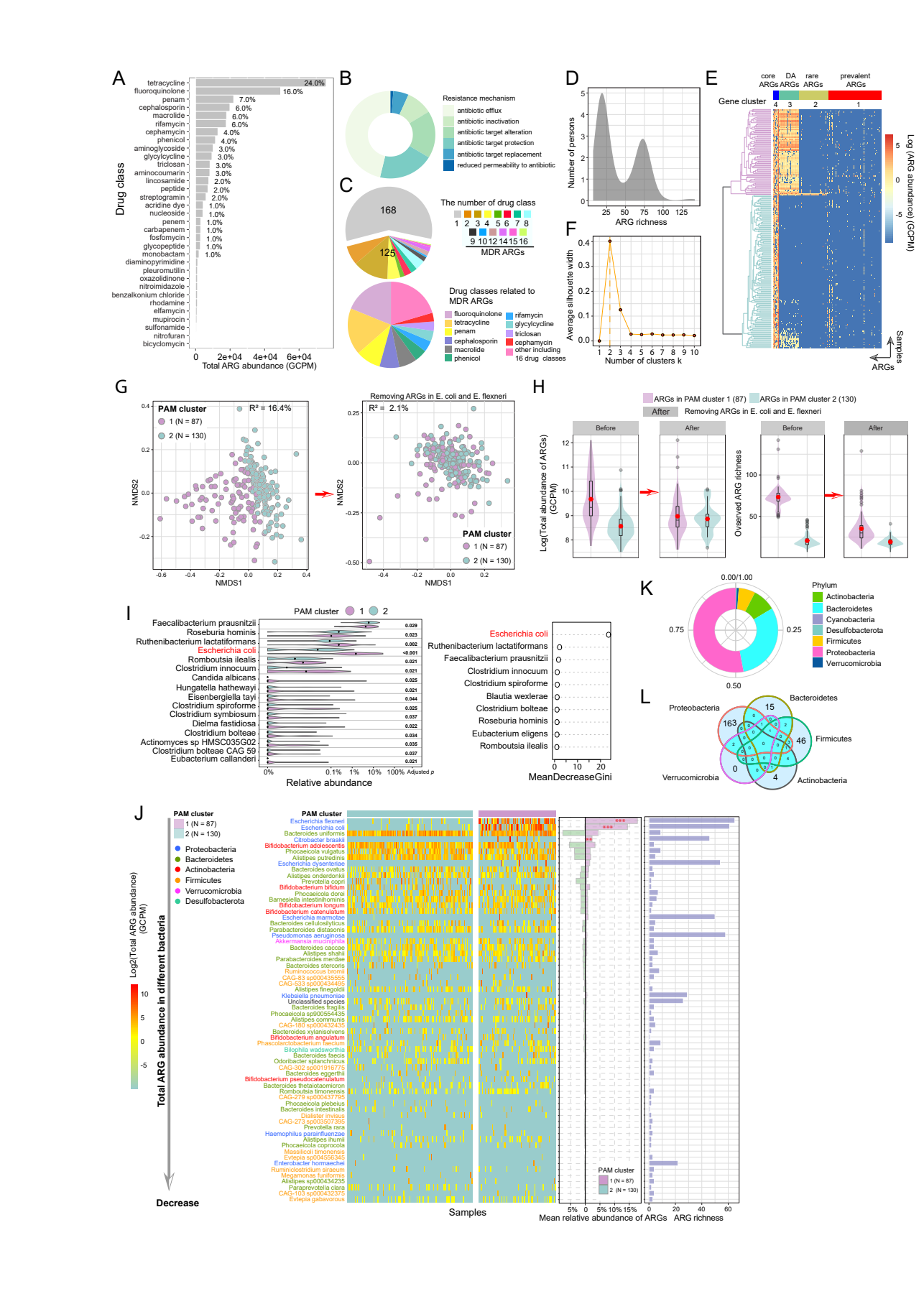
101. McMurdie, P.J., and Holmes, S. (2013). Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS One *8*. 10.1371/journal.pone.0061217.

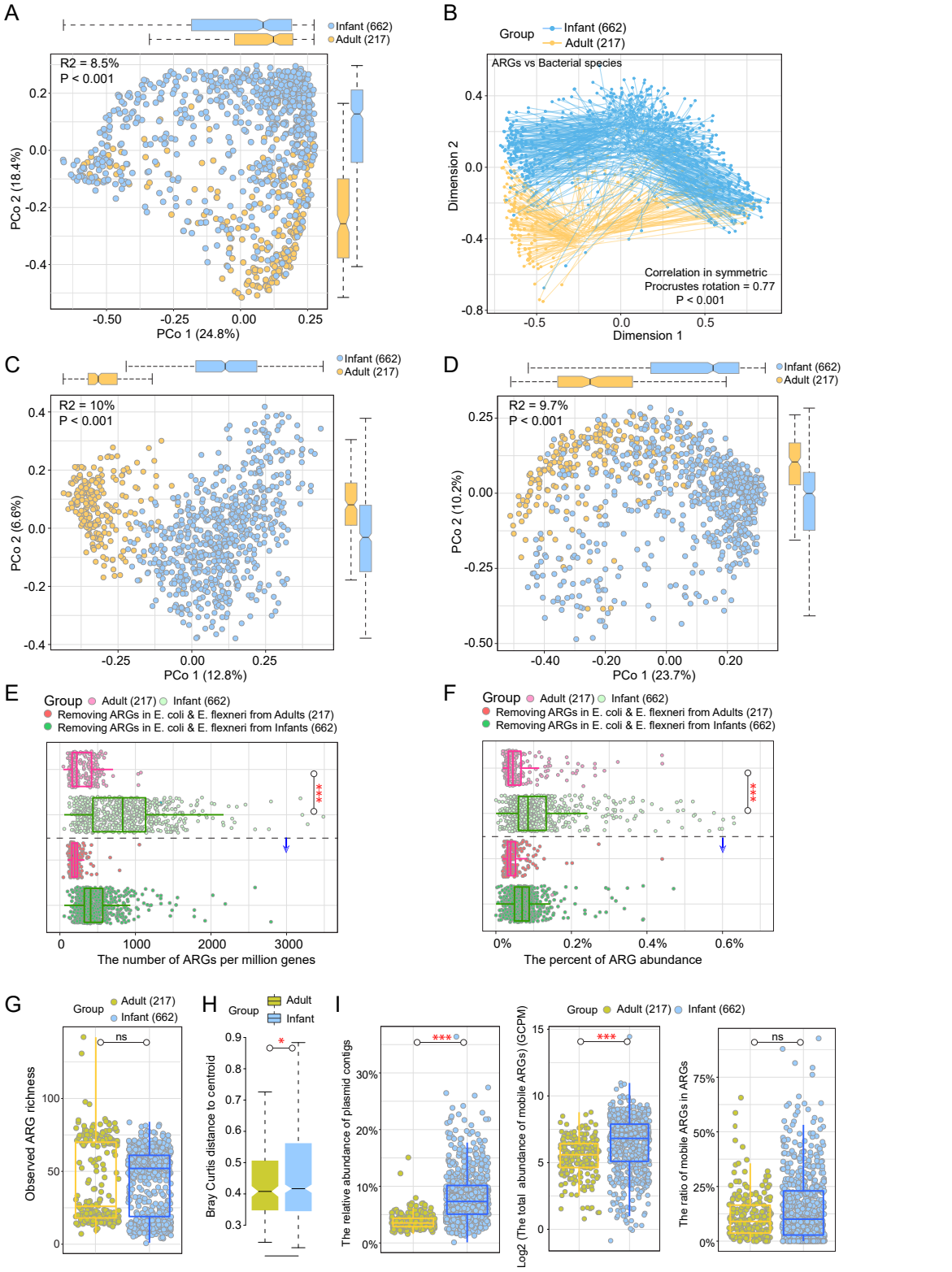
102. Reynolds, A.P., Richards, G., De La Iglesia, B., and Rayward-Smith, V.J. (2006). Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. J. Math. Model. Algorithms. 10.1007/s10852-005-9022-1.

