

# **Exploring the landscape of antimicrobial resistance in microbiomes: a computational metagenomics profiling approach**

**Achal Dhariwal**



**Ph.D. thesis**

Institute of Oral Biology  
Faculty of Dentistry  
University of Oslo  
Norway

© Achal Dhariwal, 2023

*Series of dissertations submitted to the  
Faculty of Dentistry, University of Oslo*

ISBN 978-82-8327-079-2

All rights reserved. No part of this publication may be  
reproduced or transmitted, in any form or by any means, without permission.

Cover: UiO.

Print production: Graphic center, University of Oslo.

## Table of Contents

Acknowledgements .....	4
List of Papers .....	6
List of Abbreviations .....	7
Introduction.....	10
Antibiotics and antimicrobial resistance (AMR) .....	10
Resistomes: the landscape of antimicrobial resistance genes in microbial communities .....	12
The human resistome .....	12
Impact of antibiotics on human resistome .....	14
Methods and approaches to study resistomes .....	15
Culture-dependent approach.....	16
Whole genome sequencing (WGS).....	17
Polymerase chain reaction (PCR)-based methods.....	18
Metagenomics.....	19
Functional metagenomics .....	20
Whole metagenomics.....	21
DNA sequencing and technologies .....	22
Short-read sequencing .....	24
Long-read sequencing.....	25
Bioinformatics resources for resistome profiling.....	26
Computational preprocessing .....	26
Sequence analysis .....	27
Read-based profiling.....	27
<i>De novo</i> assembly-based profiling.....	28
Read vs assembly-based approach.....	28
Antimicrobial resistance gene (ARG) databases .....	29
Downstream analysis .....	31
Data normalization.....	33
Diversity analysis.....	34
Alpha diversity.....	34

Beta diversity .....	35
Significance testing of alpha and beta diversity .....	36
Differential abundance analysis .....	38
Integrative analysis .....	40
Knowledge and Research Gap .....	43
Aims of the research .....	45
Summary of the results .....	46
Paper I .....	46
Paper II .....	47
Paper III .....	49
Methodological considerations .....	51
Paper I and II .....	51
Study designs .....	51
Sample collection and storage .....	52
DNA extraction .....	53
Library preparation and sequencing .....	54
Bioinformatics processing of sequencing reads .....	55
Sequence analysis for resistome and microbiome profiling .....	55
Statistical analysis .....	57
Ethics .....	58
Paper III .....	59
Implementation and design .....	59
Data upload and processing .....	60
Data analysis and visual exploration .....	61
Reproducibility, documentation and support .....	63
Discussion .....	65
Conclusion .....	75
References .....	77



## **Acknowledgments**

I would like to express my profound gratitude to everyone who has played a significant role in the completion of my Ph.D. thesis. Your support, guidance, and encouragement have been invaluable throughout this incredible journey.

First and foremost, I am deeply thankful to my exceptional supervisor, Fernanda Cristina Petersen. Her expertise, dedication, patience and mentorship have shaped my research and personal growth. Her constant support, constructive feedback, guidance and availability at any point of time have been indispensable. I consider myself the luckiest to have had such an amazing supervisor. I extend heartfelt appreciation to my co-supervisors, Dr. Ulf Dahle and Dr. Chen Tsute, who have been exceptional mentors, guiding my academic research and shaping my life goals and well-being. A very big thank you to Ulf for always supporting and assisting me in tough situations. I am immensely grateful for both your extensive feedback and support in shaping this thesis. Also, a big thank you to Dr. Chen for always being there to solve any bioinformatics and data analysis related challenges that I faced during my work. And also, for taking caring of me and everything during my stay in Boston.

My gratitude also extends to all the co-authors and collaborators who have contributed to this research, as well as the funding agencies that have supported my work. I extend heartfelt thanks to the Norwegian Research Council (NFR), Olav Thon Foundation, and INTPART project for their financial support and belief in the importance of this study. Your contributions have provided me with the necessary resources to pursue my academic goals.

I would like to express my appreciation to my workplace and colleagues at the Institute of Oral Biology, Faculty of Dentistry, University of Oslo, for their contributions to this thesis. I am grateful to Heidi, Sudhanshu, Kjersti, Maria, Kazi, Imen, Krystina, and all the other past colleagues for their unwavering support throughout my research journey. I would also like to give a special shoutout to Roger, Gabriela, Sushma, and Navdeep, who have not only helped me professionally but have also become like family, providing support during challenging times. Britta, thank you

for your valuable assistance in revising my thesis. To Brazil, its students, and Prof. Pedro, thank you for hosting us on two unforgettable occasions that provided me with invaluable experiences.

Finally, I want to express my deepest gratitude to my parents, in-laws, family, and friends, who have been a constant source of encouragement, support, and love. Mom and Dad, thank you for always believing in me and encouraging me to work hard each day. To my wife, Nancy, your unwavering support, understanding, and patience have been my anchor. I believe that you deserve this degree more than I do for the strength and endurance you've shown in managing everything single-handedly. I am forever grateful for all the sacrifices that you have made so that I can concentrate on my work. Special acknowledgments to my brother (Anant) and my best friends in India (Soham, Ashwini, and Sudhiksha); your love, motivation, and emotional support over the past few years have been precious. To my adorable nieces, Nia, Aadhya, and Diya, thank you for bringing daily joy and refreshing energy to my life.

To all those mentioned above and to the numerous others who have supported me in various ways, I offer my deepest gratitude. Without your unwavering support, this accomplishment would not have been possible. Thank you for being an essential part of my journey and for helping me become the person I am today.

Achal Dhariwal  
Oslo 2023

# List of Papers

The thesis includes the following papers:

- I. Dhariwal A, Haugli Bråten LC, Sturød K, Salvadori G, Bargheet A, Åmdal H, Junges R, Berild D, Zwart JA, Storheim K, Petersen FC. **Differential response to prolonged amoxicillin treatment: long-term resilience of the microbiome versus long-lasting perturbations in the gut resistome.**  
*Gut Microbes.* 2023 Dec 31;15(1):2157200
- II. Dhariwal A\*, Rajar P\*, Salvadori G, Åmdal H, Berild D, Saugstad OD, Fugelseth D, Greisen G, Dahle U, Haaland K, Petersen FC. **Unravelling the landscape of antibiotic resistance determinants in the nasopharynx and the impact of antibiotics: a longitudinal study of preterm infants.**  
*In manuscript*
- III. Dhariwal A, Junges R, Chen T, Petersen FC. **ResistoXplorer: a web-based tool for visual, statistical and exploratory data analysis of resistome data.**  
*NAR Genomics and Bioinformatics.* 2021 Mar;3(1):lqab018

---

\*Shared First authorship

Relevant papers not included in this thesis:

- IV. Rajar P, Dhariwal A, Salvadori G, Junges R, Åmdal H, Berild D, Fugelseth D, Saugstad OD, Lausten-Thomasen U, Greisen G, Haaland K, Petersen FC. **Microbial DNA extraction of high-host content and low biomass samples: Optimized protocol for nasopharynx metagenomic studies.** *Frontiers in Microbiology.* 2022;13
- V. Rajar P\*, Dhariwal A\*, Salvadori G, Åmdal H, Berild D, Dahle U, Fugelseth D, Greisen G, Lausten-Thomasen U, Saugstad OD, Petersen FC, Haaland K. **Impact of antibiotics and hospitalization on the nasopharyngeal microbiome development in preterm infants.** *In manuscript*
- VI. Sturød K, Dhariwal A, Haugli Bråten LC, Berild D, Storheim K, Zwart JA, Petersen FC. **Long-term treatment with amoxicillin – impact on the oral microbiome and resistome.** *In manuscript*
- VII. Sturød K, Dhariwal A, Dahle U, Vestrheim DF, Petersen FC. **Impact of narrow spectrum penicillin V on the oral and fecal resistome in a young child treated for otitis media.** *Journal of Global Antimicrobial Resistance.* 2020 Mar 1;20:290-7

## List of Abbreviations

ALR: Additive log transformation  
AMR: Antimicrobial resistance  
ANCOM - Analysis of composition of microbiomes  
ANOSIM: Analysis of similarities  
ANOVA: Analysis of variance  
ARGs: Antimicrobial or Antibiotic resistance genes  
AST: Antimicrobial susceptibility testing  
bp: Base pairs  
CARD: Comprehensive antibiotic resistance database  
CBMAR: Comprehensive  $\beta$ -lactamase molecular annotation resource  
CLR: Centered log ratio  
CLSI: Clinical and laboratory standards institute  
CoDa: Compositional data analysis  
CSS: Cumulative sum scaling  
DAA: Differential abundance analysis  
ddNTP: Dideoxynucleotide triphosphates  
DNA: Deoxyribonucleic acid  
ECOFF: Epidemiological cut-off value  
ESBL: Extended spectrum beta-lactamase  
EUCAST: European committee on antimicrobial susceptibility testing  
FDR - False discovery rate  
GLM: Generalized linear model  
HGT: Horizontal gene transfer  
HMM: Hidden markov model  
HMP: Human microbiome project  
IAIs: Intra-abdominal infections  
Kb: Kilobases  
LacED: Lactamase engineering database  
LDA - Linear discriminant analysis

LEfSe - Linear discriminant analysis effect size  
LME: Linear mixed-effects model  
MANOVA: Multivariate analysis of variance  
MDR: Multidrug-resistant  
MGE: Mobile genetic element  
MIC: Maximal information coefficient  
MIC: Minimum inhibitory concentration  
NB: Negative binomial  
NGS: Next-generation sequencing  
NMDS: Non-metric dimensional scaling  
ORFs: Open reading frames  
PacBio: Pacific biosciences  
PCA: Principal component analysis  
PCoA: Principal coordinate analysis  
PCR: Polymerase chain reaction  
PERMONOVA: Permutational multivariate analysis of variance  
PROTEST: Procrustean randomization test  
rCCA: Regularized canonical correlation analysis  
RLE: Relative log expression  
RNA: Ribonucleic acid  
SMRT: Single-molecule real-time  
sPLS - Sparse partial least squares  
TMM: Trimmed mean of M-values  
TSS: Total sum scaling  
u-CARE - User-friendly comprehensive antibiotic resistance repository of *Escherichia coli*  
UQ: Upper quantile  
UTIs: Urinary tract infections  
VRE: Vancomycin-resistant enterococci  
WGS: Whole genome sequencing  
WMS: Whole metagenome sequencing  
ZIG - Zero-inflated gaussian

ZILG - Zero-inflated log-gaussian

ZMWs: Zero-mode waveguides

# Introduction

## Antibiotics and antimicrobial resistance (AMR)

Antibiotics were the miracle drugs of the 20th century. They have saved millions of lives across the globe since their discovery and facilitated significant advancements in medicine and surgery <sup>1</sup>. Antibiotics treat bacterial infections by either killing the bacteria or inhibiting their growth. Most known antibiotics we use today originated from microbial sources, with a few exceptions of synthetically developed antibiotics (e.g., quinolones and sulfonamides)<sup>2,3</sup>. The majority were discovered and synthesized between 1940 and 1962, a period referred to as the golden era of antibiotics <sup>4</sup>. However, there has been a lack of novel antibiotics after this era, and only a few new classes of antibiotics have been approved for clinical use <sup>4,5</sup>. Not long after the discovery of each new antibiotic agent, bacteria developed resistance to their killing and inhibitory effects <sup>6</sup>. Unfortunately, this phenomenon arose rapidly for all other antimicrobial drugs following their introduction into the market or clinics <sup>7</sup>. One of the best examples is the first commercially used antibiotic, penicillin, for which a bacterial enzyme, penicillinase, capable of inactivating penicillin, was identified immediately after its introduction <sup>8</sup>. Today, resistance to almost all antimicrobials, including those regarded as last-resort treatments for life-threatening infections, has been documented <sup>6,9,10</sup>. Moreover, bacterial strains have evolved over time, developing from resistance to single classes of antibiotics to becoming multidrug-resistant (MDR). This leads us to the current situation, where the increasing prevalence of MDR superbugs renders common bacterial infections difficult to treat or even untreatable, thus posing a severe health crisis <sup>11,12</sup>.

Antimicrobial resistance (AMR) is the property of microorganisms to resist or thwart the effects of antibiotics, thus making drugs previously used to treat infections ineffective. AMR is one of the biggest and most urgent threats to global public health, economic development and food security <sup>13,14</sup>. According to a recent report published by Murray CJ *et al.*, it has been estimated that bacterial AMR was associated with almost 5 million deaths and was directly responsible for at least 1.27 million deaths globally in 2019 <sup>15</sup>. Several national and international public health organizations recognize AMR as an immediate danger and unanimously agree that monitoring its emergence, prevalence and spread is crucial to mitigate risks to human health <sup>16,17</sup>.

Bacterial antimicrobial resistance is primarily genetically encoded and is thought to be mediated predominantly by antibiotic resistance genes (ARGs)<sup>18</sup>. However, in some cases, bacteria are intrinsically resistant to antibiotics, i.e., resistance does not alter over time and is present in all members of particular clades. For instance, Gram-negative bacteria have outer membranes that physically constrain the action of certain antibiotics that target the cell wall<sup>19</sup>. On the other hand, genetically encoded resistance can be facilitated through several different mechanisms, including the duplication or overexpression of existing genes, point mutations, or acquisition of new ARGs through horizontal gene transfer (HGT)<sup>18,20</sup>. These acquired ARGs can be found in chromosomal DNA and as part of plasmids. Mobile genetic elements (MGEs), such as plasmids and transposons, play an essential role in the transfer of ARGs within and between bacterial species, contributing significantly to the challenge of AMR faced today<sup>21,22</sup>. In addition to the transmission of ARGs between bacteria, the resistant bacteria themselves can be disseminated throughout the environment via food and water<sup>23,24</sup>, as well as between humans and animals<sup>25,26</sup>, and among healthcare facilities (e.g., hospitals) and communities<sup>27,28</sup>. Furthermore, these bacteria can spread globally through travel<sup>29,30</sup>. All these collective factors further exacerbate the increasing AMR burden worldwide, thus making it difficult to mitigate and control it.

Irrational use of antibiotics is a major driver of AMR<sup>31,32</sup>. The misuse and overuse of antibiotic treatments have led to an increase in selection pressure, which has contributed to the rapid emergence of resistant bacteria with acquired resistance<sup>33</sup>. Antibiotics, often called a panacea for eradicating bacterial infections, have inadvertently driven the development of AMR in pathogens<sup>34</sup>. However, many human pathogens are not originally carriers of these ARGs. Instead, it is theorized that many of the ARGs found in clinical environments initially originated from natural settings<sup>35</sup>. Groundbreaking studies have discovered ARGs identical and similar to those present in pathogenic bacteria in natural environments free from human intervention, such as in glaciers<sup>36</sup> and permafrost sediments<sup>37</sup>. This suggests that AMR is an ancient and natural phenomenon. However, the situation we find ourselves in during this post-antibiotic era, with pathogenic-resistant bacteria causing treatment failures, is unnatural. Instead, it is a consequence of the prolonged imprudent use of antibiotics<sup>31</sup>. Moreover, recent studies demonstrate that environmental and commensal microbial communities harbor a vast diversity of ARGs, many of which have not yet been discovered in clinical settings<sup>38,39</sup>. AMR is ubiquitous in various



microbial communities, including humans, animals, and the environment. This widespread prevalence has prompted the development of a “One-Health” approach to address the global challenge of AMR <sup>40</sup>.

## **Resistomes: the landscape of antimicrobial resistance genes in microbial communities**

The term “resistome” refers to the collection of all types of ARGs in a microbial community, including intrinsic and acquired genes, their precursors, and potential resistance mechanisms. Among these mechanisms are silent ARGs, which require alterations in expression to confer resistance, and proto ARGs, which exhibit little or no activity until they mutate <sup>41</sup>. The abundance and diversity of resistomes are largely influenced by the microbial composition, which can host different types of ARGs, and the type of exposure the microbial community has had to anthropogenic factors, particularly antimicrobials <sup>38,42</sup>. Due to the increased availability of next-generation sequencing (NGS) techniques and significant advancements in culture-independent methods (such as metagenomics), the complex and diverse commensal and pathogenic microbes living in and on humans (referred to as the human microbiota) have gained increased attention as an important reservoir of ARGs <sup>43</sup>. The genomes of all the microbes within human microbial communities are collectively known as the human microbiome. Understanding the compositional characteristics and dynamics of the human resistome, including the factors that affect the diversity, abundance, and transmission of ARGs, is crucial for developing preventive and therapeutic strategies to prevent the development and spread of AMR <sup>1,44</sup>.

### **The human resistome**

NGS technology-based metagenomics has been utilized to study the resistomes of human microbiomes found in the gut, oral cavity, skin, and respiratory tract <sup>45-48</sup>. The different niches in the human body have varying resistome profiles with differing ARG abundances <sup>49</sup>. One functional metagenomics-based study characterizing the microbiomes and resistomes of uncontacted Amerindians in different body sites found that ARGs from all resistance classes, other than ribosomal protection proteins, were exclusively present in either oral or gut metagenomes <sup>50</sup>. There has been a recent flurry of metagenomic studies focusing on the human gut niches (stomach

and intestines). In fact, one of the first whole metagenomic studies to characterize the resistome of healthy adults worldwide was performed on fecal samples <sup>47</sup>. A total of 507 unique ARG types, conferring resistance to 20 different classes of antibiotics, were identified in 180 fecal samples collected from 11 countries. Additionally, the study also found that differences in gut microbial composition between adults are co-localized with differences in the composition of their resistomes. This finding underscores the importance of the human gut microbiome as an AMR reservoir. The well-recognized role of the human gut microbiome in health and disease, along with its abundance and accessibility, may have contributed to a greater emphasis on whole metagenomic resistome investigations in the gut compared to other body sites. As a result, the majority of recent metagenomic studies, which provide insights into the dynamics of the human resistome in early and adult life, as well as during exposure to various types of antimicrobial drugs (mostly antibiotics), are primarily focused on the gut <sup>51-53</sup>. The resistome in adults generally remains relatively stable without any significant perturbations <sup>46,54</sup>.

The period of microbiome maturation is considered the most crucial for the acquisition and establishment of the resistome. Yet, little is known about the trajectory and dynamics of resistome development in infants <sup>42</sup>. A few reports seem to indicate, however, that the diversity of ARGs in the gut can be significant already within the first months of life <sup>52,55,56</sup>, and that for some groups, such as preterm infants, the hospital environment seems to function as an important source of resistance, particularly for those exposed to antibiotics <sup>57</sup>. Such effects can be long-lasting, resulting in microbiome scars observed months after hospital discharge <sup>53</sup>. The elevated early-life resistome is associated with the taxonomic composition in the gut microbiome <sup>51,52,55,56</sup>. Additionally, the infant gut resistome can be seeded from their mother, as bacteria containing ARGs can be transferred from maternal milk to the infant gut through breastfeeding <sup>58</sup>. As a result, the infant gut resistome shares ARGs with the maternal milk and gut resistomes early in life <sup>59</sup>.

In other body sites, the respiratory tract is becoming an area of increasing interest. However, resistome studies in such sites are still scarce and mostly use single sampling time points, which limits our understanding of the dynamics of the respiratory resistome <sup>60</sup>. One of the main challenges in conducting metagenomic studies lies in the low microbial biomass and high human

DNA content in the samples <sup>61,62</sup>. Recent studies focused on optimizing methods to address these issues will likely contribute to advancements in the field <sup>63</sup>.

## **Impact of antibiotics on human resistome**

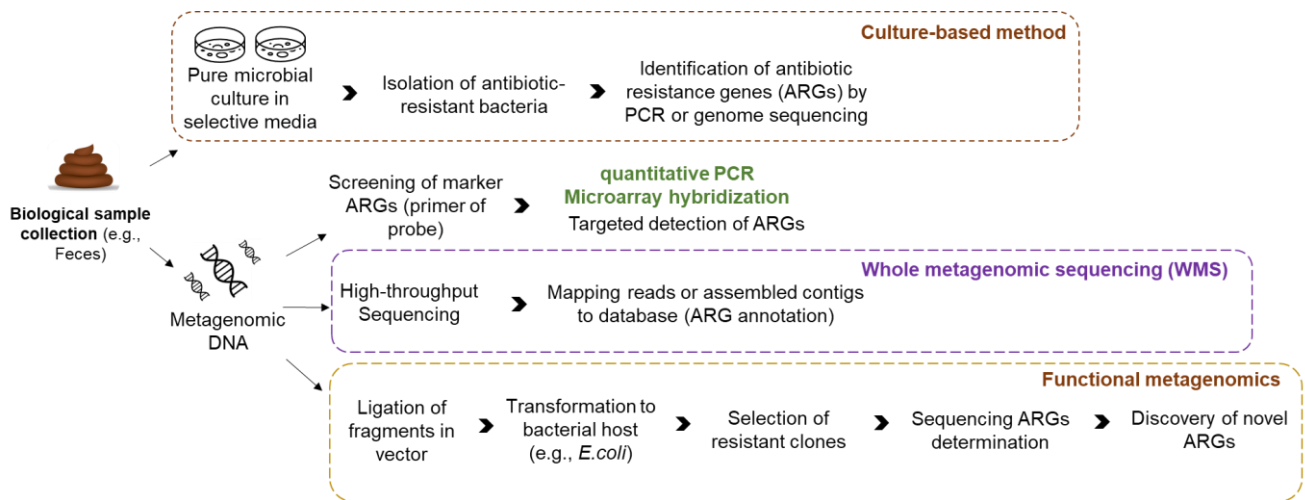
Selection pressures, such as antibiotic exposure, enhance the selection of antibiotic-resistant bacteria and increase the gene pool of resistance mechanisms, ultimately exacerbating the risk of spreading ARGs in microbial communities <sup>64</sup>. Antibiotic treatment promotes resistance in target pathogens and off-target microbes, including bystander opportunistic pathogens, which may become a source of future infections <sup>65</sup>. Studies have shown that antibiotics can significantly impact the abundance and diversity of ARGs in the human gut microbiome <sup>66-68</sup>. However, reports of lasting changes in ARGs or the resistome vary per study depending on the antibiotic spectrum, class, mode of action, duration, route, and dosage <sup>44</sup>. In addition, several host-intrinsic and other external factors such as genetics, baseline microbiome composition, age, co-morbidities, diet, and environment can also affect the impact of antibiotic treatment on the human microbiome and resistome <sup>42</sup>. For example, early antibiotic administration in hospitalized neonates with developing microbiota had a profound impact on the gut resistome diversity and composition <sup>69,70</sup>. However, the effect persisted long-term in some cases, depending on the type of antibiotics and treatment duration. During childhood, antibiotic exposures usually consist of short courses of relatively narrow-spectrum agents for respiratory tract and oropharyngeal infections <sup>71</sup>. In the previous study, a significant increase in ARG diversity in the gut of a young child was observed one month after the administration of narrow-spectrum penicillin V <sup>72</sup>. However, no such increase was observed in the oral cavity. In adults, the impact of higher acute and lower chronic antibiotic regimens can exert diverse effects on the microbiome and resistome composition, both in the short- and long term. For example, a cocktail of three last-resort, broad-spectrum antibiotics (i.e., meropenem, gentamicin, and vancomycin) was orally administered to healthy adult subjects for four days, which resulted in nine common species becoming undetectable in the gut microbiome after 180 days (6 months) <sup>73</sup>. Bacterial species harboring  $\beta$ -lactam, aminoglycoside, and glycopeptide ARGs colonized the gut for long-term after treatment completion. Although this study indicates minor yet long-term effects on healthy adults gut microbiome and resistome, it is essential to note that the research used a short course of multiple broad-spectrum antibiotics, which are not commonly administered in clinical settings. In comparison to the gut, a one-week course of the most

commonly used broad-spectrum antibiotics (clindamycin, ciprofloxacin, minocycline, or amoxicillin) in 66 healthy adults showed a remarkable recovery of the microbiome and resistome in the oral cavity after 12 months <sup>74</sup>. However, one of the main reasons the oral microbiome and resistome may be more resilient to antibiotics than the gut is the greater bioavailability and extended exposure time of orally administered antimicrobial drugs in the gut compared to the oral cavity.

In addition to inducing resistome enrichment, antibiotics have been reported to increase the frequency of HGT. Previous studies by Prudhomme *et al.* showed that antibiotics such as aminoglycosides and fluoroquinolones could induce genetic transformation of competent *Streptococcus pneumoniae* in response to general antibiotic stress <sup>75</sup>. This is also the case for other streptococci <sup>76</sup>. Increased conjugation frequency promoting ARG transmission is also reported to be associated with antibiotic exposure. For instance, exposure to several beta-lactam antibiotics increased the conjugation of plasmids containing extended-spectrum beta-lactamase ARGs in *E. coli* <sup>77</sup>.

## **Methods and approaches to study resistomes**

Current methods of surveillance for AMR in clinical settings rely on culture-based methods of determining resistance phenotype from a colony of an isolated bacterial strain, followed by sequencing of its genome to identify genetic AMR determinants. Molecular methods, such as polymerase chain reaction (PCR)-based approaches and metagenomics, can be employed to investigate genetic AMR determinants (ARGs) in microbial communities without the need for culture-based methods (Figure. 1) <sup>40,78</sup>. The relative merits and contributions of these methods for assaying ARGs (or AMR) are discussed in detail in the following section.



**Figure 1:** Overview of methodologies used for characterizing antibiotic resistance genes (ARGs) in resistome studies. The figure is based on Van Schaik, W., 2015 <sup>78</sup>.

### Culture-dependent approach

The culture-based approach is a traditional method where microorganisms are isolated from patient or environmental samples and cultured in agar or liquid broth to probe AMR phenotypes <sup>18</sup>. These culture-based tests describe the phenotype of a microorganism with AMR by exposing it to certain antimicrobial concentrations to determine how well it grows in the presence of antimicrobials. Two measurements are widely used to define this: the minimum inhibitory concentration (MIC) and the epidemiological cut-off values (ECOFFs) for resistance. The MIC is the lowest antibiotic concentration that inhibits the visible growth of a microbial strain. Culturing of the isolate is performed on agar plates or in liquid broth containing a gradient of antimicrobial concentrations, and the minimum concentration that inhibits the growth of the inoculum is reported as the MIC <sup>79</sup>. Alternatively, a disk <sup>80</sup> or paper strip <sup>81</sup> diffusion test can also be used to determine the susceptibility of a microbe to antimicrobials by calculating the distance of microbial clearance from the source of the antimicrobial. The area around the antimicrobial disc or strip with no microbial growth is known as the zone of inhibition. The concentration of the antimicrobial decreases as it diffuses away from the disk. Thus, no visible rings around the disk indicate resistance while a broader inhibition zone indicate susceptibility to low concentrations of the antimicrobial. Standards published by the Clinical and Laboratory Standards Institute (CLSI) or

the European Committee on Antimicrobial Susceptibility Testing (EUCAST) are referenced to convert MICs or zones of inhibition measurements to categorical resistance interpretations.

Antimicrobial susceptibility testing (AST) is commonly used in clinical settings because it provides actionable phenotypic information to guide the selection of optimal treatment for individual patients<sup>18</sup>. It has provided valuable insights into geographical variations and trends in antimicrobial susceptibility. For example, one study utilized the broth dilution method to demonstrate that the antimicrobial susceptibility profiles of *Acinetobacter baumannii* isolates causing intra-abdominal infections (IAIs) and urinary tract infections (UTIs) varied across six global regions but were low everywhere with no antibiotics inhibiting more than 70% of the isolates in any region<sup>82</sup>. While these phenotype-based resistance determination methods can offer important information for clinical management and surveillance, they provide little or no direct information on resistance gene epidemiology, including resistance mechanisms, transmission routes, or pathogen evolution<sup>83</sup>.

### **Whole genome sequencing (WGS)**

Phenotypic testing alone does not offer information on how the AMR is mediated at a genotypic level. WGS is a molecular-based AMR surveillance approach that can be used to determine the entire or nearly entire DNA sequence (i.e., genome) of a microorganism. By comparing DNA sequence of a microbe with a database containing reference sequences of ARGs and mutations in well-characterized microbial genomes, it is possible to infer their phenotypic traits such as AMR and virulence. Similarly, one can also differentiate between phenotypically identical isolates (same AST profile), identify putatively novel resistant genes and identify whether they were acquired through HGT. It also provides information on the location of AMR determinants on the bacterial chromosome or on plasmids, which can help in monitoring their spread and dissemination<sup>84</sup>. This method provides a higher resolution compared to other molecular-based approaches by enabling not only the determination of possible origin of the host bacteria but also the investigation of genetics of the loci responsible for resistance. It has been used successfully in AMR surveillance for pathogens such as in multidrug resistant tuberculosis<sup>85,86</sup>. Comparison of the whole microbial genomes of different isolates can facilitate reconstruction of putative transmission chains for both antimicrobial-resistant clones and mobile genetic elements of AMR as well as the evolutionary

tracing of newly characterized AMR microorganisms and disease outbreaks<sup>84</sup>. For example, WGS was applied to trace the evolutionary origin and genetic characteristics of a meropenem-resistant *Streptococcus pneumoniae* serotype 5A-ST63 after the introduction of pneumococcal conjugate vaccines in Japan<sup>87</sup>. Another WGS-based study of multiple extended-spectrum  $\beta$ -lactamase (ESBL)-producing *Klebsiella pneumoniae* strains in a specialized geriatric care ward identified that the referring hospital was the source of acquired resistance and where strain-to-strain dissemination of a blaCTX-M-15 FIBK/FIHK plasmid happened<sup>88</sup>.

Despite its application in public health AMR surveillance, WGS may lead to false positive or negative results if not complemented with phenotypic testing. This method can be used only to detect and interpret known/novel ARGs or mutations that are similar to previously known ones. Failure to identify the presence of unknown ARGs may result in incorrect predictions regarding the absence of AMR. Conversely, a result indicating a positive molecular result and a resistant phenotype denotes the expression of the ARG. However, genes (silent) or pseudo genes may be present but not expressed, which can wrongly predict the AMR if only WGS is performed. Although, WGS is a powerful approach for ARG surveillance, the need for pure culture is a major drawback, as a large proportion of microbial communities that harbor a wide variety of genetic resistance determinants (ARGs) in complex and diverse microbial communities are yet uncultured in the laboratory condition<sup>89,90</sup>.

### **Polymerase chain reaction (PCR)-based methods**

PCR is another molecular-based assay that amplifies a target nucleic acid sequence (such as an ARG) present in a sample. This is achieved using a pair of oligonucleotide primers that are complementary to each end of the target sequence. Initially, researchers targeted a small select set of ARGs, but with the ongoing reduction in the cost of NGS technologies and the consequent expansion in bacterial genome sequencing, the availability of ARG targets in databases has significantly increased. Similarly, more recent advancements in the field of PCR-based techniques have further enhanced their application in clinical diagnostics<sup>91,92</sup>. This enables the simultaneous detection and quantification of a large number of targeted ARGs (resistome) in a faster, more convenient, and high-throughput manner across various environments<sup>93-96</sup>. PCR has been extensively used to monitor the spread of Vancomycin-resistant Enterococci (VRE) and Extended

Spectrum Beta-Lactamases (ESBLs) producing Enterobacteriaceae <sup>97-99</sup>. In addition, it has been shown to reveal HGT mechanisms of ARG transfer. For instance, Class 1 integrons and *Salmonella* Genomic Island 1 (SGI1) were identified using PCR in *Salmonella* isolates that were previously non-carriers <sup>100</sup>. However, while PCR is highly sensitive, it does not provide information on bacterial hosts or the complete genetic context of the ARGs. Furthermore, this technique can only detect previously identified ARGs that are fully complementary to the primers and probes, often overlooking homologous sequences that could also contribute to similar resistance phenotypes. To overcome these limitations, a less targeted methodology is needed to detect all known and unknown ARGs, i.e., the resistome. Metagenomics sequencing offers a culture-independent solution to profile the complete resistome.

## **Metagenomics**

Metagenomics is the study of all the DNA or genomes from microorganisms present in a community or sample, collectively referred to as the metagenome. Over the past decade, metagenomic sequencing has become increasingly prevalent in assessing the composition, diversity, and structure of microbial communities in humans, animals, food, and the environment. Since a large proportion of microbes inhabiting these communities are uncultivable under standard laboratory conditions, the development of metagenomics has provided an unprecedented opportunity to explore and understand the role of these communities under unbiased conditions. With the increasing knowledge and understanding of the commensal and environmental bacterial present in these communities as key reservoirs of ARGs and their contribution to the emergence and dissemination of AMR <sup>101-103</sup>, it is crucial to focus on them to elucidate the actual abundance and diversity of ARGs, i.e., the resistome.

Metagenomic approaches for ARG profiling can be divided into two main types: functional metagenomics and whole metagenomics. These methods have been commonly used to detect novel ARGs (functional metagenomics) and to investigate ARGs at the microbiome level (whole metagenomics). However, the initial step for both these methods remains the same: to extract the metagenomic DNA from the sample of interest.

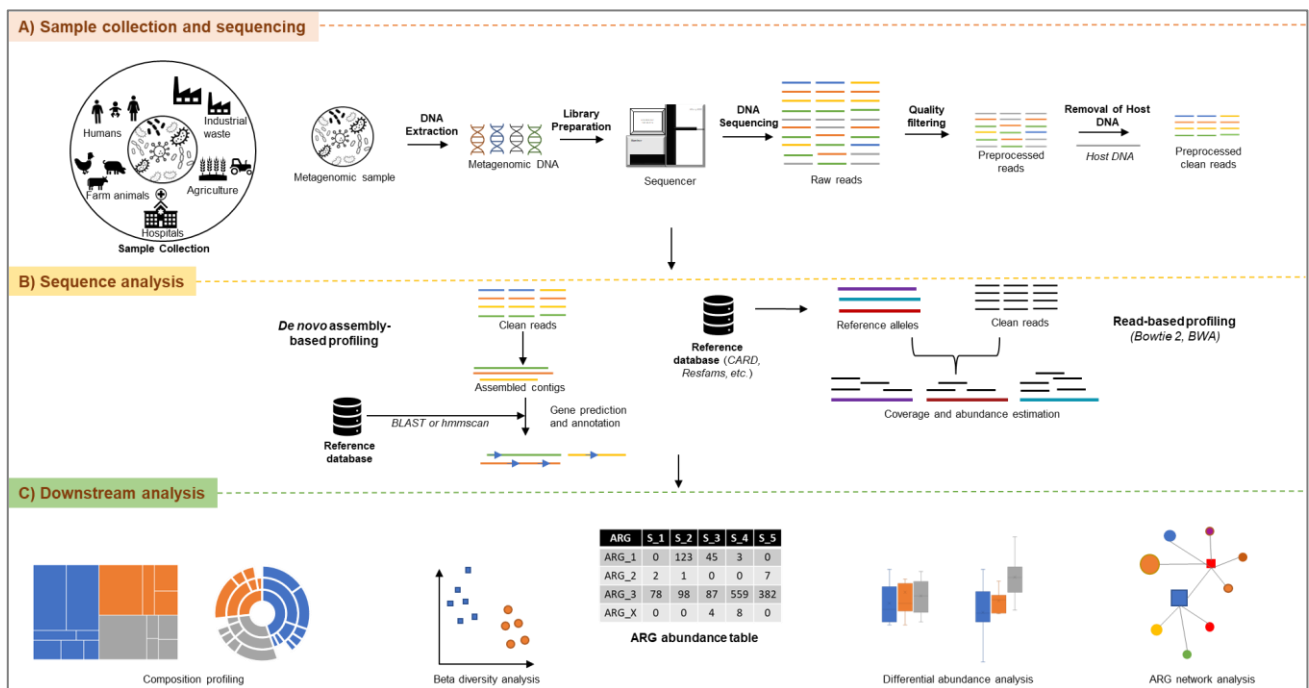


## Functional metagenomics

Functional metagenomics is a sequence-unbiased, culture-independent method for resistome characterization<sup>18</sup>. It aims to characterize the function of genes, including the mechanisms of novel ARGs, based on phenotypic expression rather than relying on bioinformatics predictions<sup>104</sup>. In this method, total microbial DNA is isolated from the sample of interest, fragmented, and subsequently cloned into a suitable expression vector (e.g., a plasmid). This vector, with the inserted DNA, is introduced into a culturable host, such as *E. coli*, to create a metagenomic library. This library, containing transformed clones, is then screened for antibiotic resistance by culturing with antibiotics at a concentration toxic to the wild-type host. The inserts from surviving recombinant, antimicrobial-resistant, or phenotypically positive clones are then sequenced to determine potential genetic determinants of that response. This method circumvents limitations imposed by bacteria that are difficult to culture and explicitly selects for ARGs that can disseminate through HGT, as only transferable phenotypes to a heterologous host are found. An earlier study revealed that a significant proportion (79%) of newly discovered ARGs in their functional metagenomic libraries had low identity to known ARGs and were not present in ARG databases<sup>51</sup>. The application of functional metagenomics is further highlighted by the discovery of novel ARGs, including tetracycline-inactivating enzymes<sup>105</sup> and rifamycin phosphorylases<sup>106</sup>. The first report of a functional metagenomic investigation of the human gut resistome screened metagenomic DNA from fecal and saliva samples from two unrelated healthy humans against 13 antibiotics<sup>39</sup>. Sequencing of 95 unique inserts containing functional ARGs, encoding resistance to all antibiotics screened, showed they were distantly similar to the closest previously known ARGs. They also discovered 10 novel  $\beta$ -lactamases in the two gut microbiomes. However, there are limitations to this method. For example, ARGs expressed in certain organisms (such as Gram-positives) may not exhibit a discernible phenotype, be efficiently expressed in heterologous hosts, or under in vitro conditions. It has also been observed that some genes confer resistance when overexpressed in a surrogate host but may not function as bona fide resistance genes in their original host<sup>107</sup>. Additionally, depending on insert size and the location of genetic resistance determinants, an ARG may be truncated, or multiple ARGs encoding multiple regulatory elements and promoters might not be captured in a vector<sup>107,108</sup>.

## Whole metagenomics

In whole metagenomics, also commonly referred to as shotgun metagenomics, the total DNA extracted from a sample is directly sequenced using a high-throughput DNA sequencing platform, effectively bypassing the culturing step (Figure 2). Consequently, this method theoretically enables the identification of all viable or non-viable microbes that can be cultured, as well as those that have not yet been cultured<sup>109</sup>. After the total DNA is extracted from the sample and the microbial community present, it is randomly sheared into small fragments. These fragments are then ligated with adapters that are identifiable by the sequencing platform prior to the sequencing process. The resulting sequence reads can then be directly or indirectly compared with public databases containing all known reference ARG sequences to profile the resistome<sup>18,110</sup>. This approach can collectively profile multiple domains, including bacterial, archaeal, eukaryotic, and viral sequences. In addition, it can also be used to predict the functional potential (or any gene profile such as mobilome) and obtain the whole genome sequences<sup>109</sup>.



**Figure 2:** Summary of a suggested metagenomics workflow for studying resistome in complex and diverse microbial environments using metagenomic sequencing data. The figure is based on M Boolchandani *et al.*, 2019, *Nature Reviews Genetics*<sup>18</sup>.

However, the many advantages of whole metagenomics are complemented by some challenges. This approach relies on existing resistance gene annotations (ARG databases), which limits its potential to identify fundamentally novel ARGs. Furthermore, the accuracy of characterized resistome profiles greatly depends on the sequencing depth, as this impacts the sensitivity to detect low abundant or rare ARGs of interest. Consequently, many studies encounter a trade-off between sequencing depth and sensitivity (undersampling), where rare sequences may remain undetected or lack coverage<sup>111</sup>. This issue becomes particularly prominent when characterizing microbial and resistome composition from metagenomes derived from complex microbial communities, such as natural environments that contain DNA from a large number of species. Additionally, the data from the sequencing process can often be highly fragmented, making it challenging to reconstruct the complete genomes of all the microorganisms found in the sampled community by assembling short reads into longer contigs<sup>18,20,60</sup>. Lastly, and most importantly, the analysis and interpretation of large and complex metagenomic data sets is challenging, requiring significant computational resources and bioinformatics expertise<sup>112</sup>.

Nonetheless, this method is being extensively used to characterize the resistome in complex microbial communities and environments, offering unprecedented knowledge about the large reservoir of ARGs and its global distribution within human<sup>113-116</sup>, animal<sup>25,117</sup>, and environmental<sup>118-120</sup> microbial communities. Furthermore, an increasing number of whole metagenomics-based resistome studies have already provided unparalleled insights into the development of the resistome in early life<sup>51,55,69,121,122</sup>, and how it can be affected by the use of antibiotics in humans<sup>70,73,74,123</sup>. This method has been implemented in the surveillance of AMR across various environments<sup>113,117,124-127</sup>. The sequencing-based metagenomics approach is becoming increasingly popular, feasible, and accessible, thanks to continuous improvements in NGS technologies, decreasing sequencing costs, and the development of bioinformatics tools and pipelines for the rapid identification and characterization of ARGs in complex and diverse metagenomes<sup>18</sup>.

## **DNA sequencing and technologies**

The process of determining the arrangement or order of nucleotides in DNA or RNA is known as sequencing. Over the past two decades, genome sequencing has played a vital role in interpreting

information encoded in bacterial genomes. The first major advancements in sequencing happened with the introduction of the first and foremost successful commercially feasible DNA sequencing method (i.e., Sanger sequencing) in 1977. Using this method, Craig Venter and colleagues performed the first whole genome sequencing of a non-pathogenic bacterial strain, i.e., *Haemophilus influenzae*, in 1995<sup>128</sup>. Sanger sequencing is based on chain termination method where the function of DNA polymerase is utilized by adding specific chain-termination dideoxynucleotide triphosphates (ddNTPs) to disrupt DNA synthesis reaction in vitro<sup>129</sup>. The ddNTPs are labelled with nucleotide-specific (ddATP, ddGTP, ddCTP, and ddTTP) radioisotopes. Alternatively, they may be labeled with fluorescent markers. When one of the ddNTPs forms a phosphodiester bond with a nucleotide from the original sequence, it inhibits the DNA polymerase from elongating the DNA. The products of such reactions with nucleotide-specific radioisotopes can then be run via polyacrylamide gel electrophoresis, and the order of nucleotide sequence can be determined according to the positions of the electrophoretic bands. If using fluorescently labeled ddNTPs, a laser excites the dye-labeled DNA fragments as they pass through a small window based on capillary electrophoresis. The excited dye produces a light at a specific wavelength that is detected by a sensor. Then, software can infer the detected signal and translate it into a call to identify the chain-terminating nucleotide. Sanger sequencing is also referred to as "first-generation" sequencing. This method has very high accuracy but is relatively expensive, time and labor-consuming, and is low throughput, thereby limiting its applications for large-scale sequencing projects.

Majorly driven by attempts to sequence the human genome as a part of the Human Genome Project, a much larger breakthrough was the development of high-throughput "second-generation", also referred to as "next-generation" sequencing (NGS) techniques<sup>130</sup>. NGS machines increased the yield of sequencing efforts significantly and enabled sequencing for large-scale, automated sequencing projects way faster and cheaper compared to Sanger sequencing. However, Sanger sequencing is still used for precise sequencing of smaller DNA molecules. The sequence reads produced with NGS techniques are typically not very long and hence, it is also known as short-read sequencing. This technique is based on parallel sequencing of spatially separated fragments (or reads) of DNA molecules. NGS technologies have been further developed with the objective of sequencing longer reads while preserving the low cost per read achieved with the second-generation techniques<sup>131</sup>.

This led to the development of third-generation sequencing technologies. High-throughput sequencing technologies have made it simpler and faster to sequence more DNA, implying metagenomes could be sequenced by short-read and long-read sequencing technologies. However, the quality of the metagenomic data is dependent on the type of technologies used. In the next sections, short-read and long-read sequencing technologies are assessed for their advantages and disadvantages in sequencing metagenomes for their resistomes.

### **Short-read sequencing**

Short-read Illumina sequencing is the current gold-standard sequencing technology, which is generally used for its high throughput, accuracy, as well as its affordability. For instance, the latest NovaSeq 6000 can practically generate up to 6000 gigabase (Gb) in a single run with a cost of less than \$0.01 per million bases and at a 0.1% error rate in base calling (process of ascribing nucleotide bases from signals). The adapter ligated DNA fragments are amplified into multiple identical copies (clusters generation) of the same fragments on the surface of a glass flow cell. After clonal amplification, the sequencing is carried out by synthesis. The four modified nucleotides (A, T, C, G) with a reversible fluorescent blocker are then washed onto the surface of a flow cell. The DNA polymerase can add only one nucleotide that complements a nucleotide of the fragment at a time due to having a reversible fluorescent blocker. The corresponding fluorescent dye of the added nucleotide is recorded and determined afterward by fluorescent imaging. After recording is done, the fluorescent blockers are removed from the newly synthesized nucleotides to enable the next round of synthesis. During sequencing, the recorded images are processed into much smaller files of nucleotide characters, which can be further interpreted and analyzed using bioinformatics tools and software. At present, Illumina dominates the NGS market by offering several high throughput sequencing platforms or models such as HiSeq, MiSeq and NextSeq, each of them varying in terms of their speed, cost, output range and other parameters <sup>132</sup>.

Illumina sequencing can be achieved in single-read or paired-end mode. Single-read sequencing means sequencing DNA fragments from only one end, while paired-end sequencing enables sequencing of both ends of the DNA fragments. Ion Torrent is another short-read sequencing technology alternative to Illumina, which has been used occasionally for its cost-effectiveness, though, it compromises on precision. In this technology, the fragment amplification is done using

emulsion PCR on micro-sized diameter beads instead of a solid surface. DNA polymerase is added, and natural nucleotides (deoxynucleoside triphosphates (dNTPs)) are washed in a similar stepwise procedure to Illumina's cycle, but the incorporation of dNTPs are monitored by a change in pH of the solution due to protons release, which is detected by electronic sensors <sup>132</sup>.

## **Long-read sequencing**

Alternative sequencing technologies capable of sequencing much longer strands are available for metagenomic sequencing. These technologies sequence a DNA molecule, thereby eliminating amplification bias and generate sufficiently long fragments that overlap and facilitate better sequence assembly. Long-read sequencing offers several advantages over short-read sequencing when investigating AMR in bacterial isolates and metagenomic samples. It enables longer *de novo* assemblies capable of resolving repeated sequence regions, thus reducing the complexity associated with assembling short reads. Moreover, it provides better resolution and allows for a more robust interpretation of genetic context, thereby facilitating more in-depth studies to understand the horizontal gene transfer of ARGs <sup>18,60</sup>.

There are two predominant long-read sequencing technologies: Pacific Biosciences single-molecule real-time (SMRT) sequencing and Oxford Nanopore Technologies. The SMRT PacBio, the first long-read sequencer, is commonly used to sequence much longer strands (10-60 kilobases (kb)) compared to Illumina sequencers, which have a typical read length of up to 250 base pairs (bp) <sup>133</sup>. It can detect a single DNA molecule in real-time. The principle behind it is based on DNA replication and the detection of fluorescent light emission signals as each nucleotide is incorporated by a DNA polymerase fixed to the bottom of 50 nm-wide wells, known as zero-mode waveguides (ZMWs) <sup>134</sup>. This allows real-time detection of nucleotide incorporation events during the elongation or synthesis of the replicated strand from the non-amplified single-stranded template. However, these technologies are still costlier than short-read sequencing. PacBio also requires a greater amount of extracted DNA as the libraries are not amplified. As such, metagenomic DNA needs to be pooled together from multiple samples <sup>135</sup>. This is more challenging for metagenomic samples with very low microbial DNA, such as the nasopharynx.

On the other hand, Oxford Nanopore Technologies platforms (Nanopore and MinIon) are routinely capable of sequencing DNA fragments that are hundreds of kb in length. They are generally less expensive and faster than PacBio and short-read technologies, but they have a considerably higher error rate <sup>136</sup>. Unlike other sequencing technologies, they detect changes in electric current or voltage as the nucleotides pass through nanopores (transmembrane proteins). Like PacBio, Nanopore technology does not require DNA amplification and allows for real-time sequencing. Due to its speed, affordability, portability, and the ability to sequence long fragments, Nanopore technology has been a breakthrough in large-scale WGS of microbes and pathogens. For instance, it has been clinically used as a diagnostic tool for the rapid detection of AMR pathogens and ARGs <sup>137-139</sup>, and for surveillance of the Ebola virus <sup>140</sup>. However, more reads are required from single microbes to have enough read depth to overcome high error rate of Nanopore technology and to obtain higher accuracy <sup>141</sup>. This could be achieved by using DNA amplification or concentration techniques, such as those used in Illumina sequencing, prior to sequencing in the Nanopore workflow. The accuracy can also be increased by sequencing not only the template but also its complement strand (2D read sequencing) <sup>142</sup>. However, for resistome profiling from complex metagenomic samples like respiratory tract, DNA amplification techniques are still needed to detect ARGs before sequencing <sup>143</sup>. While most resistome studies have relied on metagenomic data generated from short-read Illumina sequencing <sup>109</sup>, long-read technologies are proving increasingly valuable for linking ARGs with MGEs and their potential host species <sup>120,144</sup>.

In this thesis, WMS data generated using short-read sequencing technologies (shotgun metagenomics) are leveraged to explore the resistome within the human microbiome. The following section discusses the bioinformatics methods and tools used to characterize resistomes in such high-throughput metagenomic sequencing data (Figure 2).

## **Bioinformatics resources for resistome profiling**

### **Computational processing**

After sequencing, high-throughput reads are usually output as plain text files in a format called FASTQ or FASTA. Firstly, the sequencing data is organized, a crucial processing step before it is analyzed for the *in-silico* identification of ARGs. The low-quality sequences and



contamination, such as adapter sequences added during library preparation, are filtered and sequence trimmed using quality control tools<sup>20,110</sup>. FastQC<sup>145</sup> is one such tool that is commonly used to perform quality control check on sequencing reads by calculating various compositional statistics such as base quality, sequence length distribution, CG content, adapter content, etc. Quality filtering and adapter trimming of reads are typically done using software like Trimmomatic<sup>146</sup> and Cutadapt<sup>147</sup>. If the metagenomic data is derived from a host with sequenced genome, like a human or animal, it is also crucial to identify and remove any host sequences (contaminants) using sequence mapping tools before proceeding with any further analysis<sup>110</sup> (Figure 2A).

## Sequence analysis

Due to significant advancements in bioinformatics, there are several bioinformatics methods and tools publicly available to identify ARGs from clean and processed metagenomic reads.

These analysis methods can be categorized into two main categories: Read-based or *De novo* assembly-based mapping<sup>20,110,112</sup> (Figure 2B).

### Read-based profiling

To detect known ARGs, reads can be directly aligned to ARG reference databases such as the Comprehensive Antibiotic Resistance Database (CARD), AMRFinder, ResFinder, and MEGARes. Reads can be aligned using pairwise alignment-based algorithms such as Bowtie 2<sup>148</sup> or Burrow-Wheeler Alignment (BWA)<sup>149</sup>. Alternatively, reads can be split into k-mers, and then mapped to reference database for identification of ARGs using k-mer-based counting algorithm, like KMA<sup>150</sup>, which matches the coverage of k-mer frequencies between reference and query sequences. k-mer-based alignment algorithms are more accurate and sensitive in differentiating between ARGs from databases that contain redundant or highly similar ARG sequences, as these algorithms retrieve exact matches between sequences<sup>20,150</sup>. Though, read-based algorithms have the benefit of calculating the absolute abundance of an ARG present in a sample based on number of reads that are mapped, which can be useful in directly comparing the differences in abundances between samples. In case of k-mer based alignment, a KMA algorithm estimates the total number of k-mers matches with each nucleotide of a reference ARG divided by the total number of nucleotides of that ARG, which is a less precise estimation for ARG abundance. There are several freely available bioinformatics tools and pipelines available to profile ARG from metagenomic



reads using read-based approach including AMRPlusPlus<sup>151</sup>, ARG-OAPs<sup>152</sup>, KmerResistance<sup>153</sup> and shortBRED<sup>154</sup>. AMRPlusPlus is one such commonly used tool that aligns reads to a curated ARG reference database (MEGARes<sup>151</sup>) using BWA aligner. In contrast, KmerResistance splits reads into k-mers, matches them and counts the co-occurrence of k-mers between reads and a reference database to characterize ARG in the sample.

### ***De novo assembly-based profiling***

On the other hand, sequencing reads can be assembled into larger contiguous sequences (contigs) using metagenomic assembler tools such as metaSPAdes<sup>155</sup>, MEGAHIT<sup>156</sup> and MetaVelvet<sup>157</sup>. The open reading frames (ORFs) or protein-encoding regions were predicted on the assembled contigs, which are subsequently annotated for ARG by comparing against ARG reference databases using commonly used pairwise alignment tools such as BLAST<sup>158</sup>, DIAMOND<sup>159</sup> or USEARCH<sup>160,20</sup>. As opposed to the read-based approach, the abundance of ARGs from assemblies are calculated based on the coverage depth of each contig or by mapping the raw reads against the contigs. Assembly-based approach can construct longer contigs or whole genomes that enables not only annotation of ARGs but also other surrounding genetic elements to inform its genetic context; for instance, whether an ARG is a part of an MGE or taxonomic assignment of microbial hosts harboring the ARGs<sup>18,111,161</sup>. This approach can also identify ARGs that are distantly similar or lack homology to known ARG sequences in the reference databases<sup>18</sup>.

### ***Read vs assembly-based approach***

There is no clear consensus on whether read or assembly-based profiling is better, and the choice between approaches strongly depends on the availability of computational resources and research objective<sup>18,20</sup>. Both approaches have their own advantages and limitations. For instance, an assembly-based approach enables more precise identification of protein-coding ARGs and facilitates better exploration of upstream and downstream regulatory elements of detected ARGs, which cannot be accomplished through a read-based method. However, *de novo* assembly is very computationally expensive, may result in data loss and requires high genome coverage compared to read-based approaches, particularly when analyzing complex metagenomic samples with large microbial diversity and uneven composition<sup>20,162,163</sup>. As a result, read-based methods have received increased attention in the recent years, particularly in clinical ARG surveillance

applications due to their ease of computation and faster speed, as they circumvent the *de novo* assembly and ORF prediction steps. Additionally, read-based methods are more sensitive in detecting potential ARGs from low-abundant organisms present in complex communities that may remain undetected by assembly-based methods where assemblies are incomplete or poor<sup>18</sup>. However, it is important to note that direct mapping of unassembled reads to large databases may lead to high false positive predictions because of false mapping of reads to other ARGs due to local sequence similarity<sup>112</sup>. Ideally, one should use tools and pipelines that utilize both read and assembly-based profiling approaches on the same metagenomic data to achieve better resolution of resistome<sup>20</sup>. Nevertheless, both read and assembly-based methods provide *in silico* prediction of ARGs in complex microbial communities and metagenomic samples based upon comparing DNA sequences against publicly available ARG reference databases. Both use existing ARG databases to predict whether a sequence is genuinely an ARG using a sequence similarity and coverage cutoff.