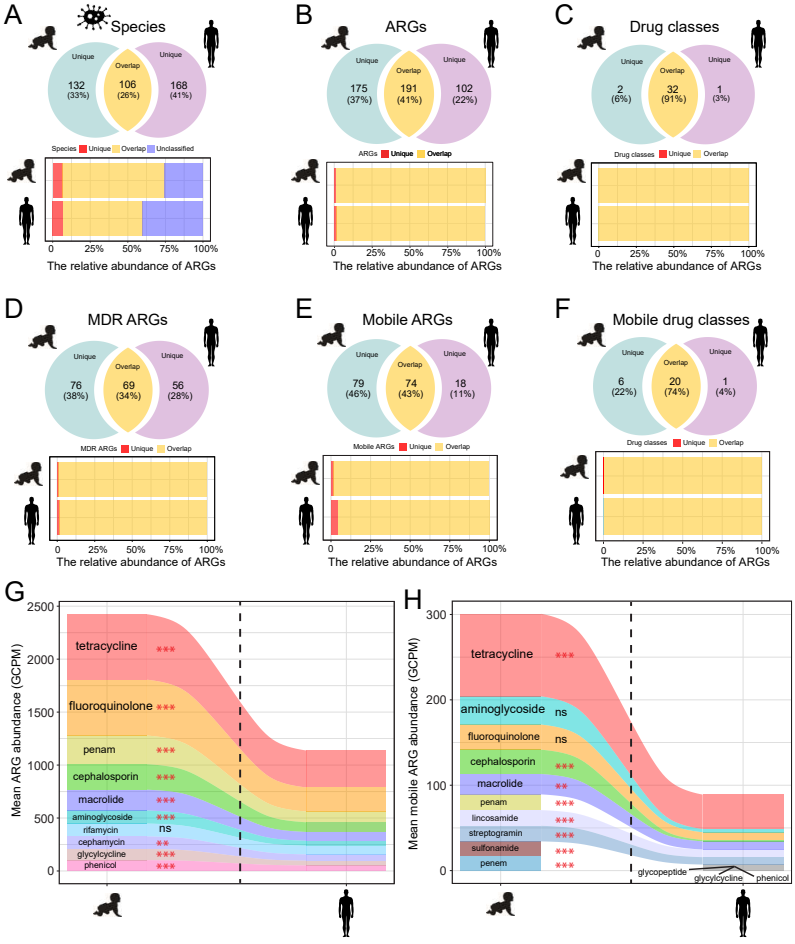
103. Martin, M., Rousseeuw, P., Struyf, A., Hubert, M., Studer, M., Roudier, P., and Gonzalez, J. (2017). Finding Groups in Data : Cluster Analysis Extended Rousseeuw et al. Cran. ISBN 0-387-95457-0.

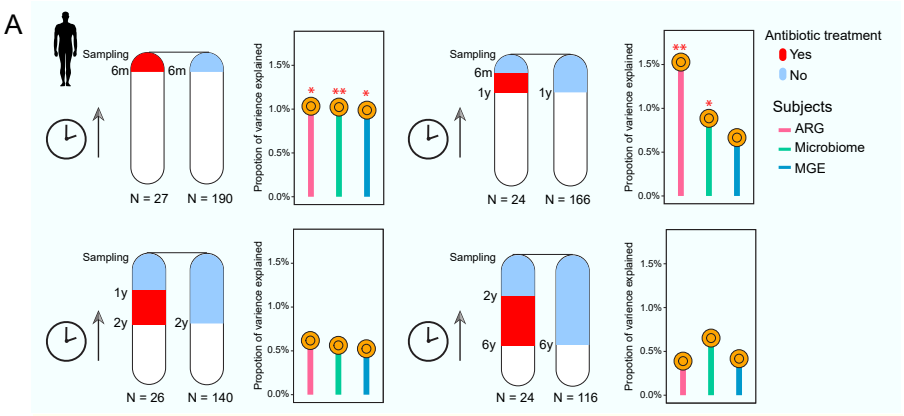
104. Kassambara, A., and Mundt, F. (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. Package Version 1.0.7. R Packag. version.

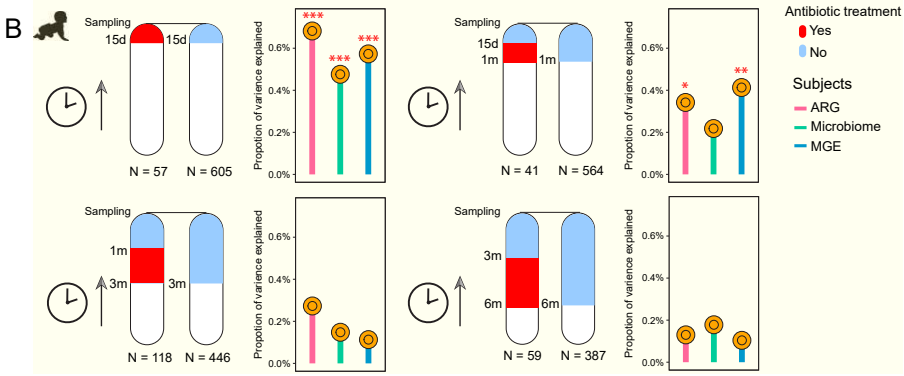
105. core Team, R. (2018). R: A Language and Environment for Statistical Computing. R Found. Stat. Comput. Vienna, Austria.