### **Antimicrobial resistance gene (ARG) databases**

There is a plethora of ARG databases available publicly as online platforms, containing information about known genetic determinants of resistance and other AMR-related data for annotation of ARGs. In general, alignment-based tools map nucleotide or protein sequences to an existing database containing reference ARG sequences and return ARG predictions<sup>20,163</sup>. These predictions usually comprise annotations with confidence scores based on the 'best-hit' approach. Thus, *in-silico* prediction of ARGs in metagenomes is heavily dependent on the accuracy and quality of available ARG databases<sup>111</sup>. The functional information on ARGs present in these databases usually represent phenotypic information gathered from several studies, including AST of microbes harboring ARGs<sup>18</sup>. However, these reference databases are considerably different in the types of information that they offer for annotations and the scope of AMR mechanisms that they encompass<sup>164,165</sup>.

These ARG databases can be primarily grouped into two categories: specialized or generalized databases. Specialized databases like Lactamase Engineering Database (LacED)<sup>166</sup>, Comprehensive  $\beta$ -lactamase Molecular Annotation Resource (CBMAR)<sup>167</sup> and User-friendly Comprehensive Antibiotic resistance Repository of *Escherichia coli* (u-CARE)<sup>168</sup> are designed to

provide comprehensive information for specific ARG families or species. On the other hand, generalized ARG databases such as CARD<sup>169</sup>, ResFinder<sup>170</sup>, PointFinder<sup>171</sup>, Antibiotic Resistance Gene-ANNOTation (ARG-ANNOT)<sup>172</sup>, NCBI-AMRFinder<sup>173</sup>, Structured ARG reference database (SARG)<sup>152</sup>, DeepARG-DB<sup>164</sup> and MEGARes<sup>151</sup> have been developed to encompass a broad spectrum of ARGs from all the sources. Nonetheless, these databases also differ in the types of resistance mechanisms that they cover. For example, ResFinder database is specialized and comprises only acquired ARGs, while the PointFinder focuses only on mutations associated ARGs. On the other hand, there are databases like CARD or NDARO, which comprise information on both AMR mechanisms<sup>20</sup>. Most of the metagenomics-based studies which aim at exploring the entire landscape of ARGs present in the microbial communities relied on the generalized databases. Hence for such community-wide studies, it is essential that the reference databases are comprehensive and comprehend all variants of the ARGs. The selection of the appropriate database become more crucial when characterizing resistome from less studied microbial communities such as soil or oceans, as resistance genes that are distantly homologous to known ARGs in database or previously unknown (novel), will remain uncharacterized<sup>174</sup>. Thus, the choice of database is primarily dependent on the objective, type of data and the community or ecosystem studied<sup>175,176</sup>.

However, many of these databases are not updated frequently, lack curation pipelines and contain multiple inconsistencies<sup>1,177</sup>. These generalized reference databases not only differ in their content and scope of mechanisms but also in the annotation and metadata information that they provide for ARG sequences<sup>176</sup>. This makes it impossible to compare the characterized resistome profiles from different reference databases. These databases also lack standardized ARG nomenclature, naming and annotation schema<sup>177,178</sup>. For instance, an aminoglycoside gene *ANT(3'')-Ia* is present under different names across the available ARG databases (*aadA*, *aadA1*, *aadA1-pm*, *aad(3'')(9)* and *ant3ia*). In particular, the classification of ARGs does not follow the standard acyclic, hierarchical taxonomy scheme used for microbial organisms, which does not allow the same microorganism to link to multiple categories or groups. Use of cyclic annotation structure may lead to inflation of counts during the downstream analyses of resistome profile at different annotation levels<sup>151</sup>. However, there are databases such as MEGARes and SARG which have manually curated a simple, acyclic and hierarchical functional scheme for annotation

of ARGs, suitable for accurate downstream analyses. More recently, most up-to-date databases like CARD and ResFinder have also adopted similar annotation structure and classification scheme.

A few comparative studies assessing the performance of different databases suggested that the most frequently updated and comprehensive database (i.e., CARD) performs better and should be the preferred choice for predicting well-annotated ARGs in most cases <sup>20,176</sup>. NDARO and ResFinder are somewhat superior resources compared to CARD in correctly predicting the ARGs related to acquired resistance. It is important to note that all these databases are biased towards ARG sequences derived from clinically relevant human pathogens and easily culturable bacteria, imposing difficulties in detecting remote homologs or novel ARG sequences found in uncultured or fastidious bacteria. A potential solution to mitigate this bias includes using Hidden Markov model (HMM)-based databases such as Resfams <sup>179</sup>, or another database like Mustard <sup>180</sup>, which is based on three-dimensional (3D) protein structure. These databases utilize structural and functional similarities for identifying resistance proteins, which might be difficult to detect based solely on sequence similarity or alignment.

## **Downstream analysis**

Both read-based and assembly-based bioinformatics pipelines involve preprocessing of raw sequencing reads, which are then summarized into ARG count abundance tables (Figure 2C). These tables represent the number of reads that align with each representative ARG sequence in each sample. Accompanied by associated sample information (i.e., metadata) and functional annotations, these ARG tables serve as the primary input for downstream analyses and functional interpretation.

However, such analyses are not straightforward due to inherent challenges associated with count data analysis <sup>181,182</sup>. The ARG abundance tables are characterized by: (i) large variation in the number of sequencing reads obtained between samples (uneven sampling depth) due to sequencing effort, bias, or inconsistencies in library preparation; (ii) excessive “zero” values (i.e., sparsity), resulting from under-sampling or the actual absence of features; (iii) skewed distribution (not normally distributed) due to a high proportion of zeros; (iv) high dimensionality, given the large number of features typically analyzed in a single sample; and (v) overdispersion (i.e., the variance

is much higher than the mean). Furthermore, the total number of counts per sample is constrained by the maximum number of sequence reads that a DNA sequencer can generate. This total count constraint induces strong interdependencies among the abundances of the different ARGs, preventing them from being interpreted independently, thus characterizing the compositional nature of metagenomic data <sup>181</sup>. Ignoring the compositionality in metagenomic data may yield biased and misleading results <sup>183</sup>. Moreover, classical statistical methods are insufficient for directly analyzing such data <sup>112,184</sup>. Recently, compositional data analysis (CoDa) methods have been proposed to overcome this issue by transforming the read counts to ratios of read counts, for example, centered log-ratio (CLR) – log ratio of each feature count to the geometric mean of all feature counts within a sample <sup>183,185</sup>. The log-ratio eliminates the restraint of the total read count capped by sequencing platform and transforms the data into an unbounded (Euclidean) space where standard statistical methods can be used. All these inherent characteristics of metagenomic data significantly influence data exploration, visualization, and statistical analysis. Consequently, it is crucial to properly normalize or transform the data before analysis and to employ statistical methods that can manage these challenges, to yield comparable samples and meaningful results <sup>111,186</sup>.

Downstream analysis of the ARG count abundance table – also referred to as resistome abundance data – typically begins with normalization. This is then followed by the exploration of resistome composition and diversity, employing a variety of visualization and diversity analysis methods that are standard practices in microbial community studies. The exploratory analysis typically involves analyzing dissimilarities between samples (using clustering and ordination techniques), assessing associations between features such as ARGs using correlation analysis, and visualizing the results using ordination plots, dendrograms, and heatmaps. These approaches can help identify potential hidden patterns or group structures within the data <sup>181</sup>. Exploratory analysis can also be performed at higher functional levels, such as ARG class or mechanisms of action, to gain a more comprehensive understanding of the resistome data. Subsequently, an inference analysis is conducted to determine if there is an association between resistome composition and a variable of interest. Association tests can be either multivariate (when investigating overall differences in resistome composition among sample groups) or univariate (when identifying differentially abundant features or ARGs between sample groups). Finally, an integrative analysis is performed

to explore the association between resistome and microbiome composition using multivariate and univariate correlation-based omics integration methods to gain insight into the complex interactions between microbial community composition, function, and antibiotic resistance.

In the following sections, various types of downstream analysis approaches for resistome data are discussed. These include normalization, diversity profiling, differential abundance testing, and integrative analysis. Some of the most popular methods for these steps are also highlighted, distinguishing between standard methods and those suited for compositional data.

### **Data normalization**

Normalization aims to remove or reduce the systematic variability from the data, which may arise due to technical factors (above mentioned) and biological factors, and thus underline the true biological differences between samples <sup>187</sup>. One commonly used normalization approach is to rarefy the count abundance table to the same depth, thereby accounting for differences in library sizes <sup>186,188</sup>. Rarefaction works by randomly subsampling the same number of reads for each sample to the size of the smallest library such that all samples have the same number of total read counts. Despite being a straightforward approach, it has been criticized due to the potential loss of useful data as it excludes sequences from samples with larger library sizes from being analyzed<sup>188</sup>.

The second set of normalization methods is based on scaling the data. These methods account for differences in sequencing depth by multiplying or dividing feature (gene) counts with a sample-specific scaling factor, transforming raw reads into relative abundances. The simplest and most commonly used approach is to divide read count abundances by the total number of counts in each sample (known as Total Sum Scaling (TSS)) <sup>184,189</sup>. Additionally, the resulting relative abundances (proportions) are sometimes multiplied by a constant sum (e.g.,  $10^6$  (million)) to obtain the feature abundance per million total read counts for easier interpretation of data. However, the resulting relative abundances after such normalization can be biased, particularly when the total number of reads in a sample is dominated by a few highly abundant features <sup>190</sup>. This can result in an overestimation of the abundance of these features and an underestimation of the abundance of other less abundant features <sup>191</sup>. To overcome this challenge, other scaling factors, such as upper

quantile (UQ <sup>192</sup>) and cumulative sum scaling (CSS <sup>193</sup>), have been recommended. For instance, CSS calculates the scaling factors as the cumulative sum of observed count abundances for each sample up to a data-derived percentile threshold to reduce the biases resulting from preferentially sampled features (genes). More complex normalization techniques, including variance stabilizing transformations such as relative log expression (RLE) and trimmed mean of M-values (TMM) methods are also commonly used to compare differential abundance between genes <sup>194</sup>. Recently, the CLR transformation is becoming a more popular method <sup>195</sup>. As described earlier, it uses a log-ratio approach rather than relative abundances between features to deal with the data compositionality.

To date, there is no consensus on which method performs optimally and should be adopted for all types of datasets and downstream analyses <sup>184,186,187</sup>. The choice of normalization methods can influence the results of downstream analyses <sup>184</sup>. Moreover, studies assessing the performance of different normalization methods have demonstrated that depending on the type of analysis and data characteristics (such as sample size, sequencing depth, etc.), some methods can perform equally well <sup>186,187,196,197</sup>. Therefore, a careful evaluation of data characteristics and the selection of an appropriate normalization method are crucial for achieving the most accurate results in metagenomic data analysis.

## **Diversity analysis**

Resistome data are multi-dimensional, represented by hundreds or thousands of different ARGs. Therefore, it is essential to evaluate not only the abundance of individual ARGs but also the overall diversity of associated ARGs to obtain meaningful insights. Similar to microbial ecology, diversity is typically described or estimated within (alpha diversity) or between (beta diversity) samples.

### ***Alpha diversity***

Alpha diversity is a measure of diversity within an individual sample. It is often characterized using various diversity indices that account for either the total number of features (i.e., richness), their relative abundances (i.e., evenness), or both <sup>111,198</sup>. For instance, richness-based indices such as 'Observed' calculates the actual number of features (in the case of resistomes, ARGs) present in

each sample. Indices like Chao1 and ACE, on the other hand, estimate richness by extrapolating the number of rare features that may be undetected due to undersampling. These indices introduce a correction factor to the observed number of features, accounting for rare ones by resampling the data and recalculating the estimator multiple times<sup>199</sup>. As these indices only consider the presence or absence of features without their abundances, they are referred to as qualitative measures. Other metrics, such as Shannon and Simpson, consider both richness and evenness to describe the diversity within a metagenomic sample, with varying importance given to evenness. Additionally, they are less sensitive to the number of sequences per sample in comparison to richness-based indices<sup>200</sup>.

However, it remains unclear which diversity indices are most suitable for estimating ARG diversity in metagenomic samples. A recent study that evaluated the predictive power of different indices suggested that the Chao1 estimator performs well in approximating ARG diversity in metagenomic datasets<sup>201</sup>. Nevertheless, each diversity index has its strengths and limitations, and the choice of the most appropriate one depends on the specific dataset, sample composition, and research question<sup>184</sup>. In general, more than one indices is presented in resistome papers, as they assist in providing a more comprehensive interpretation of resistome data.

### ***Beta diversity***

Beta diversity represents the differences between samples. It is used in resistome studies to measure the similarity or dissimilarity in the overall resistome composition between samples<sup>113</sup>. In a similar vein to alpha diversity, there are several different ecological-based indices (distances) for quantifying beta diversity. There are several different ecological-based indices (distances) for quantifying it. Each of the measures reflects different aspects of composition heterogeneity. For instance, some of the most commonly used indices, such as the Bray-Curtis distance, incorporate the abundance of features (ARGs). In contrast, qualitative indices such as the Jaccard index and Jensen-Shannon divergence (JSD) focus on the presence or absence of features, ignoring their abundance, to estimate dissimilarities in resistome composition<sup>189</sup>. Abundance-based measures are less sensitive to misclassification errors involving small numbers of reads, insufficient sampling, or variations in sequencing depth across samples<sup>202</sup>. Depending on the specific research question, some measures may more accurately reflect changes in the



overall resistome composition, whether the focus is on ARGs that are more abundant or those that are relatively rare<sup>184</sup>. Regardless, these beta diversity estimates can be summarized as a pairwise distance matrix. This matrix contains for each sample the similarity or dissimilarity score to every other sample and can be further visualized to identify patterns using ordination techniques.

Ordination methods are often used to visualize beta diversity data by summarizing large and complex dissimilarity or distance matrices between samples in lower-dimensional spaces or fewer principal components (2D-3D plots). The goal of these plots is to visualize patterns in resistome composition, identify clusters or groups of samples with similar resistome composition, and elucidate the impact of specific experimental factors on resistome composition<sup>113,181,200</sup>. Samples that are closer to each other in these plots indicate a higher degree of similarity in their resistome composition. Principal coordinate analysis (PCoA) and non-metric dimensional scaling (NMDS) are the most commonly used and widely accepted ordination methods to explore differences in resistome composition<sup>181</sup>. However, these methods require the dissimilarity or distance matrix as an input and are sensitive to the distance method used<sup>203,204</sup>. Therefore, identified clusters or patterns should be confirmed by multiple methods<sup>184</sup>. Another frequently used ordination method is principal component analysis (PCA), which is simply PCoA using Euclidean distance. PCA based on the Euclidean distances between the CLR data (i.e., Aitchison distance) can serve as an alternate CoDa-based approach for beta diversity analysis, which can account for data compositionality<sup>183</sup>.

### ***Significance testing of alpha and beta diversity***

To determine whether there are statistically significant differences in alpha diversity measures between the groups or conditions under investigation (predictor variables), parametric tests such as ANOVA and t-test or non-parametric tests such as the Wilcoxon rank-sum test or Kruskal-Wallis test are commonly used<sup>184</sup>. However, these tests may not perform as efficiently in handling continuous predictor variables, unbalanced designs, and missing data. In such instances, linear models like the generalized linear model (GLM) are often more suitable. These models also allow control for multiple confounding variables (both categorical and continuous) to yield more reliable results regarding the effect of the main (predictor) variable of interest on the alpha diversity measure<sup>202</sup>. In the context of experimental design, it is important to note that certain study designs,

such as repeated measures, may introduce correlations between samples. For instance, in longitudinal resistome studies, repeated sampling from the same individual may produce samples that are more similar to each other in terms of composition and diversity (inter-individual variation) compared to samples from different individuals<sup>205</sup>. Hence, it is crucial to use statistical methods that can account for such between-sample correlations<sup>206</sup>. More sophisticated tests, such as Linear mixed-effects models (LMEs), are gaining popularity for the comparative analysis of alpha diversity measures<sup>29,207</sup>. These models control for potential confounding variables (fixed effects) and account for between-sample correlation (included as random effects), thereby improving the power and accuracy of identifying significant differences between groups.

On the other hand, to test the statistical significance of beta diversity differences between groups, multivariate permutation-based tests such as permutational multivariate analysis of variance (PERMANOVA), analysis of similarities (ANOSIM), or the test for homogeneity of group dispersions (PERMDISP) are commonly used<sup>20</sup>. These tests are more powerful than classical methods (such as analysis of variance (ANOVA) or multivariate analysis of variance (MANOVA)) and can be used to assess global differences in resistome composition among groups<sup>184</sup>. Most of these tests compare the distances within samples of the same group to the distances between groups based on a dissimilarity (or distance) matrix. PERMANOVA is one of the most widely accepted robust tests that can handle the inclusion of multiple potential confounding variables (both categorical and continuous) and within-sample correlation to test the independent impact of predictor variable(s) on the composition<sup>208</sup>. It tests the null hypothesis that the 'centroids' of all groups are not significantly different from each other<sup>209</sup>. However, this method is sensitive to differences in variance within groups (multivariate dispersion). Therefore, the PERMDISP test can be used in conjunction with PERMANOVA to determine whether the variances of distances within groups differ from the variances of distances between groups. To address the issue of dispersion heterogeneity, it is recommended to use ANOSIM, a method that is more sensitive to such heterogeneity<sup>210</sup>. This test assesses the hypothesis that the dissimilarities between groups are greater than or equal to the average dissimilarities within groups based on the distance matrix, using the ranks of all pairwise sample distances<sup>211</sup>.

## Differential abundance analysis

The primary goal of differential abundance analysis (DAA) is to identify features with significantly different relative abundances between two or more conditions (sample groups) of interest <sup>181</sup>. In resistome studies, this analysis is commonly employed to detect significantly differentially abundant ARGs, classes or mechanisms associated with various conditions or interventions, such as antibiotic therapies. Although it may seem straightforward, several unique challenges (zero-inflation, overdispersion, skewed distribution, uneven sequencing depths, and compositionality) arise due to the inherent features of metagenomic count data, making it unsuitable for direct application of statistical methods developed in other omics fields. A widely used approach involves nonparametric (including permutations-based) tests, such as the Wilcoxon Rank-Sum or Kruskal-Wallis tests, after normalizing the count data to identify differentially abundant genes. Linear discriminant analysis (LDA) effect size (LEfSe) combines these standard nonparametric univariate tests for statistical significance with methods related to biological consistency and effect size estimation for discovering biomarkers in metagenomic datasets <sup>212</sup>. However, these standard nonparametric methods lack statistical power (high false discovery rates) and cannot model confounding factors to accommodate complex study designs <sup>188,213</sup>. Moreover, they do not account for the complex inherent features of metagenomic abundance data, which may significantly impact the results of DAA. Therefore, it is generally more appropriate and recommended to use statistical models that either assume distributions accounting for count data characteristics or employ techniques to transform the data (normalization) to fit standard test assumptions <sup>111</sup>. In recent years, significant advancements have been made in this area. For example, the metagenomeSeq (R package) tool implements a novel normalization (Cumulative sum scaling) technique combined with a statistical framework for modeling count data with zero-inflated Gaussian (ZIG) or zero-inflated Log-Gaussian (ZILG) distributions to account for sparsity (zero-inflation) and undersampling-related bias in metagenomic count data <sup>214</sup>. Other distribution-based models, such as DESeq2 <sup>215</sup> and edgeR <sup>216</sup> (originally developed for RNA-seq data analysis), are also frequently used and appear to perform better than many other methods for detecting differentially abundant genes in metagenomic datasets <sup>196</sup>. These methods fit a generalized linear model and assume that count abundances follow a negative binomial (NB) distribution after normalizing data with corresponding scaling-based methods to address uneven sequencing depths. However, these RNA-Seq inspired methods may also produce spurious results, as some of their assumptions are not

entirely valid for metagenomic data <sup>196</sup>. Additionally, these methods do not directly account for the compositional nature of sequencing data.

More recently, two CoDa methods, ANOVA-Like Differential Expression tool for compositional data (ALDEx2<sup>217</sup>) and Analysis of Composition of Microbiomes (ANCOM<sup>218</sup>), have been designed for differential abundance analysis of compositional high-throughput sequencing data. These methods intrinsically remove the effect of sampling fraction (compositionality) by transforming the observed abundances to log-ratios. ALDEx2 uses the geometric mean of observed abundance (relative) of all features within each sample as a reference for converting the abundances of each feature to log ratios for that sample (i.e., CLR transformation). ALDEx2 performs a Monte Carlo sampling with an underlying assumption that the data follow a multivariate abundance distribution based on a Dirichlet-multinomial model to account for technical and biological variation <sup>191</sup>. After CLR transformation, the log-ratio of all pairs of features is tested for differences across different sample groupings using standard parametric or non-parametric tests, such as Welch's t-test, Wilcoxon test, Kruskal-Wallis tests, or one-way ANOVA using a generalized linear model (GLM). In comparison, ANCOM uses a pre-specified feature as a reference and converts the observed count abundances to log ratios of the observed abundance of each feature relative to the reference feature (aka additive log transformation (ALR)) within each sample. ANCOM tests the log-ratios of the abundance of each feature to the abundance of every other feature for the differences in their means across groups using non-parametric tests. The proportion of significant results involving each feature is used to determine its significance <sup>181</sup>. Most of these DAA methods are also able to model confounding covariates to accommodate complex study designs. Although flexible, these methods cannot explicitly model the within-subject correlations for repeated measures (random effects) in longitudinal studies <sup>206</sup>. Alternative, more advanced tools and methods, such as MaAsLin2 <sup>219</sup> or mixed-effects models, may be more suitable for accounting for random effects in longitudinal study designs <sup>206</sup>.

To date, most benchmarking studies have evaluated the performance of different DAA methods on microbiome datasets (mainly those derived from 16S rRNA sequencing) rather than on metagenomic datasets <sup>186,191,196,220,221</sup>. Nonetheless, these studies have analyzed different sets of methods and dataset types, leading to less agreement regarding the performance of tools across studies <sup>213</sup>. Currently, there is no clear consensus on the best DAA method for a specific dataset.

Furthermore, comparative studies have suggested that no method is simultaneously robust, powerful, and flexible for all types of datasets and study designs<sup>220</sup>. The performance and outcome of different DAA methods are significantly influenced by data characteristics, such as sample (or group) size, sequencing depth, gene abundances, and effect size differences<sup>184,196</sup>. Thus, various methods may be needed for different metagenomic datasets and research questions addressed. Nevertheless, these large-scale comparative benchmarking studies can provide valuable guidance to researchers when selecting methods for their own data. Moreover, it is recommended that researchers either employ multiple methods and focus on significant features (genes) detected by most of the approaches or utilize a consensus approach based on several DAA methods to ensure the robustness and validity of the results of differential analysis tool choice<sup>213</sup>.

Regardless of the method employed for detecting significantly differentially abundant ARGs in a group of samples, a probability ( $p$ ) value will be derived for each ARG feature examined. Consequently, in multi-dimensional resistome profiles characterized using a large reference database, this process entails conducting hundreds or even thousands of tests. When multiple features (ARGs) are tested simultaneously, the likelihood of false-positive (Type I errors) observations increases<sup>222</sup>. Thus, it is essential to apply a correction for multiple testing to control the false discovery rate (FDR)<sup>111</sup>. Most DAA methods automatically perform such FDR corrections using widely accepted methods, such as the Bonferroni or Benjamini–Hochberg procedures, to adjust the  $p$ -value and obtain more reliable and robust results in differential analysis.

### **Integrative analysis**

The resistome and taxonomic composition are closely interconnected, with each influencing the other in several ways. Understanding the complex interplay between the resistome and microbiome is crucial for advancing our understanding of the ecology of AMR in microbial communities and to gain more novel insights into the interactions between microbial communities and their environments, including their response to antibiotic exposure. To identify and quantify these potential associations, microbiome and resistome count abundances of the same samples can be integrated using a variety of statistics-based computational approaches<sup>223</sup>. These integrative analyses are becoming increasingly popular to explore which microbial taxa are associated with antibiotic resistance, as well as which ARGs are present in those taxa across diverse

metagenomes, including humans <sup>47</sup>, animals <sup>224,225</sup> and environments <sup>226</sup>. Currently, such analyses are exploratory in nature with an aim to reveal inherent patterns and trends from the datasets mainly using correlation-based statistical methods. One of the simplest approaches for microbiome and resistome integration is univariate correlation analysis, such as Pearson and Spearman correlation. These methods enable us to determine if there are strong pairwise associations present between individual ARGs (resistome) and taxa (microbiome) based on different measures such as correlation and co-abundance. For example, V Carr *et al.* performed Spearman correlation analysis between the human oral microbiome and resistome to predict the <sup>46</sup>. However, they are prone to false positives due to the high dimensionality and sparsity of metagenomic datasets. Multivariate correlation methods based on dimensional reduction techniques have become the predominant methods to conduct integration of microbiome and resistome data. For example, Procrustes analysis is a statistical method that utilizes dimension reduction techniques such as PCA or PCoA for visual integration of microbiome and resistome dataset <sup>227</sup>. It essentially correlates the principal components or coordinates of two datasets at the lower-dimensional space (ordinations) rather than individual features, allowing users to assess the overall similarity between microbiome and resistome data. For instance, J Feng *et al.* used PA to establish that microbial phylogeny structured the antibiotic resistome in healthy human gut microbiota <sup>47</sup>. More advanced univariate and multivariate correlation-based methods have been recently developed for the integration of different paired omics datasets, such as CCLasso, Maximal Information Coefficient (MIC), regularized canonical correlation analysis (rCCA), and sparse partial least squares (sPLS). These methods can also be used to integrate microbiome and resistome data while addressing issues of sparsity and compositionality in metagenomic data <sup>228,229</sup>. In general, the integrative analysis of omics data remains an active research area, with an increasing availability of novel statistical methods that have yet to be extensively evaluated for metagenomic datasets.

In addition to statistical-based approaches, microbiome and resistome data can be integrated by leveraging our existing knowledge framework of the microbial hosts that harbor or carry ARGs (knowledge-driven integrations). Information on these relationships can be acquired either directly or indirectly from public ARG databases, utilizing text mining techniques or through manual data collection. Such information can be complex due to the presence of multiple ARGs in a single

microbial taxon and the distribution of single ARGs across various taxa. This can be intuitively represented as interaction networks to gain more comprehensive insights into AMR mechanisms and shed light on possible dissemination routes of ARGs. However, currently, there are no such visual analytics tools leveraging the existing knowledge present in available databases to explore the known associations between ARGs and microbial taxa.