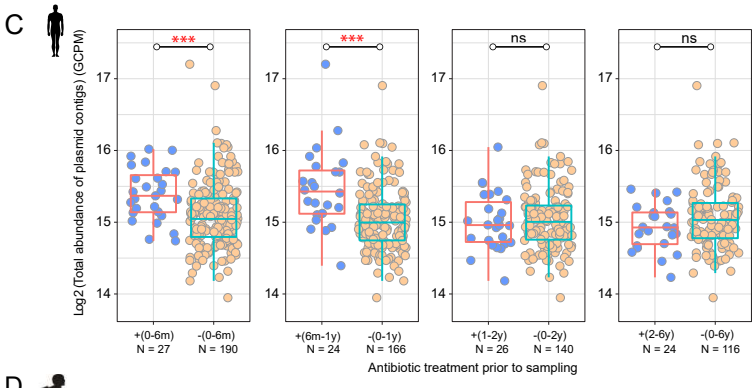

A
Drug class
tetracycline
fluoroquinolone
penam
cephalosporin
macrolide
rifamycin
cephamycin
phenicol
aminoglycoside
glycylcycline
triclosan
aminocoumarin
lincosamide
peptide
streptogramin
acridine dye
nucleoside
penem
carbapenem
fosfomycin
glycopeptide
monobactam
diaminopyrimidine
pleuromutilin
oxazolidinone
nitroimidazole
benzalkonium chloride
rhodamine
elfamycin
mupirocin
sulfonamide
nitrofuran
bicyclomycin
24.0%
16.0%
7.0%
6.0%
6.0%
6.0%
4.0%
4.0%
3.0%
3.0%
3.0%
3.0%
3.0%
2.0%
2.0%
2.0%
1.0%
1.0%
1.0%
1.0%
1.0%
1.0%
1.0%
0
2e+04
4e+04
6e+04
Total ARG abundance (GCPM)
B
Resistance mechanism
antibiotic efflux
antibiotic inactivation
antibiotic target alteration
antibiotic target protection
antibiotic target replacement
reduced permeability to antibiotic
C
168
125
The number of drug class
1 2 3 4 5 6 7 8
9 10 12 14 15 16
MDR ARGs
Drug classes related to MDR ARGs
fluoroquinolone
tetracycline
penam
cephalosporin
macrolide
phenicol
rifamycin
glycylcycline
triclosan
cephamycin
other including 16 drug classes
D
Number of persons
ARG richness
E
core ARGs
DA ARGs
rare ARGs
prevalent ARGs
Gene cluster
Log (ARG abundance) (GCPM)
Samples
ARGs
F
Average silhouette width
Number of clusters k
G
PAM cluster
1 (N = 87)
2 (N = 130)
R² = 16.4%
NMDS1
NMDS2
Removing ARGs in E. coli and E. flexneri
R² = 2.1%
H
ARGs in PAM cluster 1 (87)
ARGs in PAM cluster 2 (130)
After
Removing ARGs in E. coli and E. flexneri
Before
Log(Total abundance of ARGs) (GCPM)
Observed ARG richness
I
PAM cluster 1 2
Faecalibacterium prausnitzii
Roseburia hominis
Ruthenibacterium lactatiformans
Escherichia coli
Romboutsia ilealis
Clostridium innocuum
Candida albicans
Hungatella hathewayi
Eisenbergiella tayi
Clostridium spiroforme
Clostridium symbiosum
Dielma fastidiosa
Clostridium bolteae
Actinomyces sp HMSC035G02
Clostridium bolteae CAG 59
Eubacterium callanderi
Adjusted p
Relative abundance
Escherichia coli
Ruthenibacterium lactatiformans
Faecalibacterium prausnitzii
Clostridium innocuum
Clostridium spiroforme
Blautia wexlerae
Clostridium bolteae
Roseburia hominis
Eubacterium eligens
Romboutsia ilealis
MeanDecreaseGini
K
0.00/1.00
0.25
0.50
0.75
Phylum
Actinobacteria
Bacteroidetes
Cyanobacteria
Desulfobacterota
Firmicutes
Proteobacteria
Verrucomicrobia
L
Proteobacteria
Bacteroidetes
Firmicutes
Verrucomicrobia
Actinobacteria
15
163
46
0
4
J
PAM cluster
1 (N = 87)
2 (N = 130)
Proteobacteria
Bacteroidetes
Actinobacteria
Firmicutes
Verrucomicrobia
Desulfobacterota
Log2(Total ARG abundance) (GCPM)
Total ARG abundance in different bacteria
Decrease
Escherichia flexneri
Escherichia coli
Bacteroides uniformis
Citrobacter braakii
Bifidobacterium adolescentis
Phocaeicola vulgatus
Alistipes putredinis
Escherichia dysenteriae
Bacteroides ovatus
Alistipes onderdonkii
Prevotella copri
Bifidobacterium bifidum
Phocaeicola dorei
Barnesiella intestinihominis
Bifidobacterium longum
Bifidobacterium catenulatum
Escherichia marmotae
Bacteroides cellulosilyticus
Parabacteroides distasonis
Pseudomonas aeruginosa
Akkermansia muciniphila
Bacteroides caccae
Alistipes shahii
Parabacteroides merdae
Bacteroides stercoris
Ruminococcus bromii
CAG-83 sp000435555
CAG-533 sp000434495
Alistipes finegoldii
Klebsiella pneumoniae
Unclassified species
Bacteroides fragilis
Phocaeicola sp000554435
Alistipes communis
CAG-180 sp000432435
Bacteroides xylanisolvens
Bifidobacterium angulatum
Phascolarctobacterium faecium
Bilophila wadsworthia
Bacteroides faecis
Odoribacter splanchnicus
CAG-302 sp001916775
Bacteroides eggerthii
Bifidobacterium pseudocatenulatum
Bacteroides thetaiotaomicron
Romboutsia timonensis
CAG-279 sp000437795
Phocaeicola plebeius
Bacteroides intestinalis
Dialster invisus
CAG-273 sp003507395
Prevotella rara
Haemophilus parainfluenzae
Alistipes ihumii
Phocaeicola coprocola
Massiliicoli timonensis
Evtepia sp004556345
Enterobacter hormaechei
Ruminiclostridium siraeum
Megamonas funiformis
Alistipes sp000434235
Paraprevotella clara
CAG-103 sp000432375
Evtepia gabavorous
Samples
Mean relative abundance of ARGs
ARG richness
PAM cluster
1 (N = 87)
2 (N = 130)

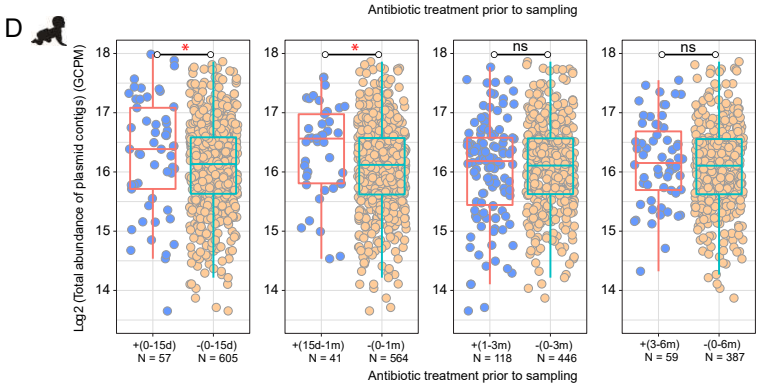

A
Infant (662)
Adult (217)
R2 = 8.5%
P < 0.001
PCo 2 (18.4%)
PCo 1 (24.8%)
B
Group
Infant (662)
Adult (217)
ARGs vs Bacterial species
Dimension 2
Dimension 1
Correlation in symmetric
Procrustes rotation = 0.77
P < 0.001
C
Infant (662)
Adult (217)
R2 = 10%
P < 0.001
PCo 2 (6.6%)
PCo 1 (12.8%)
D
Infant (662)
Adult (217)
R2 = 9.7%
P < 0.001
PCo 2 (10.2%)
PCo 1 (23.7%)
E
Group Adult (217) Infant (662)
Removing ARGs in E. coli & E. flexneri from Adults (217)
Removing ARGs in E. coli & E. flexneri from Infants (662)
***
The number of ARGs per million genes
F
Group Adult (217) Infant (662)
Removing ARGs in E. coli & E. flexneri from Adults (217)
Removing ARGs in E. coli & E. flexneri from Infants (662)
***
The percent of ARG abundance
G
Group
Adult (217)
Infant (662)
ns
Observed ARG richness
H
Group
Adult
Infant
*
Bray Curtis distance to centroid
I
***
The relative abundance of plasmid contigs
Group Adult (217) Infant (662)
***
Log2 (The total abundance of mobile ARGs) (GCPM)
ns
The ratio of mobile ARGs in ARGs

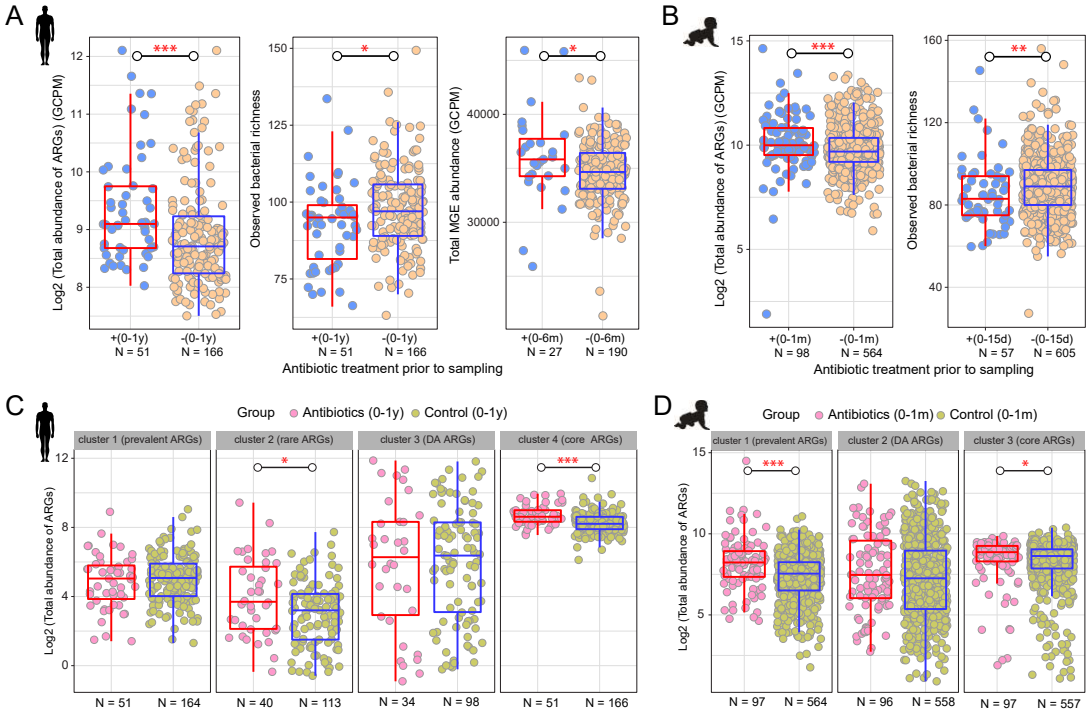

A
Species
Unique
Overlap
Unique
132
(33%)
106
(26%)
168
(41%)
Species Unique Overlap Unclassified
0 25% 50% 75% 100%
The relative abundance of ARGs
B
ARGs
175
(37%)
191
(41%)
102
(22%)
ARGs Unique Overlap
The relative abundance of ARGs
C
Drug classes
2
(6%)
32
(91%)
1
(3%)
Drug classes Unique Overlap
The relative abundance of ARGs
D
MDR ARGs
76
(38%)
69
(34%)
56
(28%)
MDR ARGs Unique Overlap
The relative abundance of ARGs
E
Mobile ARGs
79
(46%)
74
(43%)
18
(11%)
Mobile ARGs Unique Overlap
The relative abundance of ARGs
F
Mobile drug classes
6
(22%)
20
(74%)
1
(4%)
Drug classes Unique Overlap
The relative abundance of ARGs
G
Mean ARG abundance (GCPM)
2500
2000
1500
1000
500
0
tetracycline ***
fluoroquinolone ***
penam ***
cephalosporin ***
macrolide ***
aminoglycoside ***
rifamycin ns
cephamycin **
glycylcycline ***
phenicol ***
H
Mean mobile ARG abundance (GCPM)
300
200
100
0
tetracycline ***
aminoglycoside ns
fluoroquinolone ns
cephalosporin ***
macrolide **
penam ***
lincosamide ***
streptogramin ***
sulfonamide ***
penem ***
glycopeptide
glycylcycline
phenicol

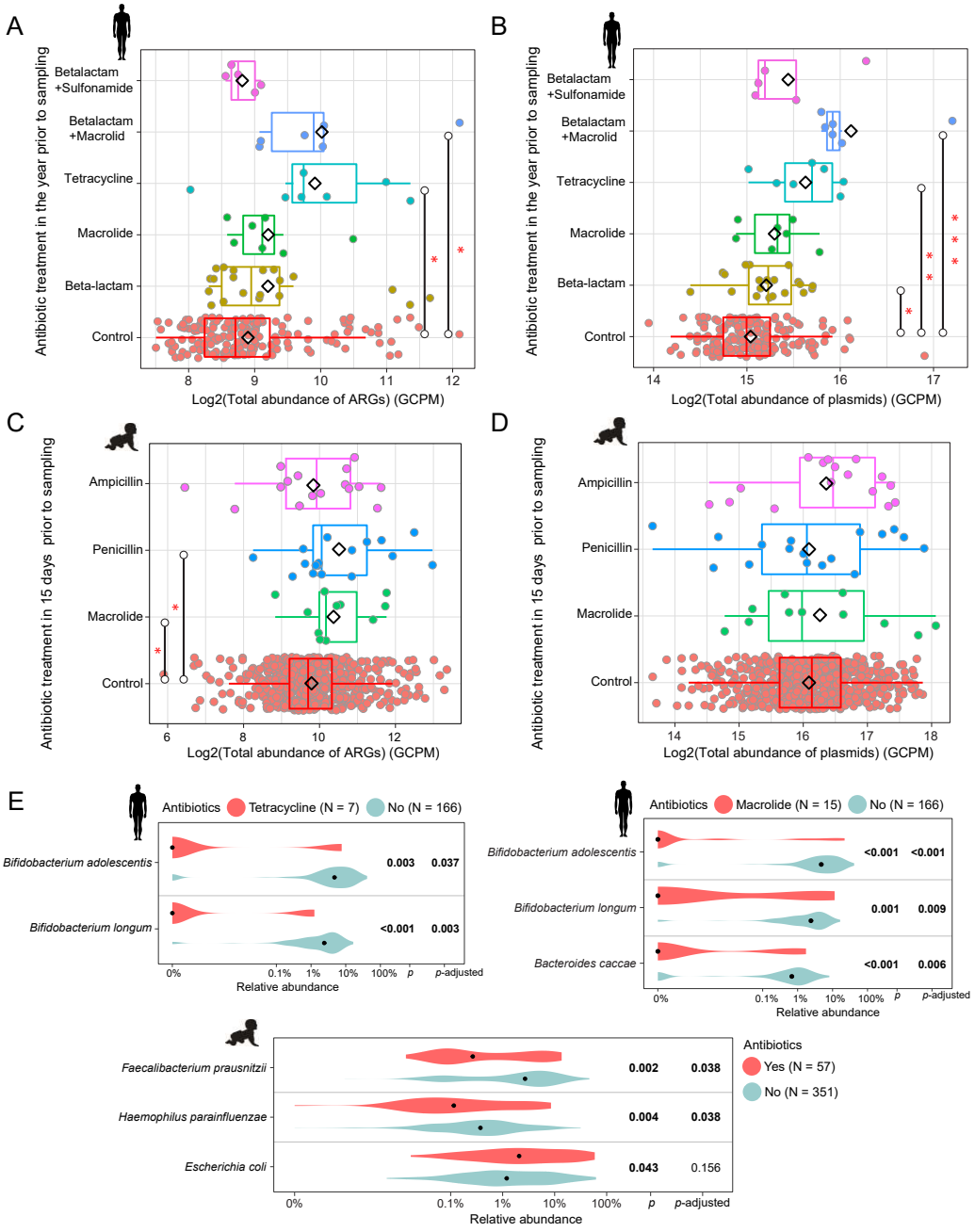

A
Sampling
6m
6m
N = 27
N = 190
Propotion of varience explained
1.5%
1.0%
0.5%
0.0%
*
**
*
Sampling
6m
1y
1y
N = 24
N = 166
**
*
Antibiotic treatment
Yes
No
Subjects
ARG
Microbiome
MGE
Sampling
1y
2y
2y
N = 26
N = 140
Sampling
2y
6y
6y
N = 24
N = 116