## Knowledge and Research Gap

AMR profiling using metagenomic sequencing data is still an evolving field with many research opportunities. In this thesis, three different research challenges are addressed, as mentioned below:

- Metagenomic-based sequencing studies have significantly enhanced our understanding of resistomes and the ecological impact of interventions like antibiotic therapy on complex microbial communities such as human microbiomes. However, our knowledge on the collateral effects of antibiotics on the selection and development of AMR is still limited compared to disturbances in microbial community composition. Furthermore, there is a lack of long-term follow-up studies examining the lasting effects of antibiotic exposure on the microbiome and resistome. Existing studies have primarily focused on short-term (<7 days) and immediate effects of antibiotic therapies <sup>66,73,74,123,230</sup>. However, longer treatments are commonly prescribed for conditions like recurrent otitis media <sup>231</sup>, urinary tract infections <sup>232</sup>, chronic low back pain <sup>233</sup>, and chronic respiratory conditions <sup>234</sup>, with the potential for long-lasting effects on microbiomes and the associated resistomes. Such effects remain largely uncharacterized, even for the gut, which has been the focus of most microbiome studies.
- Respiratory infections caused by common pathogens, including those that are commonly present in the nasopharynx, are the leading cause of deaths associated with AMR worldwide <sup>15</sup>. Nevertheless, the landscape of ARGs (i.e., the resistome) present in the microbial communities of the nasopharynx remains largely unexplored. Consequences of preterm birth and respiratory infections are the leading causes of mortality globally in early life <sup>235</sup>. Infant's microbiota is immature, and their immune system is not yet fully functional, making the negative consequences of antibiotic treatment more acute and long-lasting than in adults <sup>236,237</sup>. Preterm infants, in particular, are often given empiric antibiotics within 72 hours of birth, mainly due to suspected infections or sepsis <sup>238,239</sup>. Recent metagenomics-based studies have also shown that early antibiotic exposure during the microbiome maturation phase has persistent and detrimental effects on the microbiome and selection of AMR <sup>53</sup>. Furthermore, such exposure during infancy has been linked to numerous unintended adverse short- and long-term health consequences <sup>239-242</sup>. However, most of this knowledge comes from gut



microbiota studies. In general, the dynamics of the resistome and the factors that influence its trajectory are still poorly understood, even for the gut<sup>42</sup>. For other body sites, such as the nasopharynx (a reservoir of respiratory pathogens and a gatekeeper of respiratory health<sup>243</sup>), knowledge on the dynamics that shape resistome establishment and the effect of such early antibiotic exposure remain largely unknown.

- Significant advancements in high-throughput DNA sequencing (HTS) technologies and the development of computational tools have accelerated the rapid identification and characterization of the landscape of ARGs (i.e., resistome profiles) in complex and diverse microbial communities, including humans, animals, and the environment<sup>18</sup>. However, the data generation capabilities are currently outpacing the development of computational tools and pipelines. To date, the primary computational effort in data analysis has focused on raw data processing, annotation, and abundance estimation of ARGs in metagenomic sequencing data. Consequently, the subsequent step of data understanding and interpretation of these characterized resistome abundance profiles remains a critical bottleneck in the field<sup>112</sup>. This is mainly due to the fact that analyzing such data is exploratory in nature (no gold standards)<sup>184</sup>, and it typically requires interdisciplinary skills in both bioinformatics and biological knowledge to perform in-depth analysis. The available bioinformatics tools are designed to accomplish specific data analysis tasks and are often challenging for bench scientists and clinical researchers due to their extremely technical nature and the requirement for basic programming knowledge. However, there are currently no easy-to-use bioinformatics platforms for comprehensive visual, statistical and functional analysis of high-throughput metagenomic resistome data.

## **Aims of the research**

### **Overall aim:**

The overarching objective of this work was to deepen our understanding on the ecological consequences of antibiotics on resistomes and to develop a bioinformatics platform that facilitates resistome data analysis and interpretability.

More specifically, the aims of this thesis were to:

- assess the impact of prolonged amoxicillin therapy on the human gut microbiome and resistome. (Paper I)
- investigate the development dynamics and impact of early-life antibiotics on the nasopharyngeal resistome of preterm infants. (Paper II)
- develop a user-friendly, easily accessible and intuitive bioinformatics platform to democratize the analysis, visualization and interpretation of metagenomics resistome data. (Paper III)

## Summary of the results

The main results from each of the studies are summarized in the following section.

### Paper I

#### *Differential response to prolonged amoxicillin treatment: long-term resilience of the microbiome versus long-lasting perturbations in the gut resistome*

In this study, fecal specimens collected from 20 adult subjects, randomly assigned to either a placebo group ( $n = 12$ ) or an amoxicillin-treated group ( $n = 8$ ), were investigated to explore the long-term consequences of prolonged antibiotic treatment. These subjects were participants in a larger double-blind, placebo-controlled, randomized, multicenter trial (The AIM study), which evaluated the clinical efficacy of three months of amoxicillin treatment compared to placebo in patients with chronic low back pain and Modic changes. Fecal samples were collected from the selected patients at three time points: before the initiation of treatment (baseline), immediately after the cessation of treatment (3 months), and 9 months post-antibiotic or placebo exposure (12 months). Approximately 200 Gb of metagenomic DNA from 60 fecal samples were analyzed through shotgun metagenomic sequencing to elucidate the long-term effects of prolonged amoxicillin treatment on the human gut microbiome and resistome.

**Impact of amoxicillin on Gut microbiome:** The microbiome composition was primarily dominated either by *Bacteroides* or *Prevotella* genera across all the studied participants in both the amoxicillin and placebo groups. Our results showed that prolonged amoxicillin exposure led to a significant reduction in species richness and a shift in the overall microbiome composition immediately after cessation of treatment (3 months). Further, significant differences were detected among species with low abundance in the amoxicillin-treated group. In particular, health-associated short-chain fatty acid (particularly butyrate) producers significantly decreased in proportion immediately after exposure to amoxicillin (3 months) as compared to baseline. However, all these perturbations due to amoxicillin treatment were transient as no significant alterations in microbial diversity, composition, and abundances were observed at 9 months post-treatment.

**Impact of amoxicillin on Gut Resistome:** A total of 147 unique ARGs belonging to 15 ARG classes that confer resistance via 5 distinct mechanisms were identified across all the fecal samples. In both groups, tetracycline resistance genes were the most abundant class of ARGs in most patients. Higher inter-individual variation in the resistome than in microbiome composition ( $R^2 = 0.82$  vs  $0.79$ ) was observed, with many ARGs detected in a limited number of patients. Contrary to the microbiome results, the effects of amoxicillin exposure on the resistome persisted for long-term after treatment termination. The results showed an increase in the total AMR abundance and richness (Shannon index) of ARGs upon amoxicillin treatment, which remained significantly higher even at 9 months post-treatment compared to baseline (all *adjusted p* < 0.05). In particular, a significant portion of these long-lasting changes seemed targeted to the antibiotic treatment, as an enrichment of beta-lactamase ARGs in response to amoxicillin was observed. This included instances showing the enrichment of beta-lactamase genes associated with extended spectrum of activity (ESBLs) and those acting against carbapenems, which are clinically important as they could severely affect potential treatment outcomes. These responses to amoxicillin interventions were highly individualized and particularly evident in the resistome results.

In the placebo group, no significant changes in the microbiome and resistome diversity and composition were detected over time. This underlines the relative stability and robustness of the healthy adult human gut microbiome and resistome without antibiotic perturbations. Overall, the findings suggest that prolonged amoxicillin exposure has a more common, pronounced yet transient impact on the microbiome, compared to the more individualized and long-lasting changes in the human gut resistome.

## **Paper II**

### ***Unravelling the landscape of antibiotic resistance determinants in the nasopharynx and the impact of antibiotics: a longitudinal study of preterm infants***

This prospective observational cohort study involved preterm infants born at the Neonatal Intensive Care Unit (NICU) of Oslo University Hospital, Ullevål, Norway. To investigate the

respiratory resistome development dynamics, nasopharynx aspirate samples were collected at six distinct time points, spanning from birth to 6 months corrected age (T1-T6). This paper focuses on a subset of preterm infants, namely those who had nasopharyngeal samples collected on the same day as the initiation of broad-spectrum antibiotic treatment (amoxicillin + gentamicin) for suspected early-onset neonatal sepsis ( $n = 15$ ), as well as antibiotic-naïve infants as controls ( $n = 21$ ), to assess the temporal consequences of early antibiotics on the nasopharyngeal resistome. Deep shotgun metagenomic sequencing was utilized to characterize the resistome of 181 nasopharyngeal samples, which were longitudinally collected from 36 preterm infants.

**Nasopharynx resistome composition:** In total, 373 ARGs conferring resistance to 15 different classes of antibiotics were identified across all samples. These ARGs were detected in the majority (~95%) of the samples in both the antibiotic-treated and naïve groups, including potentially high-risk ARGs. Among these, multidrug resistance genes and efflux pump mechanisms were the most dominant types of ARGs present in the nasopharynx of preterm infants.

**Impact of early antibiotics:** Early antibiotic exposure to broad-spectrum antibiotics (amoxicillin + gentamicin) was associated with a significant increase in the diversity (Shannon) and total abundance of ARGs directly after the cessation of treatment (T2). A few high-risk and clinically relevant ARGs that appeared only at T2 were also observed in antibiotic-treated infants. Further, a minor yet significant shift in the overall resistome composition was detected between the antibiotic-treated and naïve infants immediately after the end of treatment (T2). However, no significant changes in diversity, abundance, composition, or carriage of high-risk ARGs persisted at 6 months of corrected age in antibiotic-treated compared to naïve groups, suggesting that the effects of early antibiotics on the nasopharynx resistome in preterm infants were transient.

**Factors shaping the nasopharynx resistome:** A strong correlation between ARG abundance and taxonomic abundance profiles ( $r = 0.85$ ,  $p = 0.001$ ) was observed, suggesting that the microbial community composition structured the resistome composition in the nasopharynx of preterm infants. The overall relative abundance of the resistome was moderately associated (29%, adjusted  $p = 0.01$ ) with *Streptococcus mitis* relative abundance in the nasopharynx microbiome. Inter-individual variation ( $R^2 = 30.03\%$ , adjusted  $p = 0.001$ ) and age were found to be the main factors

explaining most of the resistome compositional variation observed in the overarching cohort. Additionally, prenatal maternal antibiotic exposure ( $R^2 = 3.6\%$ , adjusted  $p = 0.001$ ) was found to be a significant clinical covariate influencing the overall nasopharynx resistome composition, which was a subgroup found only in the early antibiotic group of preterm infants.

Overall, the study provides to the best of our knowledge, the most comprehensive understanding of the characteristics and dynamics of resistome development in the nasopharynx of preterm infants, spanning the first six months of life, which is a crucial period for airway development. The results suggest that the nasopharynx in preterm infants is enriched with a diverse resistome and early antibiotic exposure can negatively impact its development, although the effects are transient. Adverse changes to the resistome during early life pose a risk, particularly for preterm infants with immature immune and organ systems. The study highlights the need to focus on this understudied yet critical reservoir of pathogens and ARGs, given the significant contribution of respiratory pathogens to the global AMR burden.

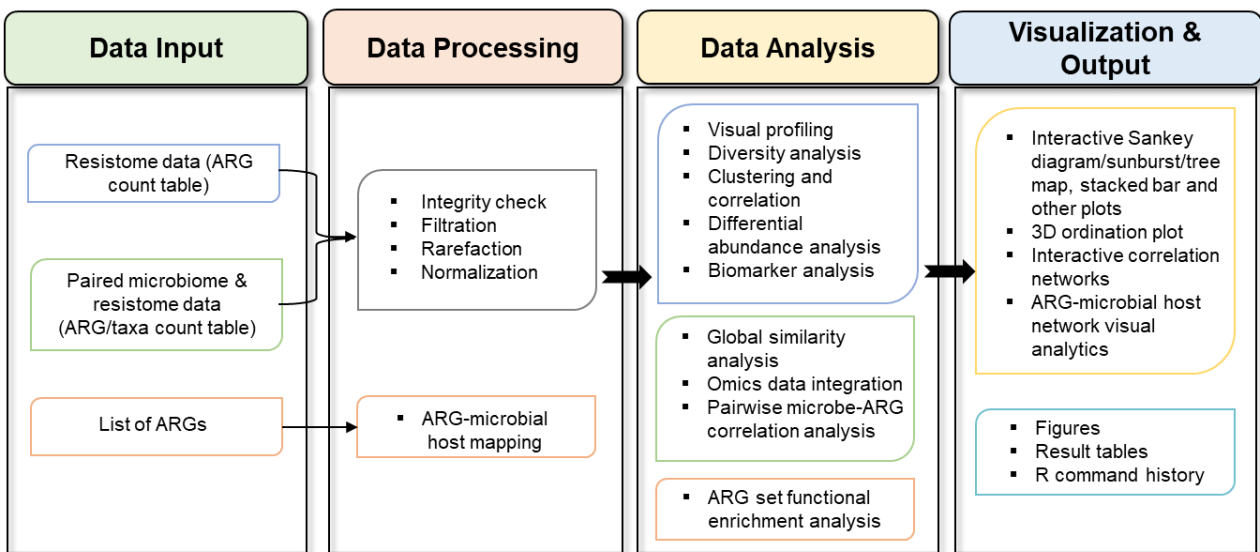
## **Paper III**

### ***ResistoXplorer: a web-based tool for visual, statistical and exploratory data analysis of resistome data***

This paper introduces ResistoXplorer, a user-friendly, web-based platform for comprehensive downstream analysis of resistome data generated from AMR metagenomic studies. The tool was implemented using three programming languages: Java, R, and JavaScript, and consists of three main modules for data processing, analysis, and result exploration. The accompanying web interface provides a wide variety of options and produces several tabular and graphical outputs to assist users in intuitively navigating through different analysis tasks. It is publicly available at [www.resistoxplorer.no](http://www.resistoxplorer.no).

ResistoXplorer supports a comprehensive suite of standards and advanced methods for composition profiling, functional profiling, comparative analysis, and statistical analysis of resistome abundance data. The tool offers several intuitive visualization methods to provide a comprehensive view of the diversity, abundance, and composition of resistome profiles. It also

includes functional annotation information from several reference databases, allowing for analysis and profiling of resistome at higher functional categories or levels. The "Integrative analysis" module in ResistoXplorer enables users to perform integrative analysis on paired microbiome and resistome data using several univariate and multivariate omics integration statistical methods. In addition, ResistoXplorer's network-based visual analytics system allows users to explore and understand the complex relationships between ARGs and microbial hosts.



**Figure 3:** Workflow of ResistoXplorer

ResistoXplorer fills an important gap between data generation and analysis by offering comprehensive support for visual, statistical, and exploratory analysis of resistome profiles and signatures generated from AMR metagenomics studies. Overall, ResistoXplorer enables researchers and members of interdisciplinary groups without prior bioinformatics expertise to easily explore and analyze resistome datasets using a variety of approaches in real-time and through interactive visualization. The tool facilitates data understanding, hypothesis generation, and knowledge discovery in the field.

## **Methodological considerations**

In the following section, methodological considerations pertaining to Papers I-III are discussed. The section is divided into two main parts due to the nature of the work and methodologies used. The first part discusses the common challenges, methodological choices, and considerations related to the application of shotgun metagenomics in clinical studies (Papers I and II). The second part discusses the considerations involved in developing an easy-to-use, comprehensive downstream bioinformatics tool, as presented in Paper III. This includes the design aspects, choices of visual and analytical methods, the selection of supported tools, methods and parameters, as well as other factors that can influence the reliability and reproducibility of the downstream results.

### **Paper I and II**

Whole metagenome sequencing (WMS) (or shotgun metagenomics) studies typically involve multiple experimental and computational steps, after the initial study design: (i) collection and storage of samples, (ii) DNA extraction, (iii) library preparation and sequencing, (iv) computational preprocessing of the reads, (v) sequence analysis to profile taxonomic and resistome features of the microbiome and (vi) statistical and biological post-processing analysis. Several approaches exist for each of these steps, and the choice of a specific approach can significantly influence the end results. This presents a considerable challenge to the reproducibility of metagenomic research, primarily due to the absence of universally accepted standardized methods and approaches <sup>112</sup>.

### **Study designs**

In the thesis, the suitable research design for both the experimental studies (I & II) was identified by keeping the main research question (i.e., window into the consequences of antibiotic exposure on the human resistome and microbiome) and ethical considerations in mind. The MODIC study leading to Paper I is an interventional study (part of a larger double-blind, placebo-controlled, randomized, multicenter trial (The AIM study<sup>233</sup>) using longitudinal data generated from shotgun metagenome sequencing of human fecal samples. Such clinical trials are considered as the gold standard and provide the most powerful evidence of causal relationships between interventions



(antibiotics) and outcomes (microbiome and resistome) <sup>244</sup>. On the other hand, the study resulting in Paper II is a prospective cohort study using clinical metadata and predominantly shotgun metagenomics data from preterm infants nasopharyngeal samples.

Longitudinal sampling from the same individual was used in both the clinical studies, owing to the fact that the human microbiome is not only highly individualized but can also be highly dynamic within the same individual <sup>245,246</sup>. These studies are often recommended to estimate and account for such intra and inter-individual variability in order to draw robust interpretations from the data. Moreover, samples taken from the same individual over time can also serve as better and additional control, particularly when it is impossible to have a separate control group free from other confounding variables for comparison. The longitudinal changes in these samples can then be correlated with associated metadata <sup>112</sup>.

Similar to microbiome, the human resistome is also known to be influenced by several factors such as host genetics, age, sex, antibiotics, diet, mode of birth, lifestyle variables, environmental surroundings, etc. These known and other unknown factors may influence the outcome between the groups, independent of the primary variable (antibiotics) being studied. In study I, the subjects were completely randomized to the antibiotics/control groups to minimize such potential confounding factors. In both the clinical studies (especially in study II where randomized treatment allocation and matching the control group is practically impossible), as much metadata (clinical covariates) about each of the treatment groups was collected as possible. The plan was to statistically adjust for confounding factors (if present) when comparing groups during the analysis.

### **Sample collection and storage**

Procedures for sample collection and preservation can impact the quality of DNA and the accuracy of metagenomics data <sup>247</sup>. Collection and storage methods were chosen according to different sample types (feces vs nasopharynx) in our two clinical studies, as methods that have been corroborated for one sample type cannot be expected to be ideal for other sample types <sup>112</sup>. Though, methods were consistent for all the samples from different time points within a given study. In Paper II, a suction device was used to collect nasopharyngeal aspirate samples from preterm infants. Although obtaining an aspirate is more invasive than practical alternatives such as

nasopharyngeal swabs, this method was used to minimize contamination with the microbiome of the anterior nostrils<sup>248</sup>. Prior to collection, standard operating procedures (SOPs) were established for obtaining aspirate samples from all preterm infants<sup>63</sup>. To further minimize potential contamination, standard protective equipment was used, and the number of people involved in the sample collection was kept to a minimum, as these samples are more prone to contamination due to low microbial biomass<sup>249</sup>. After collection, the nasopharyngeal aspirate samples were stored in a sterile tube containing glycerol and rapidly frozen for long-term storage at -80°C<sup>250</sup>.

To characterize the gut resistome of adult patients in Paper I, fecal samples were collected in sterile containers without preservatives and directly frozen at -80°C until further processing. It is critical to be aware that different collection and storage methods can influence the results, and in some conditions, the effect size of these steps can be larger than the effect size of the biological variables<sup>251</sup>.

## **DNA extraction**

DNA extraction methods can greatly influence the composition of the resulting sequencing data, as reported in numerous metagenomic studies<sup>252-254</sup>. Ideally, the optimal DNA extraction method would be able to extract DNA from all cell types present within a sample, ensuring their accurate representation in the subsequent sequence data<sup>112</sup>. Most of the available extraction methods generally rely on mechanical (bead-beating), chemical (or enzymatic) or a combination of both, with each method impacting the DNA yield, purity, integrity and degradation differently. Methods combining mechanical lysis (bead chemical and beating) and chemical lysis are considered superior for extracting DNA from fecal samples<sup>255</sup>. However, due to differences in efficiency and accuracy of different DNA extraction methods, in addition to immense variation present within the same type of samples, it is very challenging to find an extraction protocol that will work best for all sample types without bias. In Paper I, the Stratec® PSP Spin Stool DNA Kit, one of the most commonly used commercial DNA extraction kits, was used for fecal samples, consistent with several other studies<sup>256</sup>.

Nasopharyngeal aspirates collected in paper II are low-microbial biomass and high host (human) DNA samples. These samples are more vulnerable to biases or false positives due to contamination

during the processing stages <sup>249,257-259</sup>. To ensure that sufficient microbial biomass is extracted from nasopharyngeal aspirate samples of preterm infants for sequencing and comprehensive resistome profiling, a prior method optimization study was conducted and published (Paper IV) <sup>63</sup>. In this study, the efficiency of three commercially available DNA extraction protocols was compared for processing nasopharyngeal aspirate samples from premature infants. In addition, three host DNA depletion methods were also tested to enrich the microbial cell and DNA content prior to the extraction step in such samples. Based on those results, MoLYsis™ Basic5 was used for host DNA depletion, and then the MasterPure™ Gram Positive DNA Purification Kit was used for extraction of microbial DNA from preterm infant's nasopharyngeal samples in the main study (Paper II). However, prior to DNA extraction, nasopharyngeal samples (pellets) were also spiked with ZymoBIOMICS Spike-in Control II (low microbial load), which contains three different types of bacterial strains, serving as an *in situ* positive control. To account for contamination, reagent blanks were incorporated as negative controls and processed using the same kits or reagents as the actual samples. Furthermore, the level of biomass present in samples was gauged using qPCR before sequencing.

## **Library preparation and sequencing**

Similar to previous steps in the metagenomic workflow, the choice of library preparation and sequencing platform may significantly affect the results of metagenomic sequence data <sup>260,261</sup>. The process of library preparation includes fragmentation, adaptor ligation or tagmentation followed by PCR amplification (optional). Several kits are available for constructing sequencing libraries, which generally differ in the methods used for fragmenting metagenomic DNA, such as sonication, enzymatic fragmentation, and tagmentation by transposomes. The most suitable protocol available for fecal (paper I) and nasopharyngeal (paper II) samples was recommended by the Norwegian Sequencing Center, with whom the choice of library preparation and sequencing platforms was made in collaboration. In both clinical studies, the DNA sequencing libraries of samples were prepared using the most commonly used and current gold standard Nextera DNA Flex kit from Illumina. The DNA Flex kits have been widely adopted in metagenomics investigations owing to their speed and greater flexibility for input type, amount, and a wide range of supported applications <sup>262</sup>. These kits also resolved sequencing biases detected in the previous Nextera XT kit that occur in genomic regions with high GC content <sup>263</sup>. The best available (at time of

sequencing) sequencing platform with higher throughput and accuracy (low error rates) at an affordable cost was primarily used to deeply sequence metagenomic DNA libraries in our studies. Consequently, DNA libraries were sequenced earlier on Illumina HiSeq 3000 platform in Paper I, while Illumina NovaSeq S4 platform was used in more recent study (Paper II).

### **Bioinformatics processing of sequencing reads**

Preprocessing of raw reads obtained from high-throughput NGS platforms includes running variety of quality control computational tools for identification and removal of adapter sequences, low-quality reads and host DNA contaminants. There are several preprocessing tools and pipelines available for such quality control and trimming. In our projects, FastQC, Trimmomatic (paper I), and Cutadapt (paper II) were used, as they are the most commonly recommended and efficient preprocessing tools for this purpose<sup>20,110</sup>. Host DNA contamination is an important challenge in human-derived metagenomic samples, especially in nasopharyngeal samples. The proportion of human-to-microbial DNA in metagenomics data varies widely by body site. For instance, data from the Human Microbiome Project (HMP) demonstrated that more than 90% of the sequenced reads were aligned to the human genome in nasal cavity samples compared to <10% human DNA contamination in fecal samples from healthy individuals<sup>264</sup>. In our study II, a host DNA depletion kit (MoLYsis™ Basic5) was also utilized prior to DNA extraction for depletion of human reads from the nasopharyngeal samples. However, due to loss of microbial DNA and potential bias toward Gram-positive bacteria, such pre-extraction host DNA depletion techniques are not commonly used in most metagenomic studies. Nonetheless, it is common practice to computationally remove sequence reads containing human DNA before further analyzing the remaining (microbial) sequences from human-derived metagenomic data. There are numerous tools developed based on two main approaches: subtractive alignment and direct classification. In both of our clinical studies, human DNA were removed by aligning all the sequenced reads against a human reference genome (GRCh38) using one of the most commonly used, ultrafast, memory-efficient and highly precise alignment tool called Bowtie 2.

### **Sequence analysis for resistome and microbiome profiling**

There are two main methods when analyzing metagenomic sequencing data, read-based or assembly-based analysis, and the choice may affect the downstream results<sup>265</sup>. In papers I and II, read-based analysis was conducted in which high-quality clean reads were directly mapped to

reference databases using the pairwise alignment tool Bowtie 2 for identification and quantification of ARGs in metagenomic samples. The read-based approach is generally faster, requires less computational resources, and more importantly, enables the identification of potential ARGs from low-abundant organisms present in complex communities, which may remain undetected by assembly-based methods due to poor or incomplete assemblies. Importantly, this approach is more accurate, efficient and recommended for samples derived from previously well-characterized metagenomes/microbiomes with updated databases, as compared to samples obtained from more diverse and understudied environments<sup>18,112,200</sup>. This approach aligns with our aim of investigating and exploring the impact of antibiotics on the entire landscape of known ARGs (the resistome) present in the better characterized and well-studied human microbiome.

However, the accuracy and completeness of resistome profiles is mainly dependent on the reference ARG database. Several large, comprehensive and updated reference databases for annotation of sequences to ARGs are available. In both papers (I & II), the CARD database was utilized, which is the most comprehensive, commonly used, continuously updated, and rigorously curated resource, currently containing over 5000 reference ARGs. Studies have shown that CARD is often the first choice or the most preferred option for the *in-silico* identification of ARGs across metagenomes<sup>20,176</sup>.

In these studies, Metagenomic Phylogenetic Analysis (MetaPhlAn3) was used for taxonomic profiling, a tool also utilized in the HMP. This method employs a marker-gene approach, with approximately 1.1 million unique clade-specific marker genes identified from around 17,000 reference microbial genomes, to assign the reads to their appropriate taxonomic groups with species-level resolution<sup>266</sup>. This method is faster and less computationally intensive than most other approaches because the reference database only contains a small subset of the genomes (marker genes) rather than full genomes for all the species. In addition, MetaPhlAn relies on Bowtie 2 aligner to make the assignment more accurate and faster. MetaPhlAn was primarily chosen for these studies because it has a higher precision (very low false positive discovery rate) compared to other methods, as demonstrated by several benchmarking studies<sup>265-267</sup>. This is

important because false positive classifications can be a significant challenge when interpreting metagenomic data, particularly when analyzing human clinical samples<sup>268</sup>.

Selecting the most appropriate approach for analyzing WMS data is a challenging task. At present, there is no consensus on which sequence analysis approach is the best, and the choice of analysis should be ultimately determined by the aim of the study, costs and technical considerations<sup>112,184,269</sup>. However, unlike the initial steps of the WMS study workflow, the choice of methods in this step can be easily reevaluated using other methods. This extensive scope for further investigations renders WMS a valuable asset in the field of resistome and microbiome research.