B
Sampling
15d
15d
N = 57
N = 605
Propotion of varience explained
0.6%
0.4%
0.2%
0.0%
***
***
***
Sampling
15d
1m
1m
N = 41
N = 564
*
**
Antibiotic treatment
Yes
No
Subjects
ARG
Microbiome
MGE
Sampling
1m
3m
3m
N = 118
N = 446
Sampling
3m
6m
6m
N = 59
N = 387


C
Log2 (Total abundance of plasmid contigs) (GCPM)
17
16
15
14
***
***
ns
ns
+(0-6m)
N = 27
-(0-6m)
N = 190
+(6m-1y)
N = 24
-(0-1y)
N = 166
+(1-2y)
N = 26
-(0-2y)
N = 140
+(2-6y)
N = 24
-(0-6y)
N = 116
Antibiotic treatment prior to sampling


D
Log2 (Total abundance of plasmid contigs) (GCPM)
18
17
16
15
14
*
*
ns
ns
+(0-15d)
N = 57
-(0-15d)
N = 605
+(15d-1m)
N = 41
-(0-1m)
N = 564
+(1-3m)
N = 118
-(0-3m)
N = 446
+(3-6m)
N = 59
-(0-6m)
N = 387
Antibiotic treatment prior to sampling


A
Log2 (Total abundance of ARGs) (GCPM)
Observed bacterial richness
Total MGE abundance (GCPM)
+(0-1y) N = 51
-(0-1y) N = 166
+(0-6m) N = 27
-(0-6m) N = 190
Antibiotic treatment prior to sampling
B
+(0-1m) N = 98
-(0-1m) N = 564
+(0-15d) N = 57
-(0-15d) N = 605
C
Group Antibiotics (0-1y) Control (0-1y)
cluster 1 (prevalent ARGs)
cluster 2 (rare ARGs)
cluster 3 (DA ARGs)
cluster 4 (core ARGs)
Log2 (Total abundance of ARGs)
N = 51 N = 164
N = 40 N = 113
N = 34 N = 98
N = 51 N = 166
D
Group Antibiotics (0-1m) Control (0-1m)
cluster 1 (prevalent ARGs)
cluster 2 (DA ARGs)
cluster 3 (core ARGs)
Log2 (Total abundance of ARGs)
N = 97 N = 564
N = 96 N = 558
N = 97 N = 557


A
Antibiotic treatment in the year prior to sampling
Betalactam +Sulfonamide
Betalactam +Macrolid
Tetracycline
Macrolide
Beta-lactam
Control
Log2(Total abundance of ARGs) (GCPM)
B
Log2(Total abundance of plasmids) (GCPM)
C
Antibiotic treatment in 15 days prior to sampling
Ampicillin
Penicillin
Macrolide
Control
D
E
Antibiotics
Tetracycline (N = 7)
No (N = 166)
Bifidobacterium adolescentis
Bifidobacterium longum
0.003
0.037
<0.001
0.003
Relative abundance
Macrolide (N = 15)
Bacteroides caccae
<0.001
<0.001
0.001
0.009
<0.001
0.006
Faecalibacterium prausnitzii
Haemophilus parainfluenzae
Escherichia coli
0.002
0.038
0.004
0.038
0.043
0.156
Yes (N = 57)
No (N = 351)

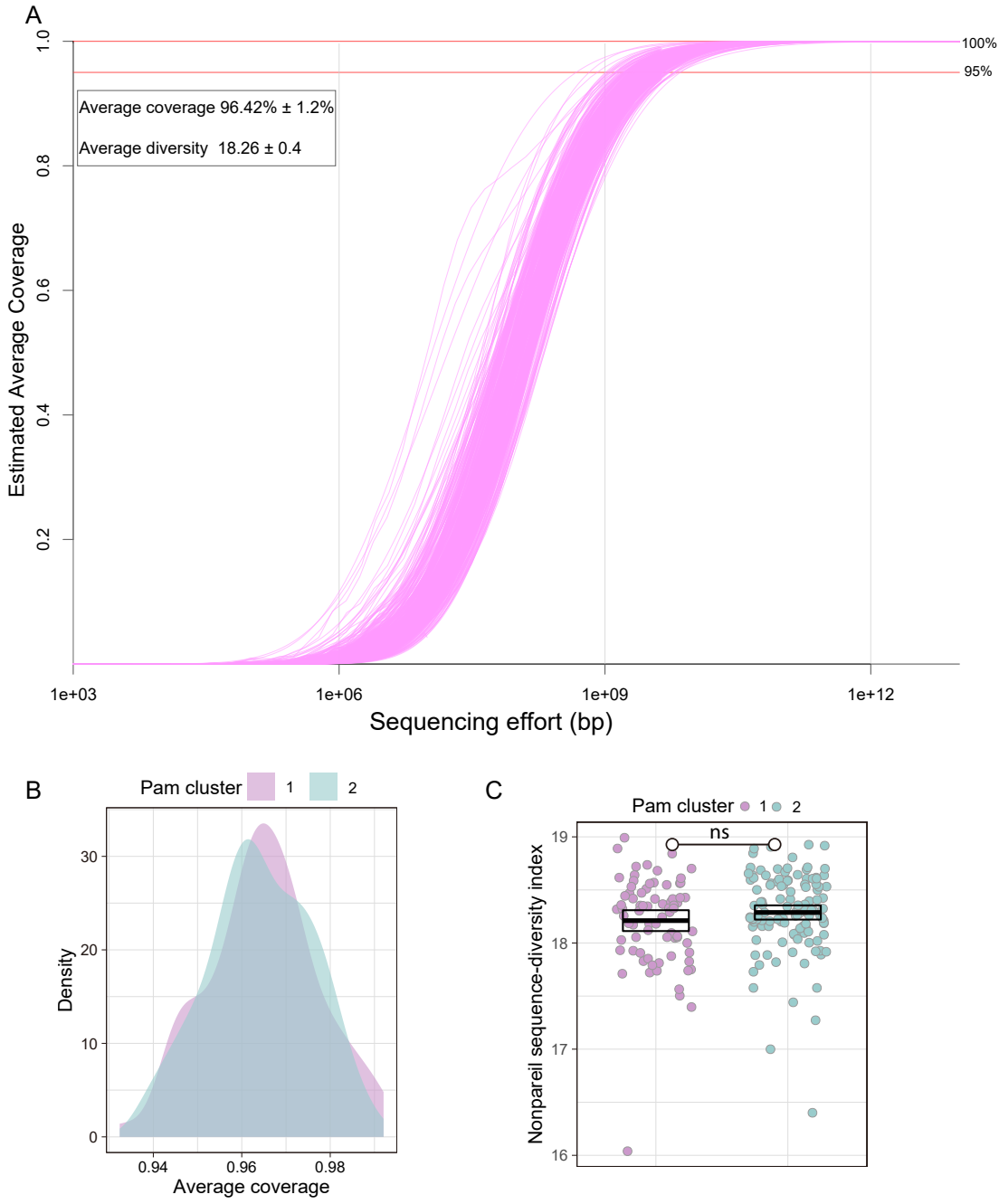


Fig. S1 Metagenomics sequencing coverage of adult samples assessed by Nonpareil using a k-mer kernel. A) Nonpareil curves showing the average coverage of all samples with sequencing effort. Average coverage and diversity (mean ± standard deviation) for all samples are shown. **B)** Density plot of average coverage in two PAM clusters. **C)** Sequence diversity in two PAM clusters. *P*-values correspond to the Wilcoxon test. ns represents no significant difference, $P > 0.05$.