### **Statistical analysis**

In papers I and II, several robust statistical testing methods were used to analyze the resistome and microbiome abundance data, based on the type of analysis or research questions to be addressed. All statistical analyses were performed in R using several CRAN and Bioconductor packages. For simple, independent comparisons of group differences, the one-way ANOVA, Wilcoxon rank-sum test, Kruskal-Wallis test, or chi-square test were used as appropriate. Due to longitudinal study designs, a linear mixed-effect model or repeated-measures ANOVA was used to assess the difference in  $\alpha$ -diversity and total AMR abundance over time within the groups. Additionally, in these models, individual subjects were incorporated as a random effect to account for interindividual variation (between-sample correlation). Further, adjustment for confounding variables, such as the age of developing preterm infants in paper II, was done to discern the effects of the primary variable more accurately and minimize the risk of false positives. PERMANOVA, a robust and flexible method, was used to test the difference between groups based on measures of  $\beta$ -diversity. This method also allowed us to adjust for covariates or confounding factors, as in paper II. However, PERMANOVA can also generate false positive results when dealing with unbalanced samples exhibiting heterogeneous dispersion across groups. Consequently, the homogeneity of multivariate dispersion between treatment groups and time points was assessed using the PERMDISP test. In paper II, due to significant heterogeneity in dispersions, more omnibus tests, such as Analysis of Similarities (ANOSIM) and Mantel tests (for continuous variables), were applied. These tests, which are particularly sensitive to heterogeneity in dispersion, helped to ensure the validity of the results. Additionally, inferential analysis was

performed only on  $\alpha$  and  $\beta$ -diversity measures that are less sensitive to the variation in sequencing depth (in paper II).

To evaluate the significance of Procrustes results (statistics) in both our studies, another permutation-based test using the *protest* function from the vegan R package was performed. Most of these permutation-based tests are also supported in ResistoXplorer (paper III) to enable users to conduct robust inferential analysis on their resistome abundance data. In paper I, differentially abundant ARGs and taxa between time points within treatment groups were identified using one of the most commonly employed Linear Discriminant Analysis Effect Size (LEfSe) algorithm. The Benjamini-Hochberg false discovery rate (FDR) procedure was used for corrections for multiple testing, which is essential for controlling the rate of false positives (Type I errors) when conducting multiple comparisons<sup>270</sup>.

## **Ethics**

As the studies presented in the thesis involved human microbiome research, our three major ethical considerations were: informed consent, confidentiality and anonymity of subjects and data sharing<sup>271</sup>.

The Regional Committees for Medical and Health Research Ethics (REK) South East Norway (2014/158/REK sør-øst) and the Norwegian Medicines Agency (SLV, reference number 14/01368-11, EudraCT Number: 2013-004505-14) approved the AIM study leading to Paper I. The AIM study was registered at <https://clinicaltrials.gov/> (ClinicalTrials.gov Identifier: NCT02323412). Written informed consent was obtained from all the subjects.

The Norwegian Regional Committee for Medical and Health Research Ethics (REK) South East Norway (Approval number: 2018/1381 REKD) approved the Born in the Twilight study resulting in Paper II. Patients received written information about this retrospective study. Written informed consent was obtained from the parents of all the preterm infants enrolled in the study.

To maintain the confidentiality and anonymity of the subjects involved in Study I and II, all their personal metadata is stored inside a secured platform (TSD, Services for Sensitive Data, University

of Oslo, Oslo, Norway). Due to the sensitive nature of human metagenomic data, computational pre-processing of sequencing data (until the removal of human contamination from the metagenomes) was conducted inside the same secured platform.

To follow the FAIR (Findable, Accessible, Interoperable and Reusable) principles<sup>272</sup>, the metadata (after proper anonymization) and clean metagenomic sequencing data (after removal of human reads) from the published paper I were made publicly available at NCBI SRA under BioProject ID: PRJNA894204. The same principles will be applied to Paper II upon its publication.

### **Paper III**

ResistoXplorer (Paper III) was developed as a comprehensive bioinformatics tool suite for the downstream analysis of resistome data from metagenomics studies. It takes into account the exploratory nature of data analyses and aims to support reliable and robust analyses while ensuring accurate and meaningful interpretation of resistome datasets. Several key steps that were considered during the development of ResistoXplorer to facilitate these objectives are presented below.

#### **Implementation and design**

ResistoXplorer is a freely accessible, web-based, comprehensive visual analytics tool suite that was developed using three open-source programming languages: Java, R, and JavaScript. The user interface of ResistoXplorer was designed using the Java Server Faces (JSF) framework and the PrimeFaces and BootsFaces component libraries, creating an intuitive and user-friendly experience. All the back-end computation, including data processing, analysis, and visualization, was performed in R (version: 4.1.0)<sup>273</sup> using more than 20 packages from the CRAN and Bioconductor projects, along with some built-in functions. JavaScript libraries such as sigma.js, CanvasXpress.js, and D3.js were also utilized to improve the user experience by generating interactive and dynamic visualizations of complex datasets on web pages. Through its modern interface design and high-performance implementation, ResistoXplorer offers a real-time visual analytics experience that allows researchers to intuitively go over the complex tasks of resistome data processing, analysis and interpretation.



The three supported modules share the same overall workflow, starting with data processing, followed by data analysis and visual exploration. In the data processing stage, the user's data are uploaded for filtering and normalization. Following this, a wide suite of standard and advanced statistical and visualization methods can be conducted on the processed data to identify, for example, overall trends and patterns, significant features, potential biomarkers, and functional insights.

## **Data upload and processing**

Users can upload a list of ARG and feature (ARG/taxa) abundance tables as input to ResistoXplorer. Several powerful and commonly used bioinformatics pipelines for metagenomic data, such as DeepARG, ARG-OAPs, AMRPlusPlus, Kraken<sup>274</sup>, and MetaPhlan3, can preprocess raw metagenomic sequencing reads. The results from these pipelines can be summarized into feature abundance tables, which are compatible with and can be easily uploaded into ResistoXplorer for comprehensive downstream analysis.

ResistoXplorer carries out two essential processing steps, namely data filtration and normalization, on uploaded metagenomic (resistome or microbiome) abundance data. These steps are crucial for removing overall systemic variability, artifacts, and noise, thus ensuring the reliability and accuracy of downstream analyses. ResistoXplorer supports a variety of typical filtering options to mainly exclude features (ARGs) having a low mean abundance or present in only a few samples, as well as samples that exhibit low variation across samples. Such features are less likely to be informative for downstream analysis, particularly comparative analyses, which aim at detecting differences between experimental conditions or groups. Filtering such features can improve the statistical power by reducing the need for stringent multiplicity corrections in differential abundance analysis<sup>275</sup>. However, some analyses, such as alpha diversity, are sensitive to data filtering. The removal of features could potentially introduce biases, thereby affecting the interpretation of such results. To avoid this, the ResistoXplorer provides the option to conduct such analysis on unfiltered data.

Normalization is another important step. An appropriate normalization method must be employed to ensure meaningful comparisons across samples, adjusting for technical and biological

variability that can arise during the processes of sample collection, DNA extraction, sequencing, and upstream data preprocessing. To account for systemic variability and the challenges associated with count data, ResistoXplorer incorporates several widely utilized and standard normalization methods. These can be broadly categorized into three main groups: data rarefaction, scaling, and transformation-based methods. Multiple methods are supported in the tool as there is no consensus on one best method that works for all datasets with different characteristics as well as for all types of downstream analysis. In ResistoXplorer, users can evaluate the impacts of various normalization methods on resistome composition. They can do this by visually exploring the potential clustering patterns or data structures with regard to experimental factors or covariates of interest using ordination plots, dendrograms, and heatmaps.

## **Data analysis and visual exploration**

ResistoXplorer provides a flexible and comprehensive framework that supports the utilization of multiple descriptive and inferential analysis methods for conducting an in-depth exploratory analysis of resistome data. Given the absence of a universally agreed-upon analysis approach or one-size-fits-all method for all types of datasets, ResistoXplorer offers a variety of methods and approaches to empower researchers in selecting the most suitable ones that align with the specific characteristics of their dataset and research question. This flexibility enhances the reliability and relevance of the analysis results, allowing for a more comprehensive exploration of the resistome data and facilitating robust interpretations. For example, the ARG Table module in ResistoXplorer currently offers more than 15 carefully selected analysis methods based on the current best practices in the field. These methods enable users to perform comprehensive visualization, statistical analysis, and exploratory analysis of their uploaded resistome abundance data.

The accompanying web interface for each analysis method enables users to adjust and choose from different key parameters and measures for interactive analysis and visual exploration of the results. For example, ResistoXplorer supports >10 commonly used alpha and beta diversity measures, providing users with the flexibility to explore different aspects of resistome structure based on their research questions. The default options are selected based on their wide acceptance and best performance in benchmarking studies, whenever available. This allows users to start their analysis

with established and validated methods, while still having the option to explore other approaches based on their specific needs and preferences.

ResistoXplorer provides a range of widely used and robust statistical approaches for inference analysis. In the context of differential abundance testing, ResistoXplorer supports both standard and advanced methods, including DESeq2, edgeR, metagenomeSeq, LefSe, ALDEx2, and ANCOM. As different methods can yield divergent results and p-values, ResistoXplorer provides users with an optimal solution by allowing them to employ a consensus-based approach. This approach utilizes multiple DAA methods to validate the results, enhancing the confidence and reliability of the interpretation. To ensure the integrity of the analysis and prevent p-hacking, ResistoXplorer restricts each DAA method to use its own default normalization. This is due to the lack of benchmark studies examining the performance of different combinations of normalization approaches with different DAA methods. Additionally, since differential analysis involves testing hundreds or thousands of features simultaneously, ResistoXplorer always applies multiplicity correction for all supported methods. This correction helps maintain the robustness of the findings by reducing the risk of Type I errors (false positives).

However, it is important to acknowledge that different statistical methods and algorithms have their own strengths and limitations. In ResistoXplorer, the default option for testing the difference between groups based on beta-diversity measures is the robust and powerful multivariate statistical method, PERMANOVA. While PERMANOVA is sensitive and recommended, it may produce false positive results in unbalanced samples that exhibit heterogeneity in dispersion across groups. To address this issue, ResistoXplorer provides alternative tests, such as ANOSIM, which are less sensitive (in general) and can enhance the validity of findings in such scenarios. Users can choose the appropriate test based on the specific characteristics of their dataset to obtain accurate and reliable results. In the statistical methods supported by ResistoXplorer for inferential analysis, the reporting goes beyond p-values. The tool also provides the effect size, which represents the likely importance of the findings. This approach enables users to evaluate the potential biological significance of their results with more reliability and interpret their findings in a more robust manner. For example, in the LefSe algorithm, ResistoXplorer provides the LDA score for differentially abundant genes, which indicates their discriminative power. In

PERMANOVA analysis, the tool offers the  $R^2$  value, which represents the proportion of variance in the overall resistome composition explained by the experimental factor or covariate of interest. These additional metrics enhance the interpretability and meaningfulness of the results obtained from ResistoXplorer.

As a visual analytics tool, ResistoXplorer incorporates robust visualization capabilities that empower users to visually explore and analyze their data. It generates a variety of compact visualizations and graphics such as charts, graphs, plots, tables, diagrams, heatmaps and dendrograms to aid users in interpreting the results of all downstream, including both descriptive and inferential analyses. Moreover, ResistoXplorer supports a range of robust and interactive visualization techniques for an in-depth visual examination of complex resistome data structures and their associations, such as Sankey diagrams, zoomable sunbursts, treemaps, and networks. This interactive approach allows users to intuitively reveal patterns, trends, and draw insights into the potential biological implications of the data.

Further, it is important to note that all statistical, exploratory, and visualization analyses of ARG profiles can be conducted at higher levels to gain functional insights, facilitating a better understanding of their biological significance. To streamline this process, ResistoXplorer offers comprehensive support by gathering and pre-compiling the functional annotation information of ARGs from nine commonly used reference AMR databases. This eliminates the traditionally tedious, time-consuming, and error-prone process of manually collecting annotation information from individual databases.

### **Reproducibility, documentation and support**

ResistoXplorer is freely accessible at <https://www.resistoxplorer.no>. ResistoXplorer thrives for transparency and reproducibility of resistome data analysis by providing the underlying R scripts, which can be found at: <https://github.com/FCPLab007/ResistoXplorerR>. Further, it also records all the steps that users have taken and provides its R command history to reproduce and validate their results locally. Further, the information on the version numbers of all software used, including all R packages, are mentioned in the ‘About’ section, and users can download any underlying data that ResistoXplorer used, such as functional annotation information of supported ARG databases from ‘Downloads’ section.

ResistoXplorer provides extensive documentation, including user manuals, FAQs, tooltips, and example datasets, to ensure users understand the tool's functionalities, limitations, and result interpretation. Detailed instructions are available on dedicated web pages for easy navigation. Furthermore, ResistoXplorer offers a user forum through Google Group, allowing users to seek assistance, share insights, and engage in collaborative discussions with other users. An introduction to ResistoXplorer is also featured in a one-week training module within a massive open online course (MOOC) entitled “Exploring the Landscape of Antibiotics in Microbiomes” ([www.futurelearn.com/courses/exploring-the-landscape-of-antibiotic-resistance-in-microbiomes](http://www.futurelearn.com/courses/exploring-the-landscape-of-antibiotic-resistance-in-microbiomes)).

## Discussion

Metagenomics approaches based on high-throughput NGS technologies are powerful and increasingly used methods for comprehensively investigating the distribution and dynamics of AMR in microbiomes. They enable unbiased cataloging and profiling of the entire landscape of ARGs in host and environmental-associated microbiomes<sup>163</sup>. In this thesis, such approaches were used to address knowledge limitations on the resistome composition and development in humans, including the ecological consequences of antibiotic therapy (Papers I and II). Further, a bioinformatics-based platform was developed to facilitate overall metagenomic resistome data analysis and interpretability (Paper III).

Human microbiomes exhibit significant plasticity, with ecological changes driven by various factors, including diet, lifestyle, medications, environmental exposures, and host genetics<sup>245,276,277</sup>. Among these, antibiotic usage is one of the most pervasive factors due to the potential to select for antibiotic resistance<sup>42,278,279</sup>. The most studied human microbiome is the one in the gut, for which advancements in NGS technologies and methodological developments are starting to capture the immediate changes caused by short-term antibiotic exposures<sup>73,74,280</sup>. However, there is a lack of evidence-based data regarding the long-term ecological consequences of prolonged antibiotic regimens on the microbiome and resistome. Other studies for which key information is still needed are those focusing on body sites other than the gut, in particular those of critical importance in relation to infections, such as the respiratory tract.

In paper I, the issue of the long-term consequences of prolonged antibiotic therapy was addressed. This was achieved by utilizing deep shotgun sequencing and functional metagenomics to analyze the microbiome and resistome in the gut of adults before, immediately after, and 9 months following a prolonged (3 months) amoxicillin treatment. The findings revealed that prolonged amoxicillin therapy leads to severe perturbations in the human gut microbiome, with a significant reduction in bacterial diversity, changes in overall community composition, and decrease in abundances of beneficial bacteria. These results contrasted with previous sequencing-based studies, which have demonstrated that the adult human gut microbiome generally exhibits resilience towards short-term amoxicillin intervention, with minimal or no impact on its microbial

composition<sup>230,281</sup>. A significant depletion of short-chain fatty acid producers was observed, in particular butyrate, immediately after the cessation of amoxicillin treatment, despite high individual variation in baseline microbiome composition among patients. These butyrate producers play a crucial role in maintaining gut health by sustaining homeostasis, protecting the intestinal epithelium from inflammation, inhibiting pathogen proliferation, and promoting pathogen clearance<sup>282,283</sup>. Conversely, *Clostridium bolteae*, a gastrointestinal opportunistic pathogen typically associated with infections in immunocompromised individuals or those with underlying health conditions<sup>284</sup>, was found to be significantly enriched in most of our patients following prolonged amoxicillin treatment. Our study corroborates previous findings that *C. bolteae* can serve as a potential marker of beta-lactam or antibiotic-induced dysbiosis<sup>280,285</sup>. Additionally, genera *Bacteroides*, well-known for harboring beta-lactamase ARGs, also systematically increased immediately after the cessation of amoxicillin treatment<sup>39,286,287</sup>. However, the effects of prolonged amoxicillin treatment seemed short-lived, as no such changes in the microbiome composition were observed 9 months post-treatment, suggesting the long-term resilience and stability of the adult gut microbiome to such perturbations. Despite their transient nature, these instances of microbial dysbiosis could potentially lead to significant long-term health consequences, including an increased susceptibility to recurrent, pathogenic, and potentially untreatable antibiotic-resistant infections such as *Clostridioides difficile*<sup>288</sup>.

Compared to the microbiome, the results in Paper I indicate that prolonged amoxicillin exposure has a long-lasting impact on the human gut resistome, leading to a more diverse and enriched resistome long after antibiotic use termination. In particular, the total abundance and diversity of ARGs significantly increased immediately after the amoxicillin treatment (3 months), which remained significantly higher even after 9 months post-treatment. A previous study by Zaura *et al.* reported a substantial enrichment of various classes of ARGs in the human adult gut resistome following short-term amoxicillin therapy<sup>74</sup>. However, the majority of their results were derived from predictive analysis, with only a single sample being analyzed through WMS. In contrast, the study performed in Paper I demonstrated that prolonged amoxicillin intervention leads to an enrichment of the resistome in ARGs that confer resistance directly against the antibiotic class used, specifically beta-lactams. This includes instances of enrichment of clinically relevant beta-lactamase ARGs associated with an extended spectrum of activity (ESBL) and those active against

carbapenems, considered last-resort antibiotics. Though the increase in abundances or emergence of beta-lactamase ARGs observed were highly individualized, the potential relevance of personalized antimicrobial therapies. Overall, these findings underscore the potential risks associated with prolonged antibiotic exposure in selecting and promoting the development of AMR in human gut microbiome. Given the persistent changes observed in the resistome, in contrast to the transient alterations in the microbiome, it is plausible that exposure to amoxicillin could independently facilitate the selection and spread of ARGs through mechanisms such as horizontal gene transfer (HGT). However, due to the inherent limitations of the short-read metagenomic sequencing and read-based profiling approaches used in our study, detailed information on the genetic context of ARGs is lacking. This restricts the ability to discern whether an ARG is present on a MGE - a factor that significantly influences the potential for HGT - and consequently hinders the understanding of how these ARGs might spread within gut microbial communities. On the other hand, no significant changes were observed in the diversity and composition of the microbiome and resistome over time in patients who did not receive long-term amoxicillin treatment, thus confirming the stability and resilience of the healthy adult human gut microbiome without significant perturbations <sup>289,290</sup>.

In Paper II, the focus was on the impact of antibiotics on the resistome development of preterm infants using samples from the nasopharynx, a site of particular importance in mediating susceptibility to respiratory tract infections <sup>291</sup>. In contrast with the gut, for which establishment and development of the infant microbiome and resistome has been studied to a certain extent <sup>45,51-53,55,56,59</sup>, little is known about the diversity and dynamics of antibiotic resistance determinants in the microbial communities of the nasopharynx. The resistome and its dynamics in the nasopharynx were explored by characterizing them using deep shotgun metagenomic sequencing. Further, in accordance with primary objective, the long-term effects of early antibiotic therapy (ampicillin + gentamicin) on the resistome were also assessed by analyzing nasopharyngeal samples collected from birth until six months corrected age.

The findings in this paper indicate that early broad-spectrum antibiotic treatment with a combination of ampicillin and gentamicin can lead to acute perturbations in the developmental trajectory of the nasopharyngeal resistome in preterm infants. A significant enrichment of the



resistome was observed, characterized by an increase in the total abundance and diversity of ARGs directly after the cessation of early antibiotic treatment. This led to minor yet significant shifts in the overall resistome composition of the early antibiotic-treated infants compared to the naive infants. A few potentially high-risk and clinically significant ARGs, encoding for aminoglycoside-modifying enzymes, extended-spectrum beta-lactamases (ESBLs) activity, and those conferring resistance against carbapenems (such as the presence of ACT-beta lactamase), also appeared in some infants immediately after the cessation of early antibiotic treatment. These compositional changes in the resistome upon early antibiotic treatment were more profound in infants whose mothers had also received prenatal antibiotics during pregnancy. However, no significant changes in diversity, composition, or carriage persistence of high-risk ARGs in treated preterm infants compared to controls at 6 months corrected age were found. This suggests the effect of early antibiotic treatment on the nasopharyngeal resistome was short-term. Interestingly, this transient nature is in line with a recent study by Reyman *et al.*, who also reported short-term effects on the pediatric intestinal resistome following the administration of various early antibiotic regimens (penicillin + gentamicin, co-amoxiclav + gentamicin, and amoxicillin + cefotaxime) in the first week of life <sup>70</sup>. Despite their transient nature, any adverse changes in the resistome during early life in such a vulnerable population are particularly concerning, especially those that favor the selection of antibiotic-resistant pathogens, commonly found in the nasopharynx, that have the potential to spread and cause invasive infections.

To the best of my knowledge, paper II is the first study to perform a comprehensive characterization of nasopharyngeal resistome development in neonates using a shotgun metagenomics approach. To date, one of the most comprehensive attempts to investigate ARGs in the nasopharynx of infants using a shotgun sequencing approach identified ARGs in 64% of the samples, but the samples were enriched for streptococci prior to DNA extraction, thus presenting biased information <sup>292</sup>. The main challenge with characterizing the nasopharyngeal resistome is the low microbial biomass and high content of human host DNA <sup>60</sup>. Prior to the main study in paper II, an optimization protocol for DNA extraction and library preparation was developed, with a key emphasis on depleting human DNA and recovering maximum amounts of microbial DNA <sup>63</sup>. Using this protocol, ARGs were identified in almost all (~95%) of the nasopharyngeal samples in the main cohort. Our results suggest that the nasopharynx of preterm infants harbors a rich and diverse resistome, which undergoes dynamic changes with age during

the first six months of life and can be negatively affected by prenatal and postnatal antibiotic exposure. The majority of the variation observed in the resistome composition of preterm infants was attributed to inter-individual differences, thus highlighting the importance of longitudinal time-series studies. Considering the disproportionate contribution of respiratory pathogens that colonize the respiratory tract (including the nasopharynx) to AMR-associated deaths <sup>15</sup>, our study could serve as a baseline for future research focusing on this relatively understudied yet critical reservoir of ARGs and pathogens.

There is a swift increase in awareness regarding the impact of antibiotics on emergence and evolution of AMR in pathogenic bacteria, the disruption of microbial communities (dysbiosis), and potential health-related risks at the population level. Currently, most hospitals enforce antimicrobial stewardship programs (ASPs) to optimize the usage of antibiotics, reducing the development of AMR and improving patient outcomes. However, these programs have not yet incorporated considerations of these ecological side effects of antibiotic therapies, primarily due to the absence of such information <sup>70</sup>. This thesis provides evidence-based data on these ecological consequences, highlighting their relevance for antimicrobial stewardship practices. Additionally, our results underscore the importance of weighing these side effects against the potential clinical benefits of antibiotic therapies. For instance, in Paper I, adult subjects from a multicenter, randomized, double-blind, placebo-controlled trial in Norway, known as the AIM study, were analyzed. This study found no significant clinical benefits of prolonged amoxicillin therapy in patients suffering from chronic low back pain and Modic changes. Our findings reinforce the lack of support for prolonged antibiotic therapy for these conditions and also stress the need to consider such risks in relation to other conditions that receive minor or no clinical benefits from prolonged antibiotic therapy. On the other hand, the cohort in our study II, i.e., preterm infants, being especially vulnerable, are frequently exposed to antibiotic interventions within the first few days or weeks following birth due to the potential risk of early-onset neonatal sepsis. In Norway, approximately 75% of preterm infants (<32 weeks of gestation) receive empirical antibiotics within 72 hours after birth due to suspicions of early-onset neonatal sepsis <sup>239</sup>. However, only 1 in 100 is estimated to develop a confirmed infection, suggesting an unnecessary over-prescription of antibiotics in preterm infants. Emerging research suggests that early antibiotic-induced perturbations to the microbiome and resistome, particularly during this critical developmental

phase, may have far-reaching consequences such as asthma development <sup>55,240,293,294</sup>. Additionally, early and prolonged antibiotic exposure in very preterm infants has been associated with adverse outcomes such as bronchopulmonary dysplasia <sup>295</sup> and late-onset sepsis <sup>296</sup>, with conflicting results regarding necrotizing enterocolitis <sup>297-299</sup> and mortality <sup>296</sup>. Therefore, the risk-benefit balance of antibiotic use in this vulnerable population necessitates more careful consideration. In general, the development and enrichment of the resistome may cause treatment failure, the spread of AMR, limited therapeutic options, and increased healthcare costs.

The microbiome in humans is complex and diverse, exhibiting distinct types and abundances of microbes and ARGs according to the different body sites and maturity stage <sup>49</sup>. Despite differences in body sites sampled, baseline microbiome composition, antibiotic dosage and duration, maturity of the microbiome (developing vs stable), diet, environment, and the co-morbidities of the cohorts in Papers I and II, the results highlight the detrimental ecological consequences of antibiotics on the human resistome. Also, in both studies, it was found that the resistomes are structured by the composition of the microbiome and that they exhibit considerable variation among individuals. Adults exhibited greater inter-individual variations in the resistome, which is expected due to the stability, complexity, and diversity of the adult gut microbiome (Paper I) compared to the developing microbiome in the nasopharynx of preterm infants (Paper II) ( $R^2$ : 79% vs 30%). This is in line with studies from gut microbiomes showing individualized resistome responses to antibiotic therapies in both infants and adults <sup>67,121,280,300</sup>. Future studies will benefit from characterizing the respiratory resistome and the impact of antibiotics in other age groups, including term infants, adults, and older adults.

Driven by the growing number of metagenomic-based investigations across different reservoirs of ARGs - including humans, animals, and environmental sources - there is a strong demand for an easy-to-use and intuitive bioinformatics platform. Such a platform would enable researchers to easily understand and interpret the large and complex datasets generated from these studies <sup>269</sup>. At present, most of these studies and their data analysis are mainly exploratory in nature. It is recommended to use multiple methods and tools for analysis and visualization rather than relying on a single method <sup>184</sup>. To address these needs, a user-friendly, web-based tool called ResistoXplorer was developed and launched, detailed in Paper III of the thesis. This tool assists

researchers in performing visual, statistical, functional, and exploratory analysis of resistome data without requiring prior programming expertise. Users can choose from a wide array of well-established methods, as well as more advanced compositionally appropriate (CoDa) methods, and explore results in real-time to gain a better and more comprehensive understanding of their data. ResistoXplorer also enables data and knowledge-driven integrative analysis of resistome and microbiome datasets using a variety of sophisticated statistics-based methods and a powerful network-based visual analytics system. Such a tool suite facilitates more complex analyses, typically restricted to experienced bioinformaticians, leading to the democratization of resistome data analysis. Over the past two years, the web server has processed over 13,000 data analysis jobs submitted by more than 3,000 users worldwide. The current features and the functional annotation information from supported ARG databases have been actively updated, as well as new functions added based on user feedback and advancements in the field.

Several functions and methods supported by ResistoXplorer were utilized for the analysis of resistome data in our clinical studies (Papers I and II). However, it is important to acknowledge that certain analyses exceeded the current capabilities of ResistoXplorer due to the complex experimental designs, including complex metadata, present in our longitudinal studies. Additionally, while our tool does support various standard and advanced statistical methods, it is worth noting that these methods may not be entirely suitable for analyzing longitudinal (time-series) data due to specific characteristics associated with such data. These characteristics comprise a low number of replicates, varying numbers of samples per subject, differing numbers of subjects per group, samples collected at inconsistent time points, and within-subject correlation resulting from repeated measures <sup>206</sup>. Consequently, in both of our longitudinal clinical studies, more sophisticated statistical methods were opted for, specifically mixed-effects models implemented in R packages. By utilizing these models, it was possible to effectively evaluate the true impact of antibiotic exposure on ARG diversity while appropriately adjusting for covariates and accounting for within-subject correlation. Looking ahead, our future plans entail further enhancing ResistoXplorer by incorporating additional novel and robust statistical methods, as well as advanced data visualization techniques tailored for the time-series (longitudinal) analysis of resistome data. In the long term, the aim is to integrate an upstream sequence analysis pipeline that directly links to downstream analysis, enabling raw sequencing data processing and ARG

annotation. The goal is that ResistoXplorer will become a one-stop-shop for high-throughput bioinformatics analysis and interpretation of metagenomics resistome data. For now, it is essential to acknowledge that the success and validity of the results obtained from any downstream analysis, including those performed in ResistoXplorer, are significantly influenced by the comprehensiveness and quality of available ARG databases, upstream bioinformatics tools and analysis parameters<sup>18,20,111</sup>.

In the methodological considerations section, possible biases introduced from sampling to sequencing steps were discussed. Other steps in the bioinformatic analysis of resistome data deserve further consideration, as they introduce a certain degree of uncertainty (Papers I and II). For instance, our analysis relied primarily on homology-based alignment of sequences with known ARGs in a reference database. This approach has inherent limitations, such as potentially missing undiscovered ARGs or those not present in the reference ARG database. Read-based profiling of short reads can be prone to incorrect mapping based on sequence homology and may lead to false-positive identification of ARGs<sup>301</sup>. However, such an approach provides higher sensitivity for identifying low-abundance ARGs compared to assembly-based approaches, which are computationally intensive and time-consuming<sup>18,20</sup>. Additionally, the assembly of nasopharyngeal samples (as in paper II) remains a significant challenge due to their low-microbial-biomass nature, which can potentially obscure the detection of a lower number of microbial reads<sup>60</sup>. Furthermore, sequence similarity-based searches with stringent conservative coverage thresholds were used to identify ARGs from short-reads by aligning them to reference sequences. This approach was employed to minimize false positives and over-classification of ARGs. The ARG is only considered to be “present” in a sample only if at least 80% of the nucleotides in a reference ARG sequence must have at least one read perfectly aligned to it. This approach, widely used in numerous publications and computational pipelines, strikes a balance between efficiency, sensitivity, and stringency<sup>20,111,151,164,302</sup>. However, this method may result in missing actual ARGs with coverage below the defined thresholds or those sharing low similarity compared to the reference sequence, potentially affecting the number and type of ARGs detected. Alternatively, recent approaches like fARGene<sup>174</sup> and DeepARG employ machine learning to improve sensitivity and accuracy in identifying new ARGs with limited sequence similarity directly from short reads.

Additionally, in Papers I and II, our reads were annotated using one of the most comprehensive, regularly updated ARG databases, i.e., CARD. However, it is important to note that the CARD database, as any other database, is biased towards ARGs characterized in clinically associated pathogens and commensals that have been experimentally validated, limiting the ability to detect environmental and non-clinical ARGs or emerging resistance mechanisms<sup>179</sup>. For instance, in Paper I, it was observed that the resistome of the adult gut microbiome, cataloged using a sequence-unbiased approach (functional metagenomics), only shared a median identity of 33% with the reference protein sequences in the CARD and AMRFinder database. This discrepancy highlights a significant gap between the actual diversity of ARGs and the known diversity captured by the databases. Moreover, the CARD database includes certain chromosomal point mutations and presumptive ARGs, which may not reliably confer resistance based solely on short-read sequencing data and are less likely to be directly involved in clinical AMR. Consequently, to minimize the inclusion of false positives in the studies, the decision was made to exclude these ARGs from the downstream analyses<sup>29</sup>. Conversely, it is important to note that not all ARGs identified in metagenomes pose the same risk to public health, as many ARGs are ubiquitous among bacteria in various environments. For instance, efflux pumps, which are abundantly found in the nasopharyngeal microbiome of preterm infants (Paper II), are widely distributed in bacteria and serve diverse biological functions beyond conferring resistance<sup>303,304</sup>. Some studies do not strictly categorize them as ARGs, considering their primary role in homeostasis and cell-cell signaling<sup>305,306</sup>. While this is a valid approach, it has the disadvantage of overlooking an important class of ARGs.

Recent efforts have focused on evaluating the human health risks associated with different types of ARGs<sup>304,307,308</sup>. Following an omics-based risk framework proposed by Zhang *et al.*, ARGs enriched in human-associated environments, mobile (present on MGEs), and harbored by ESKAPE pathogens, have been classified as high-risk ARGs (Rank I)<sup>304</sup>. In Paper II, this information was utilized to identify potential high-risk ARGs in the nasopharynx microbiome of preterm infants during their first six months of life. This study revealed the presence of a small proportion of potential high-risk ARGs at multiple time points in the nasopharynx of almost all preterm infants, irrespective of antibiotic treatment. However, these were classified as potentially

high-risk due to the lack of detailed genomic context information, including mobility and host pathogenicity, resulting from our use of direct profiling of short-reads approach. As the field progresses, more research efforts are anticipated to concentrate on the timely and cost-effective detection of these high-risk or critical ARGs among the numerous thousands identified in large and complex metagenomes.

Metagenomic sequencing data and their analysis have the potential to contribute to antibiotic stewardship in clinical settings <sup>111,309</sup>. However, there are several limitations that currently restrict its application as a surveillance or diagnostics tool, including the lack of high-resolution information on the genetic context of ARGs and individual resistomes <sup>60,310</sup>, longer turnaround time for processing clinical samples into actionable results, and the need for significant sample handling expertise. Advancements in sequencing technologies, particularly long-read and chromosome conformation capture (Hi-C) sequencing, have the potential to revolutionize the field by enabling the analysis of ARGs in the context of their genetic surroundings, allowing for the linkage of ARGs, mobile genetic elements (MGEs), and bacterial hosts <sup>311</sup>. This capability would enhance our understanding of the spread and transmission of ARGs and the dynamics of resistomes within and between microbial communities. Recent efforts have focused on the application of long-read sequencing technologies, such as Nanopore and SMRT sequencing, for rapid and effective surveillance of ARGs and as point-of-care diagnostic tools in clinical settings <sup>309,312</sup>. For instance, studies have demonstrated the use of MinION (Oxford Nanopore Technologies) sequencing-based clinical metagenomics to accurately characterize and analyze pathogenic gut bacteria associated with sepsis or necrotizing enterocolitis (NEC), along with their corresponding resistome profiles, within just 5 hours using fecal samples <sup>313</sup>. Looking into the near future, it is highly likely that the integration of resistome data generated from long-read sequencing into AI models will greatly enhance the metagenomic surveillance of ARGs in clinical settings <sup>309</sup>. This approach will likely play a significant role in guiding antibiotic stewardship and potentially accelerate appropriate antibiotic administration (personalized medicine) in time-critical settings such as NICUs, and informing future clinical decision-making around AMR. Furthermore, the development of bioinformatics tools and pipelines, optimization of ARG databases, and standardization of methodologies are still necessary to improve the reliability of metagenomic resistome data and its analysis, thereby maximizing its potential for AMR surveillance <sup>60</sup>.



## Conclusion

The work in this thesis broadens our understanding of AMR from an ecological perspective by comprehensively profiling the landscape of ARGs (resistomes) in the human microbiome using metagenomic sequencing approaches. Several laboratory protocols, along with various bioinformatics tools, statistical and integrative analysis methods, as well as visualization approaches, were utilized in the clinical studies (Papers I and II) to investigate the impact of antibiotic therapies on human resistomes. Though Papers I and II investigate different antibiotic regimens with varying degrees of perturbation, they share a common theme: understanding how selective pressures induced by antibiotic exposure affect the emergence, transmission, and evolution of AMR from an ecological perspective. The findings underscore the need for reinforced surveillance of antibiotic therapies and the reduction of unnecessary antibiotic exposure. This knowledge is invaluable for the design of personalized prescription strategies and rational administration guidelines, and could potentially strengthen antibiotic stewardship programs. In addition to providing a common framework for understanding the collateral impact of antibiotics, these papers utilize whole shotgun metagenomics analysis to elucidate the adverse ecological effects of antibiotic therapies on the human microbiome and resistome, including dysbiosis and the emergence and development of AMR. This thesis provides, for the first time, a comprehensive characterization of the development of the nasopharyngeal resistome in infants, an understudied reservoir for the emergence and spread of ARGs.

Interpreting and understanding the large and complex datasets generated from any metagenomic-based resistome studies remains a key bottleneck, requiring substantial computational tools and bioinformatics expertise. In the work of this thesis, this gap was addressed by introducing the development of a user-friendly, web-based bioinformatics tool, ResistoXplorer. This tool allows researchers to conveniently visualize, analyze, and explore their resistome datasets, promoting hypothesis generation and knowledge discovery in the field. Several statistical and integrative analysis methods, as well as visualization approaches supported by this tool, were utilized in our clinical studies (Papers I and II). The development of such high-throughput analysis tools promotes the application of metagenomics-based resistome investigations for surveillance, enhancing our



understanding of AMR and its dynamics in humans, animals, and environmental microbial communities.

It is expected that the data, results, and bioinformatics resources generated as part of this thesis may provide valuable insights and contribute to the field, enhancing our understanding and facilitating future studies in this area to help combat the global threat of AMR.

## References

1. Crofts TS, Gasparrini AJ, Dantas G. Next-generation approaches to understand and combat the antibiotic resistome. *Nature Reviews Microbiology*. 2017;15(7):422-34.
2. Walsh C. Antibiotics: actions, origins, resistance: American Society for Microbiology (ASM); 2003.
3. Fair RJ, Tor Y. Antibiotics and bacterial resistance in the 21st century. *Perspectives in medicinal chemistry*. 2014;6:PMC. S14459.
4. Aminov RI. A brief history of the antibiotic era: lessons learned and challenges for the future. *Frontiers in microbiology*. 2010;1:134.
5. Coates AR, Halls G, Hu Y. Novel classes of antibiotics or more of the same? *British journal of pharmacology*. 2011;163(1):184-94.
6. Ventola CL. The antibiotic resistance crisis: part 1: causes and threats. *Pharmacy and therapeutics*. 2015;40(4):277.
7. Alekshun MN, Levy SB. Molecular mechanisms of antibacterial multidrug resistance. *Cell*. 2007;128(6):1037-50.
8. Barber M. Staphylococcal infection due to penicillin-resistant strains. *British medical journal*. 1947;2(4534):863.
9. Poirel L, Nordmann P. Carbapenem resistance in *Acinetobacter baumannii*: mechanisms and epidemiology. *Clinical Microbiology and Infection*. 2006;12(9):826-36.
10. Liu Y-Y, Wang Y, Walsh TR, Yi L-X, Zhang R, Spencer J, et al. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. *The Lancet infectious diseases*. 2016;16(2):161-8.
11. Prestinaci F, Pezzotti P, Pantosti A. Antimicrobial resistance: a global multifaceted phenomenon. *Pathogens and global health*. 2015;109(7):309-18.
12. Bloom DE, Black S, Salisbury D, Rappuoli R. Antimicrobial resistance and the role of vaccines. *Proceedings of the National Academy of Sciences*. 2018;115(51):12868-71.
13. Centers for Disease Control and Prevention. Antibiotic: antimicrobial Resistance. <http://www.cdc.gov/drugresistance/index.html>. 2011.
14. World Health Organization. Global action plan on antimicrobial resistance. 2015.
15. Murray CJ, Ikuta KS, Sharara F, Swetschinski L, Aguilar GR, Gray A, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*. 2022;399(10325):629-55.
16. Cassini A, Högberg LD, Plachouras D, Quattrocchi A, Hoxha A, Simonsen GS, et al. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *The Lancet infectious diseases*. 2019;19(1):56-66.
17. Wernli D, Jørgensen PS, Morel CM, Carroll S, Harbarth S, Levrat N, et al. Mapping global policy discourse on antimicrobial resistance. *BMJ Global Health*. 2017;2(2):e000378.
18. Boolchandani M, D'Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. *Nature Reviews Genetics*. 2019;20(6):356-70.
19. Breijyeh Z, Jubeh B, Karaman R. Resistance of gram-negative bacteria to current antibacterial agents and approaches to resolve it. *Molecules*. 2020;25(6):1340.

20. Gupta CL, Tiwari RK, Cytryn E. Platforms for elucidating antibiotic resistance in single genomes and complex metagenomes. *Environment international*. 2020;138:105667.
21. Stokes HW, Gillings MR. Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens. *FEMS microbiology reviews*. 2011;35(5):790-819.
22. Von Wintersdorff CJ, Penders J, Van Niekerk JM, Mills ND, Majumder S, Van Alphen LB, et al. Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Frontiers in microbiology*. 2016;7:173.
23. Nnadozie CF, Odume ON. Freshwater environments as reservoirs of antibiotic resistant bacteria and their role in the dissemination of antibiotic resistance genes. *Environmental Pollution*. 2019;254:113067.
24. Sletteemås JS, Urdahl A-M, Mo SS, Johannessen GS, Grave K, Norström M, et al. Imported food and feed as contributors to the introduction of plasmid-mediated colistin-resistant Enterobacteriaceae to a 'low prevalence' country. *Journal of Antimicrobial Chemotherapy*. 2017;72(9):2675-7.
25. Van Gompel L, Luiken RE, Hansen RB, Munk P, Bouwknegt M, Heres L, et al. Description and determinants of the faecal resistome and microbiome of farmers and slaughterhouse workers: a metagenome-wide cross-sectional study. *Environment international*. 2020;143:105939.
26. Allen HK, Donato J, Wang HH, Cloud-Hansen KA, Davies J, Handelsman J. Call of the wild: antibiotic resistance genes in natural environments. *Nature reviews microbiology*. 2010;8(4):251-9.
27. Oteo J, Navarro C, Cercenado E, Delgado-Iribarren A, Wilhelmi I, Orden B, et al. Spread of Escherichia coli strains with high-level cefotaxime and ceftazidime resistance between the community, long-term care facilities, and hospital institutions. *Journal of Clinical Microbiology*. 2006;44(7):2359-66.
28. Zhang J, Zheng B, Zhao L, Wei Z, Ji J, Li L, et al. Nationwide high prevalence of CTX-M and an increase of CTX-M-55 in Escherichia coli isolated from patients with community-onset infections in Chinese county hospitals. *BMC infectious diseases*. 2014;14(1):1-10.
29. D'Souza AW, Boolchandani M, Patel S, Galazzo G, van Hattem JM, Arcilla MS, et al. Destination shapes antibiotic resistance gene acquisitions, abundance increases, and diversity changes in Dutch travelers. *Genome medicine*. 2021;13(1):1-21.
30. Östholm-Balkhed Å, Tärnberg M, Nilsson M, Nilsson LE, Hanberger H, Hällgren A, et al. Travel-associated faecal colonization with ESBL-producing Enterobacteriaceae: incidence and risk factors. *Journal of Antimicrobial Chemotherapy*. 2013;68(9):2144-53.
31. Davies J, Davies D. Origins and evolution of antibiotic resistance. *Microbiology and molecular biology reviews*. 2010;74(3):417-33.
32. Cecchini M. Antimicrobial resistance in G7 countries and beyond: Economic issues, policies and options for action. (*No Title*). 2015.
33. Tenover FC. Mechanisms of antimicrobial resistance in bacteria. *The American journal of medicine*. 2006;119(6):S3-S10.
34. Llor C, Bjerrum L. Antimicrobial resistance: risk associated with antibiotic overuse and initiatives to reduce the problem. *Therapeutic advances in drug safety*. 2014;5(6):229-41.
35. Walsh F. Investigating antibiotic resistance in non-clinical environments. *Frontiers in microbiology*. 2013;4:19.

36. Segawa T, Takeuchi N, Rivera A, Yamada A, Yoshimura Y, Barcaza G, et al. Distribution of antibiotic resistance genes in glacier environments. *Environmental microbiology reports*. 2013;5(1):127-34.
37. D'Costa VM, King CE, Kalan L, Morar M, Sung WW, Schwarz C, et al. Antibiotic resistance is ancient. *Nature*. 2011;477(7365):457-61.
38. Forsberg KJ, Patel S, Gibson MK, Lauber CL, Knight R, Fierer N, et al. Bacterial phylogeny structures soil resistomes across habitats. *Nature*. 2014;509(7502):612-6.
39. Sommer MO, Dantas G, Church GM. Functional characterization of the antibiotic resistance reservoir in the human microflora. *science*. 2009;325(5944):1128-31.
40. Kim D-W, Cha C-J. Antibiotic resistome from the One-Health perspective: understanding and controlling antimicrobial resistance transmission. *Experimental & Molecular Medicine*. 2021;53(3):301-9.
41. Perry JA, Westman EL, Wright GD. The antibiotic resistome: what's new? *Current opinion in microbiology*. 2014;21:45-50.
42. Thänert R, Sawhney SS, Schwartz DJ, Dantas G. The resistance within: Antibiotic disruption of the gut microbiome and resistome dynamics in infancy. *Cell Host & Microbe*. 2022;30(5):675-83.
43. Brinkac L, Voorhies A, Gomez A, Nelson KE. The threat of antimicrobial resistance on the human microbiome. *Microbial ecology*. 2017;74(4):1001-8.
44. Schwartz DJ, Langdon AE, Dantas G. Understanding the impact of antibiotic perturbation on the human microbiome. *Genome medicine*. 2020;12(1):1-12.
45. Allemann A, Kraemer JG, Korten I, Ramsey K, Casaulta C, Group SS, et al. Nasal resistome development in infants with cystic fibrosis in the first year of life. *Frontiers in microbiology*. 2019;10:212.
46. Carr VR, Witherden EA, Lee S, Shoaie S, Mullany P, Proctor GB, et al. Abundance and diversity of resistomes differ between healthy human oral cavities and gut. *Nature communications*. 2020;11(1):693.
47. Feng J, Li B, Jiang X, Yang Y, Wells GF, Zhang T, et al. Antibiotic resistome in a large-scale healthy human gut microbiota deciphered by metagenomic and network analyses. *Environmental microbiology*. 2018;20(1):355-68.
48. Li Z, Xia J, Jiang L, Tan Y, An Y, Zhu X, et al. Characterization of the human skin resistome and identification of two microbiota cutotypes. *Microbiome*. 2021;9(1):1-18.
49. Maestre-Carballa L, Navarro-López V, Martínez-García M. A resistome roadmap: from the human body to pristine environments. *Frontiers in Microbiology*. 2022:1107.
50. Clemente JC, Pehrsson EC, Blaser MJ, Sandhu K, Gao Z, Wang B, et al. The microbiome of uncontacted Amerindians. *Science advances*. 2015;1(3):e1500183.
51. Gibson MK, Wang B, Ahmadi S, Burnham CA, Tarr PI, Warner BB, et al. Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. *Nature Microbiology*. 2016;1:16024.
52. Lebeaux RM, Coker MO, Dade EF, Palys TJ, Morrison HG, Ross BD, et al. The infant gut resistome is associated with *E. coli* and early-life exposures. *BMC Microbiol*. 2021;21(1):201.
53. Gasparrini AJ, Wang B, Sun X, Kennedy EA, Hernandez-Leyva A, Ndao IM, et al. Persistent metagenomic signatures of early-life hospitalization and antibiotic treatment in the infant gut microbiota and resistome. *Nature microbiology*. 2019;4(12):2285-97.

54. Sun J, Liao X-P, D'Souza AW, Boolchandani M, Li S-H, Cheng K, et al. Environmental remodeling of human gut microbiota and antibiotic resistome in livestock farms. *Nature communications*. 2020;11(1):1427.
55. Li X, Stokholm J, Brejnrod A, Vestergaard GA, Russel J, Trivedi U, et al. The infant gut resistome associates with *E. coli*, environmental exposures, gut microbiome maturity, and asthma-associated bacterial composition. *Cell Host & Microbe*. 2021;29(6):975-87. e4.
56. Yassour M, Vatanen T, Siljander H, Hämäläinen A-M, Härkönen T, Ryhänen SJ, et al. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Science translational medicine*. 2016;8(343):343ra81-ra81.
57. Christoff AP, Sereia AFR, Cruz GNF, Bastiani DCd, Silva VL, Hernandez C, et al. One year cross-sectional study in adult and neonatal intensive care units reveals the bacterial and antimicrobial resistance genes profiles in patients and hospital surfaces. *PLoS One*. 2020;15(6):e0234127.
58. Sosa-Moreno A, Comstock SS, Sugino KY, Ma TF, Paneth N, Davis Y, et al. Perinatal risk factors for fecal antibiotic resistance gene patterns in pregnant women and their infants. *PLoS One*. 2020;15(6):e0234751.
59. Pärnänen K, Karkman A, Hultman J, Lyra C, Bengtsson-Palme J, Larsson DJ, et al. Maternal gut and breast milk microbiota affect infant gut antibiotic resistome and mobile genetic elements. *Nature Communications*. 2018;9(1):3891.
60. O'Connor L, Heyderman R. The challenges of defining the human nasopharyngeal resistome. *Trends in Microbiology*. 2023.
61. Nelson MT, Pope CE, Marsh RL, Wolter DJ, Weiss EJ, Hager KR, et al. Human and extracellular DNA depletion for metagenomic analysis of complex clinical infection samples yields optimized viable microbiome profiles. *Cell reports*. 2019;26(8):2227-40. e5.
62. Shi Y, Wang G, Lau HC-H, Yu J. Metagenomic sequencing for microbial DNA in human samples: emerging technological advances. *International Journal of Molecular Sciences*. 2022;23(4):2181.
63. Rajar P, Dhariwal A, Salvadori Da Silva G, Junges R, Åmdal HA, Berild D, et al. Microbial DNA extraction of high-host content and low biomass samples: Optimized protocol for nasopharynx metagenomic studies. *Frontiers in Microbiology*. 2022;13.
64. Modi SR, Collins JJ, Relman DA. Antibiotics and the gut microbiota. *The Journal of clinical investigation*. 2014;124(10):4212-8.
65. Morley VJ, Woods RJ, Read AF. Bystander selection for antimicrobial resistance: implications for patient health. *Trends in microbiology*. 2019;27(10):864-77.
66. Anthony WE, Wang B, Sukhum KV, D'Souza AW, Hink T, Cass C, et al. Acute and persistent effects of commonly used antibiotics on the gut microbiome and resistome in healthy adults. *Cell reports*. 2022;39(2):110649.
67. Forslund K, Sunagawa S, Kultima JR, Mende DR, Arumugam M, Typas A, et al. Country-specific antibiotic use practices impact the human gut resistome. *Genome research*. 2013;23(7):1163-9.
68. Jakobsson HE, Jernberg C, Andersson AF, Sjölund-Karlsson M, Jansson JK, Engstrand L. Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. *PloS one*. 2010;5(3):e9836.
69. Gasparrini AJ, Crofts TS, Gibson MK, Tarr PI, Warner BB, Dantas G. Antibiotic perturbation of the preterm infant gut microbiome and resistome. *Gut Microbes*. 2016;7(5):443-9.

70. Reyman M, van Houten MA, Watson RL, Chu M, Arp K, de Waal WJ, et al. Effects of early-life antibiotics on the developing infant gut microbiome and resistome: a randomized trial. *Nat Commun.* 2022;13(1):893.
71. Dawson-Hahn EE, Mickan S, Onakpoya I, Roberts N, Kronman M, Butler CC, et al. Short-course versus long-course oral antibiotic treatment for infections treated in outpatient settings: a review of systematic reviews. *Family practice.* 2017;34(5):511-9.
72. Sturød K, Dhariwal A, Dahle UR, Vestrheim DF, Petersen FC. Impact of narrow-spectrum penicillin V on the oral and faecal resistome in a young child treated for otitis media. *Journal of Global Antimicrobial Resistance.* 2020;20:290-7.
73. Palleja A, Mikkelsen KH, Forslund SK, Kashani A, Allin KH, Nielsen T, et al. Recovery of gut microbiota of healthy adults following antibiotic exposure. *Nature microbiology.* 2018;3(11):1255-65.
74. Zaura E, Brandt BW, Teixeira de Mattos MJ, Buijs MJ, Caspers MP, Rashid M-U, et al. Same exposure but two radically different responses to antibiotics: resilience of the salivary microbiome versus long-term microbial shifts in feces. *MBio.* 2015;6(6):e01693-15.
75. Prudhomme M, Attaiech L, Sanchez G, Martin B, Claverys J-P. Antibiotic stress induces genetic transformability in the human pathogen *Streptococcus pneumoniae*. *Science.* 2006;313(5783):89-92.
76. Sturød K, Salvadori G, Junges R, Petersen F. Antibiotics alter the window of competence for natural transformation in streptococci. *Molecular oral microbiology.* 2018;33(5):378-87.
77. Liu G, Bogaj K, Bortolaia V, Olsen JE, Thomsen LE. Antibiotic-induced, increased conjugative transfer is common to diverse naturally occurring ESBL plasmids in *Escherichia coli*. *Frontiers in microbiology.* 2019;10:2119.
78. Van Schaik W. The human gut resistome. *Philosophical Transactions of the Royal Society B: Biological Sciences.* 2015;370(1670):20140087.
79. Wiegand I, Hilpert K, Hancock RE. Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nature protocols.* 2008;3(2):163-75.
80. Bauer A. Antibiotic susceptibility testing by a standardized single diffusion method. *Am J Clin Pathol.* 1966;45:493-6.
81. Brown DF, Brown L. Evaluation of the E test, a novel method of quantifying antimicrobial activity. *Journal of Antimicrobial Chemotherapy.* 1991;27(2):185-90.
82. Lob SH, Hoban DJ, Sahm DF, Badal RE. Regional differences and trends in antimicrobial susceptibility of *Acinetobacter baumannii*. *International journal of antimicrobial agents.* 2016;47(4):317-23.
83. NIHR Global Health Research Unit on Genomic Surveillance of AMR. Whole-genome sequencing as part of national and international surveillance programmes for antimicrobial resistance: A roadmap. *BMJ Global Health.* 2020;5(11):e002244.
84. World Health Organization. GLASS whole-genome sequencing for surveillance of antimicrobial resistance. 2020.
85. World Health Organization. Guidelines for surveillance of drug resistance in tuberculosis: World Health Organization; 2009.
86. Walker TM, Kohl TA, Omar SV, Hedge J, Elias CDO, Bradley P, et al. Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *The Lancet infectious diseases.* 2015;15(10):1193-202.