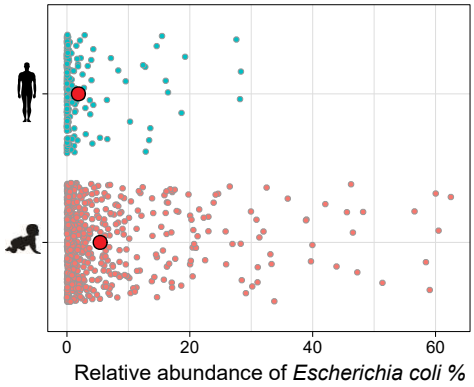


Fig. S2 The relative abundance of E. coli in adult and infant samples.

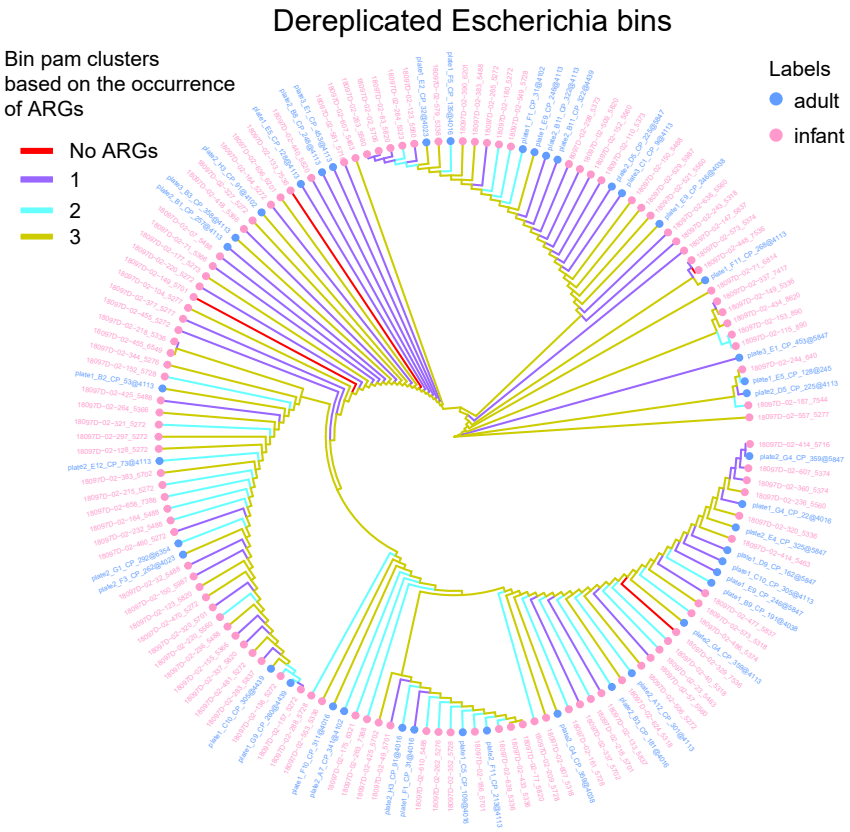


Fig. S3 Phylogenetic tree of Escherichia metagenome-assembled genomes (MAGs) in adult and infant gut based on 99% ANI analysis. *Escherichia* MAGs are classified into four categories using PAM clustering based on the presence/absence of ARGs in MAGs. The different colored branches represent these four ARG profiles.

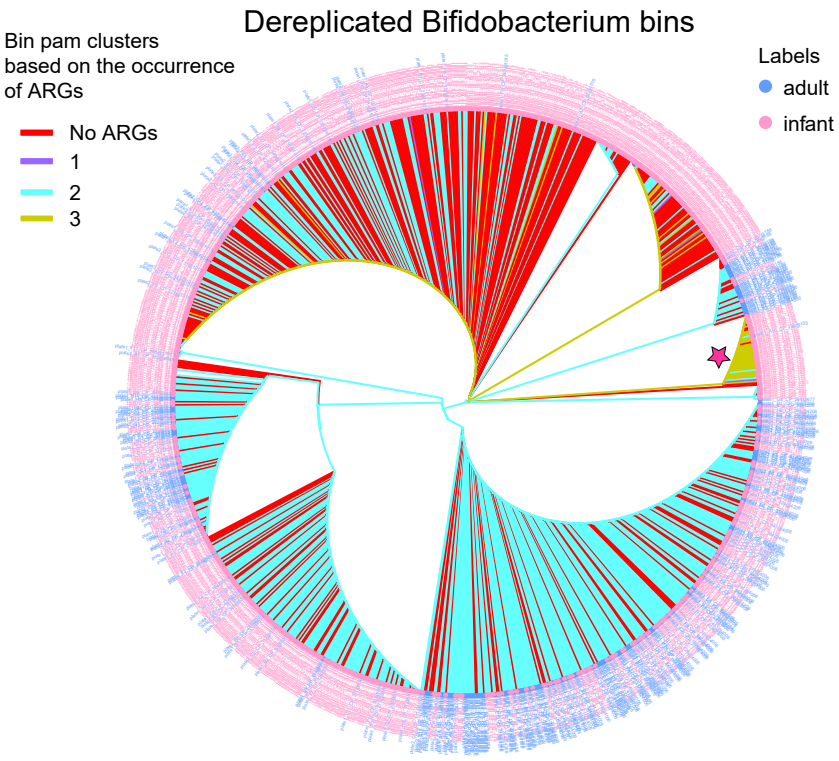


Fig. S4 Phylogenetic tree of Bifidobacterium metagenome-assembled genomes (MAGs) in adult and infant gut based on 99% ANI analysis. *Bifidobacterium* MAGs are classified into four categories using PAM clustering based on the presence/absence of ARGs in MAGs. The different colored branches represent these four ARG profiles. ARG cluster 3 in infants is heavily distributed in one MAG cluster, marked with an asterisk.