87. Nakano S, Fujisawa T, Ito Y, Chang B, Matsumura Y, Yamamoto M, et al. Spread of meropenem-resistant *Streptococcus pneumoniae* serotype 15A-ST63 clone in Japan, 2012–2014. *Emerging Infectious Diseases*. 2018;24(2):275.
88. Gorrie CL, Mirceta M, Wick RR, Judd LM, Wyres KL, Thomson NR, et al. Antimicrobial-resistant *Klebsiella pneumoniae* carriage and infection in specialized geriatric care wards linked to acquisition in the referring hospital. *Clinical infectious diseases*. 2018;67(2):161-70.
89. Steen AD, Crits-Christoph A, Carini P, DeAngelis KM, Fierer N, Lloyd KG, et al. High proportions of bacteria and archaea across most biomes remain uncultured. *The ISME journal*. 2019;13(12):3126-30.
90. D'Costa VM, McGrann KM, Hughes DW, Wright GD. Sampling the antibiotic resistome. *Science*. 2006;311(5759):374-7.
91. Hu X, Xu B, Yang Y, Liu D, Yang M, Wang J, et al. A high throughput multiplex PCR assay for simultaneous detection of seven aminoglycoside-resistance genes in Enterobacteriaceae. *BMC microbiology*. 2013;13(1):1-9.
92. Li Y, Shen Z, Ding S, Wang S. A TaqMan-based multiplex real-time PCR assay for the rapid detection of tigecycline resistance genes from bacteria, faeces and environmental samples. *BMC microbiology*. 2020;20:1-7.
93. Lee K, Kim D-W, Lee D-H, Kim Y-S, Bu J-H, Cha J-H, et al. Mobile resistome of human gut and pathogen drives anthropogenic bloom of antibiotic resistance. *Microbiome*. 2020;8(1):1-14.
94. Jernberg C, Löfmark S, Edlund C, Jansson JK. Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. *The ISME journal*. 2007;1(1):56-66.
95. An X-L, Su J-Q, Li B, Ouyang W-Y, Zhao Y, Chen Q-L, et al. Tracking antibiotic resistome during wastewater treatment using high throughput quantitative PCR. *Environment international*. 2018;117:146-53.
96. Do TT, Nolan S, Hayes N, O'Flaherty V, Burgess C, Brennan F, et al. Metagenomic and HT-qPCR analysis reveal the microbiome and resistome in pig slurry under storage, composting, and anaerobic digestion. *Environmental Pollution*. 2022;305:119271.
97. Gijón D, Curiao T, Baquero F, Coque TM, Cantón R. Fecal carriage of carbapenemase-producing Enterobacteriaceae: a hidden reservoir in hospitalized and nonhospitalized patients. *Journal of clinical microbiology*. 2012;50(5):1558-63.
98. Palladino S, Kay ID, Flexman JP, Boehm I, Costa AMG, Lambert EJ, et al. Rapid detection of vanA and vanB genes directly from clinical specimens and enrichment broths by real-time multiplex PCR assay. *Journal of Clinical Microbiology*. 2003;41(6):2483-6.
99. Roger M, Faucher M-C, Forest P, St-Antoine P, Coutlée Fo. Evaluation of a vanA-specific PCR assay for detection of vancomycin-resistant *Enterococcus faecium* during a hospital outbreak. *Journal of Clinical Microbiology*. 1999;37(10):3348-9.
100. Vo AT, Van Duijkeren E, Gaastra W, Fluit AC. Antimicrobial resistance, class 1 integrons, and genomic island 1 in *Salmonella* isolates from Vietnam. *PloS one*. 2010;5(2):e9440.
101. Finley RL, Collignon P, Larsson DJ, McEwen SA, Li X-Z, Gaze WH, et al. The scourge of antibiotic resistance: the important role of the environment. *Clinical infectious diseases*. 2013;57(5):704-10.

102. Surette MD, Wright GD. Lessons from the environmental antibiotic resistome. *Annual review of microbiology*. 2017;71:309-29.
103. Wright GD. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nature reviews microbiology*. 2007;5(3):175-86.
104. Pehrsson EC, Forsberg KJ, Gibson MK, Ahmadi S, Dantas G. Novel resistance functions uncovered using functional metagenomic investigations of resistance reservoirs. *Frontiers in microbiology*. 2013;4:145.
105. Forsberg KJ, Patel S, Wencewicz TA, Dantas G. The tetracycline destructases: a novel family of tetracycline-inactivating enzymes. *Chemistry & biology*. 2015;22(7):888-97.
106. Spanogiannopoulos P, Waglechner N, Koteva K, Wright GD. A rifamycin inactivating phosphotransferase family shared by environmental and pathogenic bacteria. *Proceedings of the National Academy of Sciences*. 2014;111(19):7102-7.
107. Mullany P. Functional metagenomics for the investigation of antibiotic resistance. *Virulence*. 2014;5(3):443-7.
108. Penders J, Stobberingh EE, Savelkoul PH, Wolffs PF. The human microbiome as a reservoir of antimicrobial resistance. *Frontiers in microbiology*. 2013;4:87.
109. Imchen M, Moopantakath J, Kumavath R, Barh D, Tiwari S, Ghosh P, et al. Current trends in experimental and computational approaches to combat antimicrobial resistance. *Frontiers in Genetics*. 2020;11:563975.
110. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. *Briefings in bioinformatics*. 2019;20(4):1125-36.
111. Bengtsson-Palme J, Larsson DJ, Kristiansson E. Using metagenomics to investigate human and environmental resistomes. *Journal of Antimicrobial Chemotherapy*. 2017;72(10):2690-703.
112. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. *Nature biotechnology*. 2017;35(9):833-44.
113. Pehrsson EC, Tsukayama P, Patel S, Mejía-Bautista M, Sosa-Soto G, Navarrete KM, et al. Interconnected microbiomes and resistomes in low-income human habitats. *Nature*. 2016;533(7602):212-6.
114. Hu Y, Yang X, Qin J, Lu N, Cheng G, Wu N, et al. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nature communications*. 2013;4(1):2151.
115. Qiu Q, Wang J, Yan Y, Roy B, Chen Y, Shang X, et al. Metagenomic analysis reveals the distribution of antibiotic resistance genes in a large-scale population of healthy individuals and patients with varied diseases. *Frontiers in Molecular Biosciences*. 2020;7:590018.
116. Wang R, Van Dorp L, Shaw LP, Bradley P, Wang Q, Wang X, et al. The global distribution and spread of the mobilized colistin resistance gene *mcr-1*. *Nature communications*. 2018;9(1):1179.
117. Munk P, Knudsen BE, Lukjancenko O, Duarte ASR, Van Gompel L, Luiken RE, et al. Abundance and diversity of the faecal resistome in slaughter pigs and broilers in nine European countries. *Nature microbiology*. 2018;3(8):898-908.
118. Yin X, Li L, Chen X, Liu Y-Y, Lam TT-Y, Topp E, et al. Global environmental resistome: Distinction and connectivity across diverse habitats benchmarked by metagenomic analyses. *Water Research*. 2023;235:119875.



119. Martiny H-M, Munk P, Brinch C, Szarvas J, Aarestrup FM, Petersen TN. Global distribution of mcr gene variants in 214K metagenomic samples. *Msystems*. 2022;7(2):e00105-22.
120. Che Y, Xia Y, Liu L, Li A-D, Yang Y, Zhang T. Mobile antibiotic resistome in wastewater treatment plants revealed by Nanopore metagenomic sequencing. *Microbiome*. 2019;7(1):1-13.
121. Lebeaux RM, Coker MO, Dade EF, Palys TJ, Morrison HG, Ross BD, et al. The infant gut resistome is associated with E. coli and early-life exposures. *BMC microbiology*. 2021;21(1):201.
122. Parnanen K, Karkman A, Hultman J, Lyra C, Bengtsson-Palme J, Larsson DGJ, et al. Maternal gut and breast milk microbiota affect infant gut antibiotic resistome and mobile genetic elements. *Nat Commun*. 2018;9(1):3891.
123. Willmann M, Vehreschild MJ, Biehl LM, Vogel W, Dörfel D, Hamprecht A, et al. Distinct impact of antibiotics on the gut microbiome and resistome: a longitudinal multicenter cohort study. *BMC biology*. 2019;17(1):1-18.
124. Danko D, Bezdán D, Afshin EE, Ahsanuddin S, Bhattacharya C, Butler DJ, et al. A global metagenomic map of urban microbiomes and antimicrobial resistance. *Cell*. 2021;184(13):3376-93. e17.
125. Hendriksen RS, Munk P, Njage P, Van Bunnik B, McNally L, Lukjancenko O, et al. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nature communications*. 2019;10(1):1124.
126. Munk P, Andersen VD, de Knecht L, Jensen MS, Knudsen BE, Lukjancenko O, et al. A sampling and metagenomic sequencing-based methodology for monitoring antimicrobial resistance in swine herds. *Journal of Antimicrobial Chemotherapy*. 2017;72(2):385-92.
127. Munk P, Brinch C, Møller FD, Petersen TN, Hendriksen RS, Seyfarth AM, et al. Genomic analysis of sewage from 101 countries reveals global landscape of antimicrobial resistance. *Nature Communications*. 2022;13(1):7251.
128. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *science*. 1995;269(5223):496-512.
129. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*. 1977;74(12):5463-7.
130. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. *Genomics*. 2016;107(1):1-8.
131. Van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. *Trends in genetics*. 2014;30(9):418-26.
132. Gupta N, Verma VK. Next-generation sequencing and its application: empowering in public health beyond reality. *Microbial Technology for the Welfare of Society*. 2019:313-41.
133. Rhoads A, Au KF. PacBio sequencing and its applications. *Genomics, proteomics & bioinformatics*. 2015;13(5):278-89.
134. Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. Zero-mode waveguides for single-molecule analysis at high concentrations. *science*. 2003;299(5607):682-6.
135. Suzuki Y, Nishijima S, Furuta Y, Yoshimura J, Suda W, Oshima K, et al. Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut. *Microbiome*. 2019;7:1-16.

136. Pearman WS, Freed NE, Silander OK. Testing the advantages and disadvantages of short-and long-read eukaryotic metagenomics using simulated reads. *BMC bioinformatics*. 2020;21(1):1-15.
137. Taxt AM, Avershina E, Frye SA, Naseer U, Ahmad R. Rapid identification of pathogens, antibiotic resistance genes and plasmids in blood cultures by nanopore sequencing. *Scientific reports*. 2020;10(1):1-11.
138. Zhang H, Wang M, Han X, Wang T, Lei Y, Rao Y, et al. The application of targeted nanopore sequencing for the identification of pathogens and resistance genes in lower respiratory tract infections. *Frontiers in Microbiology*. 2022;13.
139. Chapman R, Jones L, D'Angelo A, Suliman A, Anwar M, Bagby S. Nanopore-Based Metagenomic Sequencing in Respiratory Tract Infection: A Developing Diagnostic Platform. *Lung*. 2023:1-9.
140. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature*. 2016;530(7589):228-32.
141. Tyler AD, Mataseje L, Urfano CJ, Schmidt L, Antonation KS, Mulvey MR, et al. Evaluation of Oxford Nanopore's MinION sequencing device for microbial whole genome sequencing applications. *Scientific reports*. 2018;8(1):1-12.
142. Wang Y, Zhao Y, Bollas A, Wang Y, Au KF. Nanopore sequencing technology, bioinformatics and applications. *Nature biotechnology*. 2021;39(11):1348-65.
143. Charalampous T, Kay GL, Richardson H, Aydin A, Baldan R, Jeanes C, et al. Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nature biotechnology*. 2019;37(7):783-92.
144. Qian X, Gunturu S, Sun W, Cole JR, Norby B, Gu J, et al. Long-read sequencing revealed cooccurrence, host range, and potential mobility of antibiotic resistome in cow feces. *Proceedings of the National Academy of Sciences*. 2021;118(25):e2024464118.
145. A  
om; 2010.
146. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-20.
147. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*. 2011;17(1):10-2.
148. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature methods*. 2012;9(4):357-9.
149. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*. 2009;25(14):1754-60.
150. Clausen PT, Aarestrup FM, Lund O. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC bioinformatics*. 2018;19:1-8.
151. Doster E, Lakin SM, Dean CJ, Wolfe C, Young JG, Boucher C, et al. MEGARes 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. *Nucleic acids research*. 2020;48(D1):D561-D9.
152. Yin X, Jiang X-T, Chai B, Li L, Yang Y, Cole JR, et al. ARGs-OAP v2. 0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. *Bioinformatics*. 2018;34(13):2263-70.

153. Clausen PT, Zankari E, Aarestrup FM, Lund O. Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. *Journal of Antimicrobial Chemotherapy*. 2016;71(9):2484-8.
154. Kaminski J, Gibson MK, Franzosa EA, Segata N, Dantas G, Huttenhower C. High-specificity targeted functional profiling in microbial communities with ShortBRED. *PLoS computational biology*. 2015;11(12):e1004557.
155. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. *Genome research*. 2017;27(5):824-34.
156. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. 2015;31(10):1674-6.
157. Namiki T, Hachiya T, Tanaka H, Sakakibara Y, editors. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Proceedings of the 2nd ACM conference on bioinformatics, computational biology and biomedicine; 2011.
158. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of molecular biology*. 1990;215(3):403-10.
159. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nature methods*. 2015;12(1):59-60.
160. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010;26(19):2460-1.
161. Hendriksen RS, Bortolaia V, Tate H, Tyson GH, Aarestrup FM, McDermott PF. Using genomics to track global antimicrobial resistance. *Frontiers in public health*. 2019;7:242.
162. Henson J, Tischler G, Ning Z. Next-generation sequencing and large genome assemblies. *Pharmacogenomics*. 2012;13(8):901-15.
163. Lee K, Kim D-W, Cha C-J. Overview of bioinformatic methods for analysis of antibiotic resistome from genome and metagenome data. *Journal of Microbiology*. 2021;59:270-80.
164. Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*. 2018;6:1-15.
165. Lakin SM, Dean C, Noyes NR, Dettenwanger A, Ross AS, Doster E, et al. MEGARes: an antimicrobial resistance database for high throughput sequencing. *Nucleic acids research*. 2017;45(D1):D574-D80.
166. Thai QK, Pleiss J. SHV Lactamase Engineering Database: a reconciliation tool for SHV  $\beta$ -lactamases in public databases. *BMC genomics*. 2010;11:1-8.
167. Srivastava A, Singhal N, Goel M, Viridi JS, Kumar M. CBMAR: a comprehensive  $\beta$ -lactamase molecular annotation resource. *Database*. 2014;2014.
168. Saha SB, Uttam V, Verma V. u-CARE: user-friendly Comprehensive Antibiotic resistance Repository of Escherichia coli. *Journal of clinical pathology*. 2015;68(8):648-51.
169. Alcock BP, Raphenya AR, Lau TT, Tsang KK, Bouchard M, Edalatmand A, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic acids research*. 2020;48(D1):D517-D25.
170. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, et al. ResFinder 4.0 for predictions of phenotypes from genotypes. *Journal of Antimicrobial Chemotherapy*. 2020;75(12):3491-500.
171. Zankari E, Allesøe R, Joensen KG, Cavaco LM, Lund O, Aarestrup FM. PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with

- chromosomal point mutations in bacterial pathogens. *Journal of Antimicrobial Chemotherapy*. 2017;72(10):2764-8.
172. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrobial agents and chemotherapy*. 2014;58(1):212-20.
173. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, et al. Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrobial agents and chemotherapy*. 2019;63(11):e00483-19.
174. Berglund F, Österlund T, Boulund F, Marathe NP, Larsson D, Kristiansson E. Identification and reconstruction of novel antibiotic resistance genes from metagenomes. *Microbiome*. 2019;7(1):1-14.
175. Galhano BS, Ferrari RG, Panzenhagen P, de Jesus ACS, Conte-Junior CA. Antimicrobial resistance gene detection methods for bacteria in animal-based foods: A brief review of highlights and advantages. *Microorganisms*. 2021;9(5):923.
176. Papp M, Solymosi N. Review and comparison of antimicrobial resistance gene databases. *Antibiotics*. 2022;11(3):339.
177. Xavier BB, Das AJ, Cochrane G, De Ganck S, Kumar-Singh S, Aarestrup FM, et al. Consolidating and exploring antibiotic resistance gene data resources. *Journal of clinical microbiology*. 2016;54(4):851-9.
178. Arango-Argoty GA, Guron G, Garner E, Riquelme MV, Heath LS, Pruden A, et al. ARGminer: a web platform for the crowdsourcing-based curation of antibiotic resistance genes. *Bioinformatics*. 2020;36(9):2966-73.
179. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *The ISME journal*. 2015;9(1):207-16.
180. Ruppé E, Ghoulane A, Tap J, Pons N, Alvarez A-S, Maziers N, et al. Prediction of the intestinal resistome by a three-dimensional structure-based method. *Nature microbiology*. 2019;4(1):112-23.
181. Calle ML. Statistical analysis of metagenomics data. *Genomics & informatics*. 2019;17(1).
182. Li H. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*. 2015;2:73-94.
183. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. *Frontiers in microbiology*. 2017;8:2224.
184. Pérez-Cobas AE, Gomez-Valero L, Buchrieser C. Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. *Microbial genomics*. 2020;6(8).
185. Aitchison J. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1982;44(2):139-60.
186. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*. 2017;5:1-18.
187. Pereira MB, Wallroth M, Jonsson V, Kristiansson E. Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC genomics*. 2018;19:1-17.

188. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS computational biology*. 2014;10(4):e1003531.
189. Chong J, Liu P, Zhou G, Xia J. Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. *Nature protocols*. 2020;15(3):799-821.
190. Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics*. 2013;14(6):671-83.
191. Lin H, Peddada SD. Analysis of microbial compositions: a review of normalization and differential abundance analysis. *NPJ biofilms and microbiomes*. 2020;6(1):60.
192. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics*. 2010;11(1):1-13.
193. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. *Nature methods*. 2013;10(12):1200-2.
194. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*. 2010;11(3):1-9.
195. Sisk-Hackworth L, Kelley ST. An application of compositional data analysis to multiomic time-series data. *NAR Genomics and Bioinformatics*. 2020;2(4):lqaa079.
196. Jonsson V, Österlund T, Nerman O, Kristiansson E. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. *BMC genomics*. 2016;17(1):1-14.
197. Poulsen CS, Aarestrup FM, Brinch C, Ekstrøm CT. Feature matrix normalization, transformation and calculation of  $\beta$ -diversity in metagenomics: Theoretical and applied perspectives on your decisions. *bioRxiv*. 2019:859157.
198. Morgan XC, Huttenhower C. Chapter 12: Human microbiome analysis. *PLoS computational biology*. 2012;8(12):e1002808.
199. Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ. Counting the uncountable: statistical approaches to estimating microbial diversity. *Applied and environmental microbiology*. 2001;67(10):4399-406.
200. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. *Nature Reviews Microbiology*. 2018;16(7):410-22.
201. Bengtsson-Palme J. The diversity of uncharacterized antibiotic resistance genes can be predicted from known gene variants—but not always. *Microbiome*. 2018;6:1-12.
202. Savva G. 8. Statistical Analysis of Microbiome Data.
203. Goodrich JK, Di Rienzi SC, Poole AC, Koren O, Walters WA, Caporaso JG, et al. Conducting a microbiome study. *Cell*. 2014;158(2):250-62.
204. Kuczynski J, Liu Z, Lozupone C, McDonald D, Fierer N, Knight R. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nature methods*. 2010;7(10):813-9.
205. Kodikara S, Ellul S, Lê Cao K-A. Statistical challenges in longitudinal microbiome data analysis. *Briefings in Bioinformatics*. 2022;23(4):bbac273.
206. Aldirawi H, Morales FG. Univariate and Multivariate Statistical Analysis of Microbiome Data: An Overview. *Applied Microbiology*. 2023;3(2):322-38.
207. Ghanbari M, Klose V, Crispie F, Cotter PD. The dynamics of the antibiotic resistome in the feces of freshly weaned pigs following therapeutic administration of oxytetracycline. *Scientific reports*. 2019;9(1):1-11.



208. Yue Y, Hu Y-J. Extension of PERMANOVA to Testing the Mediation Effect of the Microbiome. *Genes*. 2022;13(6):940.
209. Anderson MJ. A new method for non-parametric multivariate analysis of variance. *Austral ecology*. 2001;26(1):32-46.
210. Anderson MJ, Walsh DC. PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing? *Ecological monographs*. 2013;83(4):557-74.
211. Clarke KR. Non-parametric multivariate analyses of changes in community structure. *Australian journal of ecology*. 1993;18(1):117-43.
212. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. *Genome biology*. 2011;12:1-18.
213. Nearing JT, Douglas GM, Hayes MG, MacDonald J, Desai DK, Allward N, et al. Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications*. 2022;13(1):342.
214. Paulson JN, Pop M, Bravo HC. metagenomeSeq: Statistical analysis for sparse high-throughput sequencing. *Bioconductor package*. 2013;1(0):191.
215. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*. 2014;15(12):1-21.
216. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*. 2010;26(1):139-40.
217. Gloor G. ALDEx2: ANOVA-Like Differential Expression tool for compositional data. *ALDEX manual modular*. 2015;20:1-11.
218. Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial ecology in health and disease*. 2015;26(1):27663.
219. Mallick H, Rahnavard A, McIver LJ, Ma S, Zhang Y, Nguyen LH, et al. Multivariable association discovery in population-scale meta-omics studies. *PLoS computational biology*. 2021;17(11):e1009442.
220. Yang L, Chen J. A comprehensive evaluation of microbial differential abundance analysis methods: current status and potential solutions. *Microbiome*. 2022;10(1):130.
221. Thorsen J, Brejnrod A, Mortensen M, Rasmussen MA, Stokholm J, Al-Soud WA, et al. Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome*. 2016;4:1-14.
222. Goeman JJ, Solari A. Multiple hypothesis testing in genomics. *Statistics in medicine*. 2014;33(11):1946-78.
223. Chong J, Xia J. Computational approaches for integrative analysis of the metabolome and microbiome. *Metabolites*. 2017;7(4):62.
224. Cao J, Hu Y, Liu F, Wang Y, Bi Y, Lv N, et al. Metagenomic analysis reveals the microbiome and resistome in migratory birds. *Microbiome*. 2020;8:1-18.
225. Noyes NR, Yang X, Linke LM, Magnuson RJ, Dettenwanger A, Cook S, et al. Resistome diversity in cattle and the environment decreases during beef production. *Elife*. 2016;5:e13195.
226. Ma L, Li B, Jiang X-T, Wang Y-L, Xia Y, Li A-D, et al. Catalogue of antibiotic resistome and host-tracking in drinking water deciphered by a large scale survey. *Microbiome*. 2017;5(1):1-12.
227. Gower JC. Generalized procrustes analysis. *Psychometrika*. 1975;40:33-51.

228. Rohart F, Gautier B, Singh A, Lê Cao K-A. mixOmics: An R package for ‘omics feature selection and multiple data integration. *PLoS computational biology*. 2017;13(11):e1005752.
229. You Y, Liang D, Wei R, Li M, Li Y, Wang J, et al. Evaluation of metabolite-microbe correlation detection methods. *Analytical biochemistry*. 2019;567:106-11.
230. Elvers KT, Wilson VJ, Hammond A, Duncan L, Huntley AL, Hay AD, et al. Antibiotic-induced changes in the human gut microbiota for the most commonly prescribed antibiotics in primary care in the UK: a systematic review. *BMJ open*. 2020;10(9):e035677.
231. Leach AJ, Morris PS, Mathews JD, au COMITOGAME. Compared to placebo, long-term antibiotics resolve otitis media with effusion (OME) and prevent acute otitis media with perforation (AOMwiP) in a high-risk population: a randomized controlled trial. *BMC pediatrics*. 2008;8:1-9.
232. Ahmed H, Davies F, Francis N, Farewell D, Butler C, Paranjothy S. Long-term antibiotics for prevention of recurrent urinary tract infection in older adults: systematic review and meta-analysis of randomised trials. *BMJ open*. 2017;7(5):e015233.
233. Bråten LCH, Rolfsen MP, Espeland A, Wigemyr M, Aßmus J, Froholdt A, et al. Efficacy of antibiotic treatment in patients with chronic low back pain and Modic changes (the AIM study): double blind, randomised, placebo controlled, multicentre trial. *bmj*. 2019;367.
234. Albert RK, Connett J, Bailey WC, Casaburi R, Cooper Jr JAD, Criner GJ, et al. Azithromycin for prevention of exacerbations of COPD. *New England Journal of Medicine*. 2011;365(8):689-98.
235. Liu L, Oza S, Hogan D, Chu Y, Perin J, Zhu J, et al. Global, regional, and national causes of under-5 mortality in 2000–15: an updated systematic analysis with implications for the Sustainable Development Goals. *The Lancet*. 2016;388(10063):3027-35.
236. Simon AK, Hollander GA, McMichael A. Evolution of the immune system in humans from infancy to old age. *Proceedings of the Royal Society B: Biological Sciences*. 2015;282(1821):20143085.
237. Vangay P, Ward T, Gerber JS, Knights D. Antibiotics, pediatric dysbiosis, and disease. *Cell host & microbe*. 2015;17(5):553-64.
238. Flannery DD, Ross RK, Mukhopadhyay S, Tribble AC, Puopolo KM, Gerber JS. Temporal trends and center variation in early antibiotic use among premature infants. *JAMA network open*. 2018;1(1):e180164-e.
239. Vatne A, Hapnes N, Stensvold HJ, Dalen I, Guthe HJ, Støen R, et al. Early empirical antibiotics and adverse clinical outcomes in infants born very preterm: a population-based cohort. *The Journal of Pediatrics*. 2023;253:107-14. e5.
240. Arrieta M-C, Stiemsma LT, Dimitriu PA, Thorson L, Russell S, Yurist-Doutsch S, et al. Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Science translational medicine*. 2015;7(307):307ra152-307ra152.
241. Novitsky A, Tuttle D, Locke RG, Saiman L, Mackley A, Paul DA. Prolonged early antibiotic use and bronchopulmonary dysplasia in very low birth weight infants. *American journal of perinatology*. 2014:043-8.
242. Bailey LC, Forrest CB, Zhang P, Richards TM, Livshits A, DeRusso PA. Association of antibiotics in infancy with early childhood obesity. *JAMA pediatrics*. 2014;168(11):1063-9.
243. Man WH, de Steenhuijsen Piters WA, Bogaert D. The microbiota of the respiratory tract: gatekeeper to respiratory health. *Nature Reviews Microbiology*. 2017;15(5):259-70.