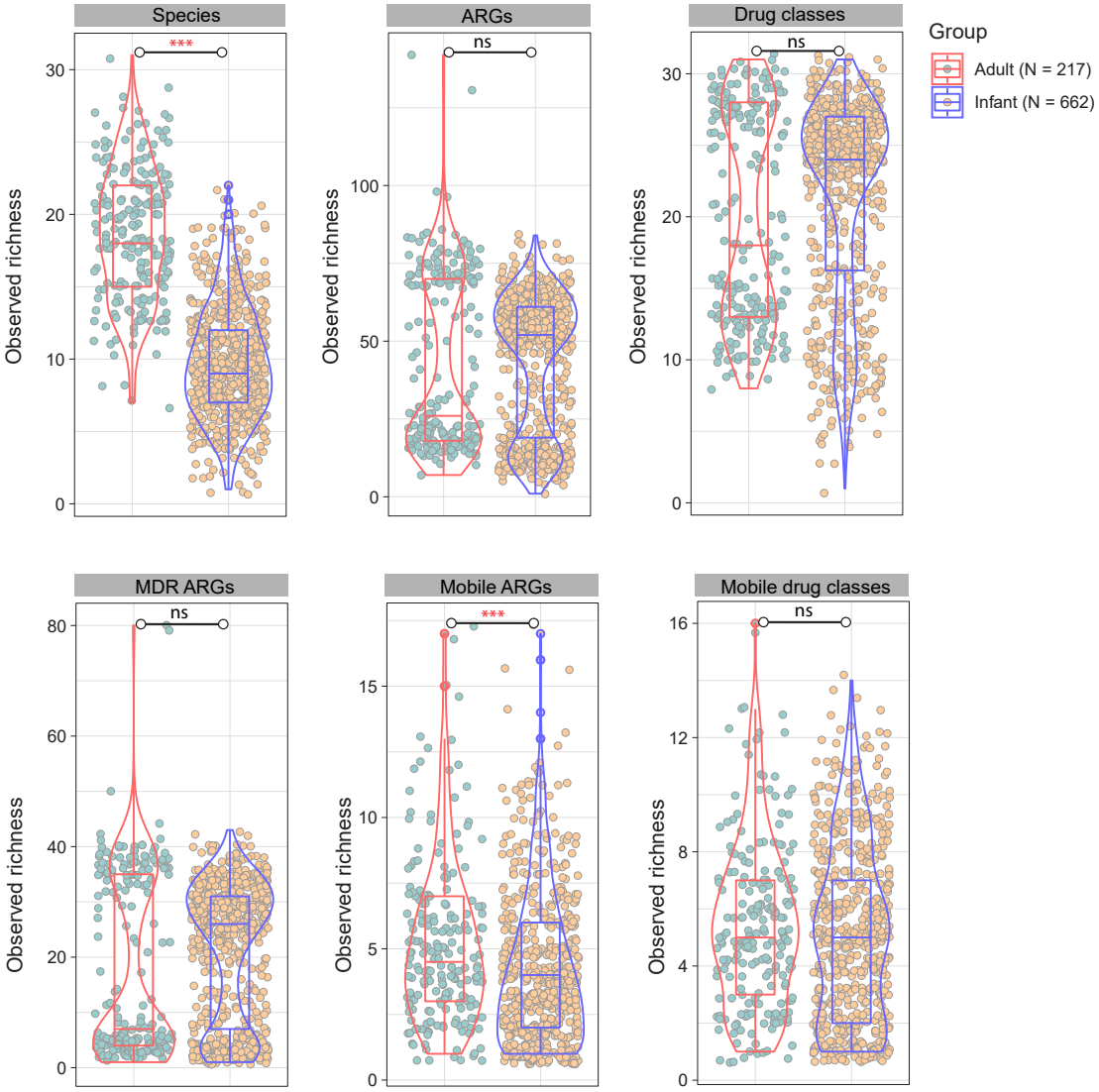


Fig. S5 The observed richness of bacterial species carrying ARGs, ARGs, drug classes, MDR ARGs, mobile ARGs, and mobile drug classes in the adult and infant gut, as a measure of alpha diversity. *P*-value < 0.001 and *P*-value > 0.05 obtained from the Wilcoxon test are indicated by three asterisks and ns.

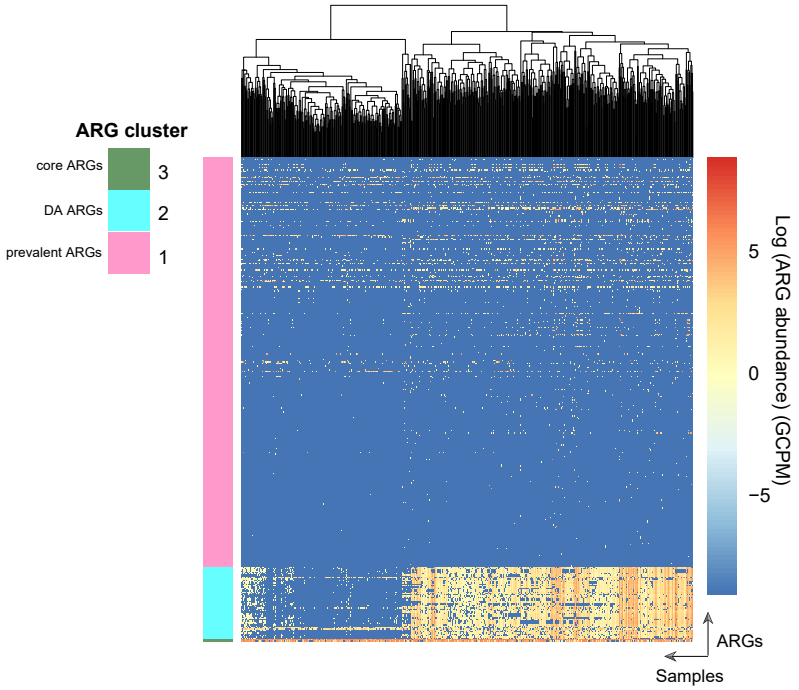


Fig. S6 Heatmap with the abundance of 366 ARGs across infant samples. Samples were clustered with Euclidean distance by complete linkage hierarchical clustering. ARGs were clustered into three categories with Euclidean distance by PAM clustering; Cluster 3 (core ARGs) contains high abundant and prevalent ARGs (N = 2) in the samples. Cluster 2 (DA ARGs) contains ARGs (N = 55) with significant abundance differences between samples. Cluster 1 (prevalent ARGs) contains ARGs (N = 311) whose abundance in the samples falls between the ARGs in cluster 3 and those in cluster 2.

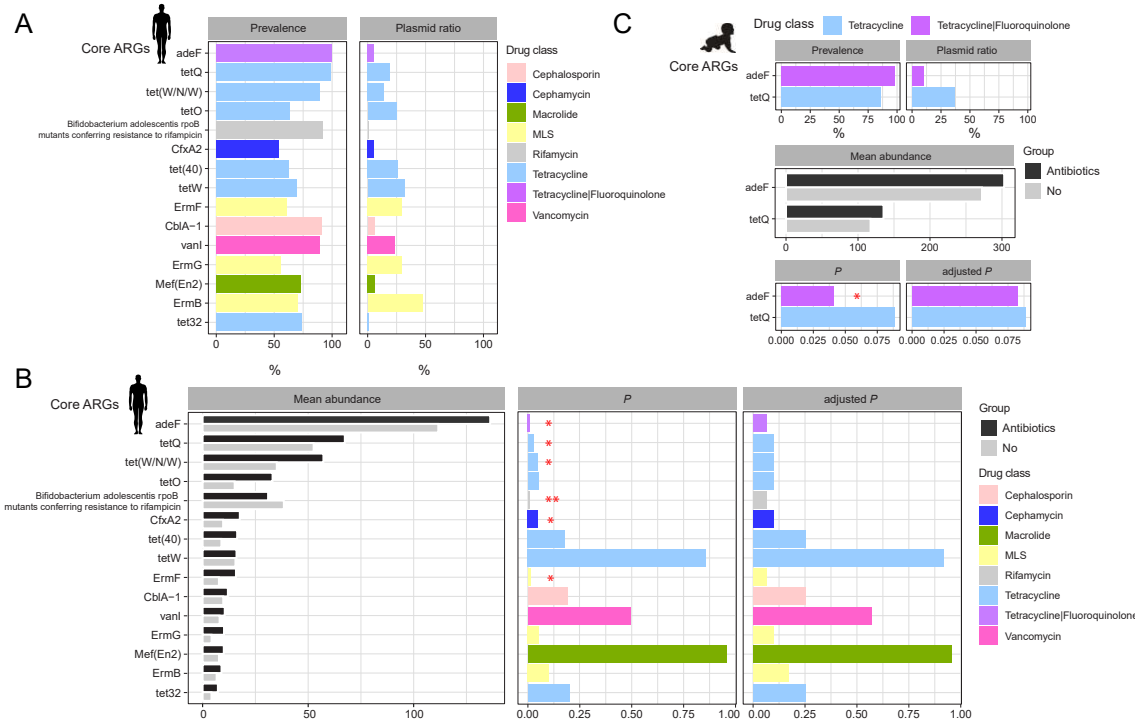


Fig. S7 An overview of core ARGs in the adult and infant gut and the impact of antibiotic treatment on core ARGs in both guts. A) Prevalence of core ARGs in the adult gut and the abundance proportion of core ARGs on plasmids. **B)** Effect of antibiotics on the mean abundance of core ARGs in the adult gut, and *P*-values and FDR-adjusted *P*-values obtained by the Wilcoxon test for comparisons. **C)** Prevalence of core ARGs and the abundance proportion of core ARGs on plasmids in the infant gut, and the effect of antibiotics on the mean abundance of core ARGs and *P*-values and FDR-adjusted *P*-values obtained by the Wilcoxon test for comparisons.

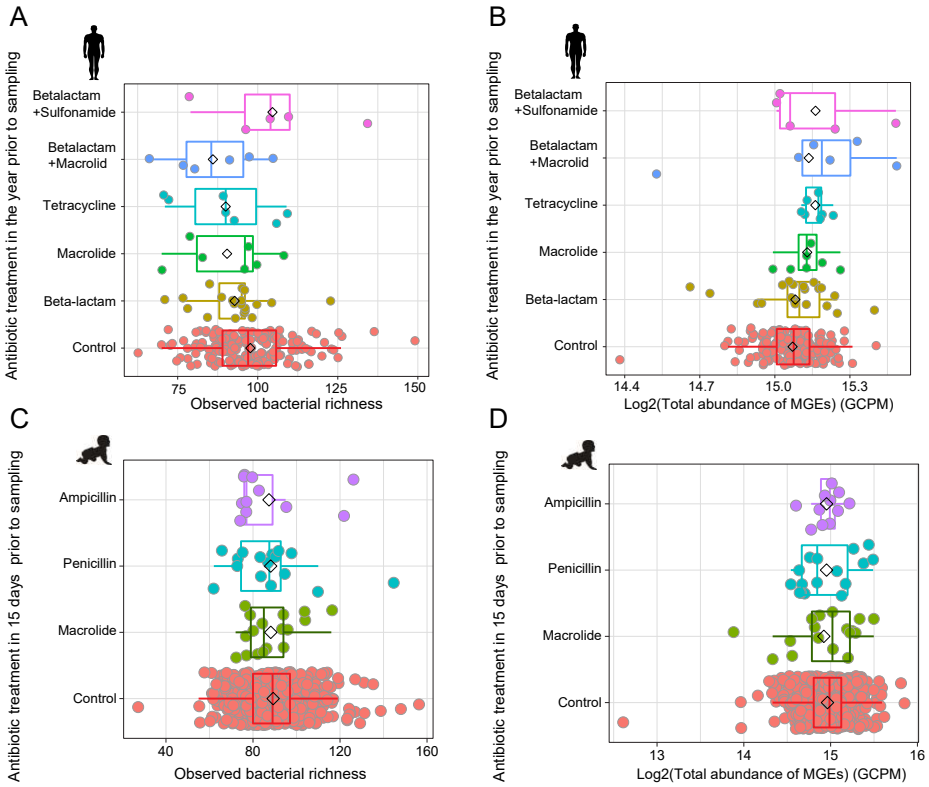


Fig. S8 The effects of various antibiotic exposures on bacterial observed richness and MGE abundance. **(A & B)** Changes in bacterial observed richness **(A)** and MGE abundance **(B)** in the gut of adults who had taken five major antibiotics and antibiotic combinations in one year before sampling. Controls are those samples that had not taken antibiotics within one year. All *P*-values obtained by the Wilcoxon test adjusted with FDR are greater than 0.05, for all pairwise comparisons. **C & D)** Changes in bacterial observed richness **(C)** and MGE abundance **(D)** in the gut of infants who had taken three major antibiotics in 15 days before sampling. To exclude interactions between antibiotics, only samples that had taken a single antibiotic were included. Controls are those samples that had not taken antibiotics within 15 days before sampling. All *P*-values obtained by the Wilcoxon test adjusted with FDR are greater than 0.05, for all pairwise comparisons. The black diamond refers to the mean value.

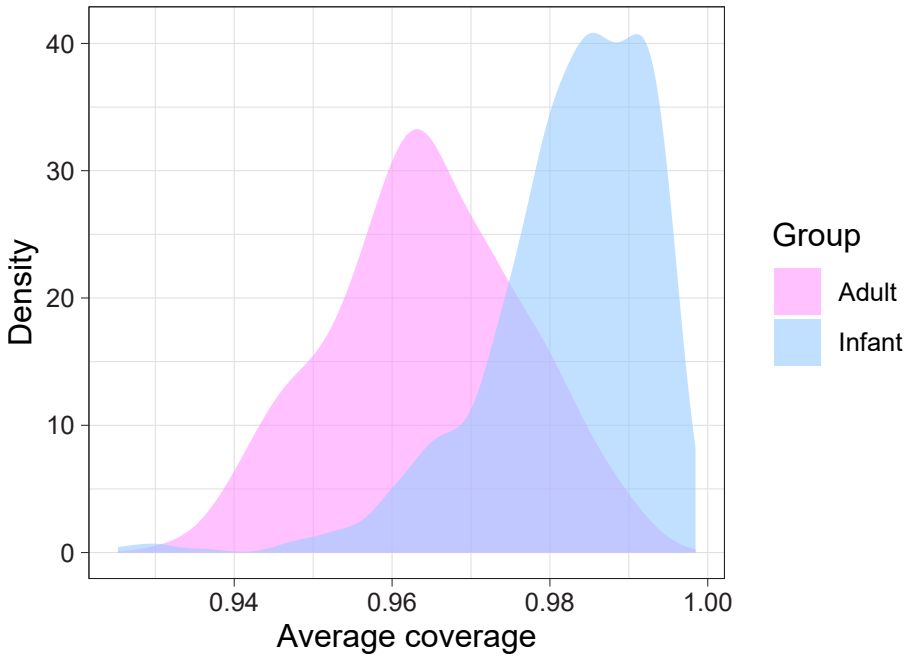


Fig. S9 Density plot of average coverage in adult and infant metagenomics samples calculated by Nonpareil using a k-mer kernel.

Technical University of Denmark
Health Technology
Section of Bioinformatics

Kemitorvet 204, 257
2800 Kgs. Lyngby

www.healthtech.dtu.dk