244. Swann J, Rajilic-Stojanovic M, Salonen A, Sakwinska O, Gill C, Meynier A, et al. Considerations for the design and conduct of human gut microbiota intervention studies relating to foods. *European journal of nutrition*. 2020;59:3347-68.
245. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. *Nature medicine*. 2018;24(4):392-400.
246. Flores GE, Caporaso JG, Henley JB, Rideout JR, Domogala D, Chase J, et al. Temporal variability is a personalized feature of the human microbiome. *Genome biology*. 2014;15:1-13.
247. Poulsen CS, Kaas RS, Aarestrup FM, Pamp SJ. Standard sample storage conditions have an impact on inferred microbiome composition and antimicrobial resistance patterns. *Microbiology Spectrum*. 2021;9(2):e01387-21.
248. Luna PN, Hasegawa K, Ajami NJ, Espinola JA, Henke DM, Petrosino JF, et al. The association between anterior nares and nasopharyngeal microbiota in infants hospitalized for bronchiolitis. *Microbiome*. 2018;6:1-14.
249. Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. Contamination in low microbial biomass microbiome studies: issues and recommendations. *Trends in microbiology*. 2019;27(2):105-17.
250. McKain N, Genc B, Snelling TJ, Wallace RJ. Differential recovery of bacterial and archaeal 16S rRNA genes from ruminal digesta in response to glycerol as cryoprotectant. *Journal of microbiological methods*. 2013;95(3):381-3.
251. Lozupone CA, Stombaugh J, Gonzalez A, Ackermann G, Wendel D, Vázquez-Baeza Y, et al. Meta-analyses of studies of the human microbiota. *Genome research*. 2013;23(10):1704-14.
252. Sui H-y, Weil AA, Nuwagira E, Qadri F, Ryan ET, Mezzari MP, et al. Impact of DNA extraction method on variation in human and built environment microbial community and functional profiles assessed by shotgun metagenomics sequencing. *Frontiers in microbiology*. 2020;11:953.
253. Wesolowska-Andersen A, Bahl MI, Carvalho V, Kristiansen K, Sicheritz-Pontén T, Gupta R, et al. Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome*. 2014;2(1):1-11.
254. Li A-D, Metch JW, Wang Y, Garner E, Zhang AN, Riquelme MV, et al. Effects of sample preservation and DNA extraction on enumeration of antibiotic resistance genes in wastewater. *FEMS microbiology ecology*. 2018;94(2):fix189.
255. Yuan S, Cohen DB, Ravel J, Abdo Z, Forney LJ. Evaluation of methods for the extraction and purification of DNA from the human microbiome. *PloS one*. 2012;7(3):e33865.
256. Wu W-K, Chen C-C, Panyod S, Chen R-A, Wu M-S, Sheen L-Y, et al. Optimization of fecal sample processing for microbiome study—The journey from bathroom to bench. *Journal of the Formosan Medical Association*. 2019;118(2):545-55.
257. Douglas CA, Ivey KL, Papanicolas LE, Best KP, Muhlhausler BS, Rogers GB. DNA extraction approaches substantially influence the assessment of the human breast milk microbiome. *Scientific Reports*. 2020;10(1):1-10.
258. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC biology*. 2014;12:1-12.
259. Kennedy KM, de Goffau MC, Perez-Muñoz ME, Arrieta M-C, Bäckhed F, Bork P, et al. Questioning the fetal microbiome illustrates pitfalls of low-biomass microbial studies. *Nature*. 2023;613(7945):639-49.



260. Poulsen CS, Ekstrøm CT, Aarestrup FM, Pamp SJ. Library Preparation and Sequencing Platform Introduce Bias in Metagenomic-Based Characterizations of Microbiomes. *Microbiology Spectrum*. 2022;10(2):e00090-22.
261. Jones MB, Highlander SK, Anderson EL, Li W, Dayrit M, Klitgord N, et al. Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proceedings of the National Academy of Sciences*. 2015;112(45):14024-9.
262. Gaulke CA, Schmeltzer ER, Dasenko M, Tyler BM, Vega Thurber R, Sharpton TJ. Evaluation of the effects of library preparation procedure and sample characteristics on the accuracy of metagenomic profiles. *Msystems*. 2021;6(5):e00440-21.
263. Sato MP, Ogura Y, Nakamura K, Nishida R, Gotoh Y, Hayashi M, et al. Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. *DNA Research*. 2019;26(5):391-8.
264. Marotz CA, Sanders JG, Zuniga C, Zaramela LS, Knight R, Zengler K. Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome*. 2018;6:1-9.
265. McIntyre AB, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome biology*. 2017;18(1):1-19.
266. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife*. 2021;10:e65088.
267. Poussin C, Khachatryan L, Sierro N, Narsapuram VK, Meyer F, Kaikala V, et al. Crowdsourced benchmarking of taxonomic metagenome profilers: lessons learned from the sbv IMPROVER Microbiomics challenge. *BMC genomics*. 2022;23(1):1-19.
268. Simon HY, Siddle KJ, Park DJ, Sabeti PC. Benchmarking metagenomics tools for taxonomic classification. *Cell*. 2019;178(4):779-94.
269. Bharti R, Grimm DG. Current challenges and best-practice protocols for microbiome analysis. *Briefings in bioinformatics*. 2021;22(1):178-93.
270. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*. 1995;57(1):289-300.
271. McGuire AL, Achenbaum LS, Whitney SN, Slashinski MJ, Versalovic J, Keitel WA, et al. Perspectives on human microbiome research ethics. *Journal of Empirical Research on Human Research Ethics*. 2012;7(3):1-14.
272. Mirzayi C, Renson A, Consortium GS, Analysis M, 84 QCSFCSS-A, Zohra F, et al. Reporting guidelines for human microbiome research: the STORMS checklist. *Nature medicine*. 2021;27(11):1885-92.
273. R Core Team R. R: A language and environment for statistical computing. 2013.
274. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*. 2014;15(3):1-12.
275. Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*. 2010;107(21):9546-51.
276. Tamburini S, Shen N, Wu HC, Clemente JC. The microbiome in early life: implications for health outcomes. *Nature medicine*. 2016;22(7):713-22.
277. Thriene K, Michels KB. Human Gut Microbiota Plasticity throughout the Life Course. *International Journal of Environmental Research and Public Health*. 2023;20(2):1463.

278. Ramirez J, Guarner F, Bustos Fernandez L, Maruy A, Sdepanian VL, Cohen H. Antibiotics as major disruptors of gut microbiota. *Frontiers in cellular and infection microbiology*. 2020;10:572912.
279. Holmes AH, Moore LS, Sundsfjord A, Steinbakk M, Regmi S, Karkey A, et al. Understanding the mechanisms and drivers of antimicrobial resistance. *The Lancet*. 2016;387(10014):176-87.
280. Raymond F, Ouameur AA, Déraspe M, Iqbal N, Gingras H, Dridi B, et al. The initial state of the human gut microbiome determines its reshaping by antibiotics. *The ISME journal*. 2016;10(3):707-20.
281. Pallav K, Dowd SE, Villafuerte J, Yang X, Kabbani T, Hansen J, et al. Effects of polysaccharopeptide from *Trametes versicolor* and amoxicillin on the gut microbiome of healthy volunteers: a randomized clinical trial. *Gut microbes*. 2014;5(4):458-67.
282. Sorbara MT, Dubin K, Littmann ER, Moody TU, Fontana E, Seok R, et al. Inhibiting antibiotic-resistant Enterobacteriaceae by microbiota-mediated intracellular acidification. *Journal of Experimental Medicine*. 2019;216(1):84-98.
283. Tedelind S, Westberg F, Kjerrulf M, Vidal A. Anti-inflammatory properties of the short-chain fatty acids acetate and propionate: a study with relevance to inflammatory bowel disease. *World journal of gastroenterology*. 2007;13(20):2826.
284. Dehoux P, Marvaud JC, Abouelleil A, Earl AM, Lambert T, Dauga C. Comparative genomics of *Clostridium bolteae* and *Clostridium clostridioforme* reveals species-specific genomic properties and numerous putative antibiotic resistance determinants. *BMC genomics*. 2016;17:1-16.
285. MacPherson CW, Mathieu O, Tremblay J, Champagne J, Nantel A, Girard S-A, et al. Gut bacterial microbiota and its resistome rapidly recover to basal state levels after short-term amoxicillin-clavulanic acid treatment in healthy adults. *Scientific reports*. 2018;8(1):11192.
286. Crits-Christoph A, Hallowell HA, Koutouvalis K, Suez J. Good microbes, bad genes? The dissemination of antimicrobial resistance in the human microbiome. *Gut Microbes*. 2022;14(1):2055944.
287. Yan W, Hall AB, Jiang X. Bacteroidales species in the human gut are a reservoir of antibiotic resistance genes regulated by invertible promoters. *npj Biofilms and Microbiomes*. 2022;8(1):1.
288. Peng Z, Jin D, Kim HB, Stratton CW, Wu B, Tang Y-W, et al. Update on antimicrobial resistance in *Clostridium difficile*: resistance mechanisms and antimicrobial susceptibility testing. *Journal of clinical microbiology*. 2017;55(7):1998-2008.
289. Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, et al. The long-term stability of the human gut microbiota. *Science*. 2013;341(6141):1237439.
290. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. *Nature*. 2012;489(7415):220-30.
291. Huang YJ. Nasopharyngeal microbiota: gatekeepers or fortune tellers of susceptibility to respiratory tract infections? : American Thoracic Society; 2017. p. 1504-5.
292. Manenzhe RI, Dube FS, Wright M, Lennard K, Zar HJ, Mounaud S, et al. Longitudinal changes in the nasopharyngeal resistome of South African infants using shotgun metagenomic sequencing. *PloS one*. 2020;15(4):e0231887.
293. Teo SM, Mok D, Pham K, Kusel M, Serralha M, Troy N, et al. The infant nasopharyngeal microbiome impacts severity of lower respiratory infection and risk of asthma development. *Cell host & microbe*. 2015;17(5):704-15.

294. Toivonen L, Schuez-Havupalo L, Karppinen S, Waris M, Hoffman KL, Camargo Jr CA, et al. Antibiotic treatments during infancy, changes in nasal microbiota, and asthma development: population-based cohort study. *Clinical Infectious Diseases*. 2021;72(9):1546-54.
295. Cantey JB, Huffman LW, Subramanian A, Marshall AS, Ballard AR, Lefevre C, et al. Antibiotic exposure and risk for death or bronchopulmonary dysplasia in very low birth weight infants. *The Journal of Pediatrics*. 2017;181:289-93. e1.
296. Ting JY, Synnes A, Roberts A, Deshpandey A, Dow K, Yoon EW, et al. Association between antibiotic use and neonatal mortality and morbidities in very low-birth-weight infants without culture-proven sepsis or necrotizing enterocolitis. *JAMA pediatrics*. 2016;170(12):1181-7.
297. Abdel Ghany EA, Ali AA. Empirical antibiotic treatment and the risk of necrotizing enterocolitis and death in very low birth weight neonates. *Annals of Saudi medicine*. 2012;32(5):521-6.
298. Cotten CM, Taylor S, Stoll B, Goldberg RN, Hansen NI, Sánchez PJ, et al. Prolonged duration of initial empirical antibiotic treatment is associated with increased rates of necrotizing enterocolitis and death for extremely low birth weight infants. *Pediatrics*. 2009;123(1):58-66.
299. Li Y, Shen RL, Ayede AI, Berrington J, Bloomfield FH, Busari OO, et al. Early use of antibiotics is associated with a lower incidence of necrotizing enterocolitis in preterm, very low birth weight infants: the NEOMUNE-NeoNutriNet cohort study. *The Journal of pediatrics*. 2020;227:128-34. e2.
300. Dethlefsen L, Relman DA. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences*. 2011;108(supplement\_1):4554-61.
301. Carr R, Borenstein E. Comparative analysis of functional metagenomic annotation and the mappability of short reads. *PloS one*. 2014;9(8):e105776.
302. Noyes NR, Weinroth ME, Parker JK, Dean CJ, Lakin SM, Raymond RA, et al. Enrichment allows identification of diverse, rare elements in metagenomic resistome-virulome sequencing. *Microbiome*. 2017;5(1):1-13.
303. Maestre-Carballa L, Navarro-Lopez V, Martinez-Garcia M. A Resistome Roadmap: From the Human Body to Pristine Environments. *Front Microbiol*. 2022;13:858831.
304. Zhang A-N, Gaston JM, Dai CL, Zhao S, Poyet M, Groussin M, et al. An omics-based framework for assessing the health risk of antimicrobial resistance genes. *Nature communications*. 2021;12(1):4765.
305. Sukumar S, Wang F, Simpson CA, Willet CE, Chew T, Hughes TE, et al. Development of the oral resistome during the first decade of life. *Nature Communications*. 2023;14(1):1291.
306. Talat A, Blake KS, Dantas G, Khan AU. Metagenomic Insight into Microbiome and Antibiotic Resistance Genes of High Clinical Concern in Urban and Rural Hospital Wastewater of Northern India Origin: a Major Reservoir of Antimicrobial Resistance. *Microbiology spectrum*. 2023;11(2):e04102-22.
307. Zhang Z, Zhang Q, Wang T, Xu N, Lu T, Hong W, et al. Assessment of global health risk of antibiotic resistance genes. *Nature communications*. 2022;13(1):1553.
308. Martínez JL, Coque TM, Baquero F. What is a resistance gene? Ranking risk in resistomes. *Nature Reviews Microbiology*. 2015;13(2):116-23.
309. Behling AH, Wilson BC, Ho D, Virta M, O'Sullivan JM, Vatanen T. Addressing antibiotic resistance: computational answers to a biological problem? *Current Opinion in Microbiology*. 2023;74:102305.

310. Dantas G, Sommer MO. Context matters—the complex interplay between resistome genotypes and resistance phenotypes. *Current opinion in microbiology*. 2012;15(5):577-82.
311. Kent AG, Vill AC, Shi Q, Satlin MJ, Brito IL. Widespread transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial Hi-C. *Nature Communications*. 2020;11(1):4379.
312. Serpa PH, Deng X, Abdelghany M, Crawford E, Malcolm K, Caldera S, et al. Metagenomic prediction of antimicrobial resistance in critically ill patients with lower respiratory tract infections. *Genome medicine*. 2022;14(1):74.
313. Leggett RM, Alcon-Giner C, Heavens D, Caim S, Brook TC, Kujawska M, et al. Rapid MinION profiling of preterm microbiota and antimicrobial-resistant pathogens. *Nature Microbiology*. 2020;5(3):430-42.







## Differential response to prolonged amoxicillin treatment: long-term resilience of the microbiome versus long-lasting perturbations in the gut resistome

Achal Dhariwal, Lars Christian Haugli Bråten, Kjersti Sturød, Gabriela Salvadori, Ahmed Bargheet, Heidi Åmdal, Roger Junges, Dag Berild, John-Anker Zwart, Kjersti Storheim & Fernanda Cristina Petersen

To cite this article: Achal Dhariwal, Lars Christian Haugli Bråten, Kjersti Sturød, Gabriela Salvadori, Ahmed Bargheet, Heidi Åmdal, Roger Junges, Dag Berild, John-Anker Zwart, Kjersti Storheim & Fernanda Cristina Petersen (2023) Differential response to prolonged amoxicillin treatment: long-term resilience of the microbiome versus long-lasting perturbations in the gut resistome, Gut Microbes, 15:1, 2157200, DOI: [10.1080/19490976.2022.2157200](https://doi.org/10.1080/19490976.2022.2157200)

To link to this article: <https://doi.org/10.1080/19490976.2022.2157200>



© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC.



[View supplementary material](#)



Published online: 28 Dec 2022.



[Submit your article to this journal](#)



Article views: 1896



[View related articles](#)



[View Crossmark data](#)





## Differential response to prolonged amoxicillin treatment: long-term resilience of the microbiome versus long-lasting perturbations in the gut resistome

Achal Dhariwal<sup>a</sup>, Lars Christian Haugli Bråten<sup>b</sup>, Kjersti Sturød<sup>a</sup>, Gabriela Salvadori<sup>a</sup>, Ahmed Bargheet<sup>a</sup>, Heidi Åmdal<sup>a</sup>, Roger Junges<sup>c,d</sup>, Dag Berild<sup>c,d</sup>, John-Anker Zwart<sup>b,d</sup>, Kjersti Storheim<sup>b,e</sup>, and Fernanda Cristina Petersen<sup>b</sup>

<sup>a</sup>Institute of Oral Biology, Faculty of Dentistry, University of Oslo, Oslo, Norway; <sup>b</sup>Department of Research and Innovation, Division of Clinical Neuroscience, Oslo University Hospital HF, Oslo, Norway; <sup>c</sup>Department of Infectious Diseases, Oslo University Hospital HF, Oslo, Norway; <sup>d</sup>Institute of Clinical Medicine, University of Oslo, Oslo, Norway; <sup>e</sup>Department of Physiotherapy, Faculty of Health Science, Oslo Metropolitan University, Oslo, Norway

### ABSTRACT

The collateral impact of antibiotics on the microbiome has attained increasing attention. However, the ecological consequences of long-term antibiotic exposure on the gut microbiome, including antibiotic resistance, are still limited. Here, we investigated long-term exposure effects to amoxicillin on the human gut microbiome and resistome. Fecal samples were collected from 20 patients receiving 3-months of amoxicillin or placebo treatment as part of a Norwegian multicenter clinical trial on chronic low back pain (AIM study). Samples were collected at baseline, last day of treatment, and 9 months after antibiotic cessation. The abundance and diversity of microbial and resistome composition were characterized using whole shotgun and functional metagenomic sequencing data. While the microbiome profiles of placebo subjects were stable over time, discernible changes in diversity and overall microbiome composition were observed after amoxicillin treatment. In particular, health-associated short-chain fatty acid producing species significantly decreased in proportion. However, these changes were short-lived as the microbiome showed overall recovery 9 months post-treatment. On the other hand, exposure to long-term amoxicillin was associated with an increase in total antimicrobial resistance gene load and diversity of antimicrobial resistance genes, with persistent changes even at 9 months post-treatment. Additionally, beta-lactam resistance was the most affected antibiotic class, suggesting a targeted response to amoxicillin, although changes at the gene level varied across individuals. Overall, our results suggest that the impact of prolonged amoxicillin exposure was more explicit and long-lasting in the fecal resistome than in microbiome composition. Such information is relevant for designing rational administration guidelines for antibiotic therapies.

### ARTICLE HISTORY

Received 26 August 2022  
Revised 27 October 2022  
Accepted 30 November 2022

### KEYWORDS

Antimicrobial resistance; microbiome; resistome; antimicrobial resistance gene; amoxicillin; antibiotics; chronic low back pain; Modic changes; functional metagenomics; shotgun

## Introduction


Global health-care systems are under severe threat due to the increasing prevalence of antibiotic-resistant pathogens, causing treatment failure, higher mortality, and increased economic burden.<sup>1,2</sup> Imprudent use of antibiotics is one of the significant drivers for the emergence and spread of antibiotic resistance.<sup>3–6</sup> Increased consumption of antibiotics may not only affect us at the individual patient level but may also lead to greater resistance at regional, national and global levels.<sup>7</sup>

The advancements in next-generation sequencing technologies have led to an increased understanding of the gut microbiome and its role in modulating human health and physiology.<sup>8,9</sup> Our gut microbiome

is a complex ecosystem of microbial communities and an important well-known reservoir for a vast number of antibiotic resistance genes (ARGs). Moreover, it has been shown to contribute to the spread of resistance by promoting the horizontal gene transfer (HGT) or exchange of ARGs to opportunistic pathogenic bacteria.<sup>10–13</sup> Since the human gut microbiome is a crucial player in the emergence and dissemination of antibiotic resistance, it is important to characterize the landscape of antibiotic resistance (resistome) in this microbial environment to contribute to personalized antimicrobial stewardship strategies.

Antibiotics can target not only the infective pathogen but also commensal and opportunistic bacteria inhabiting the human gut. Such collateral off-target

**CONTACT** Fernanda Cristina Petersen  [fcpetersen@odont.uio.no](mailto:fcpetersen@odont.uio.no)  Institute of Oral Biology, Faculty of Dentistry, University of Oslo, Oslo, Norway

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/19490976.2022.2157200>

© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

effects can disrupt the composition or diversity of the gut ecosystem and are associated with significant health consequences.<sup>14–17</sup> Antibiotic administration can also influence the gut microbiome by selecting antibiotic-resistant bacteria,<sup>18</sup> increasing the abundance of particular ARGs<sup>19,20</sup>, and altering the resistome composition.<sup>21,22</sup> Antibiotic exposure not only gives rise to resistance to the antibiotic used but also to other classes of antibiotics via cross-resistance or selection of co-localized genes that confer resistance to multiple antibiotics.<sup>23,24</sup> In addition, the use of antibiotics also affects the dynamics of overall HGT, and thus enables the global spread of resistance by promoting the dissemination of ARGs located on mobile genetic elements (MGEs) within and between bacterial species.<sup>25</sup> These disruptive effects on the microbiome and resistome can be transient or persistent, depending upon the antibiotics class, mode of action, dosage, duration, pharmacokinetic properties, and baseline taxonomic composition.<sup>15,18,22,26–28</sup>

Amoxicillin is one of the most commonly prescribed antibiotics worldwide, used to treat common bacterial infections in the ear, nose and throat.<sup>29</sup> It is a broad-spectrum beta-lactam antibiotic that interferes with bacterial cell wall synthesis. So far, several culture and sequencing-based studies have shown that the adult human gut microbiome is generally resilient toward amoxicillin intervention, with little or no impact on its microbial composition.<sup>30,31</sup> However, these studies are largely limited to taxonomic characterization and cannot elucidate the repertoire of ARGs, as they are mainly based on 16S rRNA gene amplicon sequencing. For identifying the landscape of antibiotic resistance (repertoire of ARGs) in microbiomes, whole metagenomic sequencing stands as the most powerful high-throughput method. This approach was recently used to assess the impact of 5-days amoxicillin therapy on the adult microbiome and its associated resistome. Although the results indicated a substantial enrichment of multiple classes of ARGs in the human gut due to amoxicillin treatment, most of the results were based on predictive analysis, and only one sample was analyzed using whole metagenomic sequencing.<sup>32</sup> The impact of different class of antibiotics has also been examined, but primarily focusing on short-term (<7 days) treatments.<sup>18,20–22,30,32,33</sup> For prolonged antibiotic

therapy regimens, the effects on the human gut microbiome and resistome remain largely unknown.

Conditions for which prolonged amoxicillin treatment are generally used include recurrent otitis media, prophylaxis of urinary tract infection, and chronic respiratory conditions. Long-term amoxicillin treatment is also used to treat patients with chronic low back pain and vertebral endplate changes (Modic changes) visible on magnetic resonance imaging. The underlying rationale is a theory that Modic changes are often due to a low-grade bacterial infection.<sup>34</sup> However, a recent multicenter, randomized, double-blind, placebo-controlled trial (the AIM study)<sup>35</sup> concluded that the 3 months of amoxicillin use did not provide a clinically significant benefit in patients with chronic low back pain and Modic changes. Such prolonged antibiotic therapies are of special concern as the pressure in selecting and promoting the development of antibiotic resistance is expected to be higher than for short-term therapies.

In the present study, we have utilized next-generation deep shotgun sequencing and functional metagenomics to investigate the prolonged consequences of long-term antibiotic treatment on the gut microbiome and resistome of patients with chronic low back pain and Modic changes as compared to placebo. Here, we have examined the changes in the abundance, diversity and composition of microbial taxa and ARGs as well as explored the potential associations between them by analyzing the fecal samples from the subset of patients included in the AIM study. Overall, our results suggest that the impact of amoxicillin was severe yet short-lived in the microbiome composition compared to the changes in the resistome, which were more explicit and persistent, leading to a more diverse and abundant resistome even after 9 months post-antibiotic exposure.

## Materials and methods

### Ethics statement

The clinical trial was ethically approved by the regional Committees for Medical Research Ethics in Norway (2014/158/REK sør-øst) and the Norwegian Medicines Agency (SLV, reference

number 14/01368-11, EudraCT Number: 2013-004505-14). It was registered at ClinicalTrials.gov (NCT02323412) on December 23, 2014. The risks and benefits of the study were explained, and written informed consent was obtained from all the participants.

### **Study design and participants**

To investigate the impact of prolonged broad-spectrum antibiotic (amoxicillin) exposure on the gut microbiome and resistome, we have included 20 patients who were randomly assigned to either amoxicillin ( $n = 8$ ) or the placebo ( $n = 12$ ) group (Table 1). The patients were recruited from a larger double-blind, placebo-controlled, randomized, multicenter trial (The AIM study), testing the efficacy of antibiotic treatment in 180 patients with chronic low back pain and Modic changes.<sup>35</sup> The 20 selected patients were those enrolled at two of the study centers, Oslo University Hospital and Østfold Hospital, from April 2016 to September 2017. Fourteen patients did not provide samples for all time points and an additional two patients in the placebo group took antibiotics during the trial. These were excluded from the study. In this study, patients were relatively healthy adult human subjects aged between 27 and 62 y (mean  $\pm$  SD:  $45 \pm 11$ ), 15 (68%) were women, and their BMI ranged from 18 to  $35 \text{ kg/m}^2$  (mean  $\pm$  SD:  $24 \pm 4$ ). At baseline, 6 patients (27%) smoked while 8 (36%) were former smokers. All 20 patients in the amoxicillin group were compliant, i.e., took more than 80% of their study medication. Two patients in the amoxicillin group and four in the placebo group had gastroenteritis/diarrhea during the 3-month treatment period. The detailed inclusion and exclusion criteria for the main trial have already been described in the previously published protocol.<sup>36</sup> In brief, participants had not taken any antibiotics 1 month before treatment intervention, nor did they travel abroad for a period exceeding 4 weeks. These patients were treated thrice a day with an oral dose of 750 mg amoxicillin or placebo for 3 months. All patients and research staff were blinded, and study medication was encapsulated to secure equal taste and smell. Fecal samples were collected at three time points: before the antibiotic treatment (baseline), post-treatment (3 months after the start of intervention), and 9 months (12 months from baseline) after

the cessation of the treatment. In total, 60 fecal samples were collected in sterile tubes and stored at  $-80^\circ\text{C}$  within 24 h until further processing and analysis. The basic metadata information for samples is also provided as Supplementary Table S8.

### **DNA isolation and metagenomic sequencing**

The microbial DNA from the fecal samples was extracted manually using the PSP® Spin Stool DNA Kit (Stratec molecular, Berlin, Germany) as per the manufacturer's protocol. The quantity and quality of extracted DNA were assessed using a NanoDrop™ 2000c spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA), Qubit (Thermo Fisher Scientific) and agarose gel electrophoresis. Libraries for shotgun metagenomic sequencing were prepared using the Illumina Nextera Flex DNA library prep kit according to the manufacturer's instructions. The DNA libraries were then sequenced with 150-nucleotide-long paired-end reads on an Illumina HiSeq 3000 platform (Illumina Inc., San Diego, CA, USA) at the Norwegian Sequencing Center (Oslo, Norway).

### **Bioinformatics preprocessing**

Low-quality and adapter sequences from paired-end reads were filtered using Trimmomatic<sup>37</sup> (v.0.35) with the following parameters: ILLUMINACLIP: Nextera PE:2:30:10 LEADING:3 TRAILING:3 SLIDING WINDOW:4:15 MINLEN:36. After quality filtering, the human DNA contaminant sequences were discarded from all samples by filtering out the reads that mapped against the human reference genome (GRCh38, downloaded from NCBI GenBank) using Bowtie2<sup>38</sup> (v.2.3.4) with parameters `q -N 1 -k 1 -fr -end-to-end -phred33 -very-sensitive -no-discordant`. The quality of raw and clean reads was assessed using FastQC<sup>39</sup> (v.0.11.8).

### **Microbiome and resistome profiling**

To investigate the microbial community composition, the clean, high-quality reads were subjected to taxonomic classification using MetaPhlAn 3.0<sup>40</sup> containing ~17000 reference genomes of bacteria, archaea, eukaryotes and viruses. The resistome annotation of metagenomic reads was performed

**Table 1.** Patient Demographics table.

	Amoxicillin (n = 8)	Placebo (n = 12)
Age, mean (SD)	41.0 (11)	46.6 (10)
Women (%)	7 (88%)	8 (67%)
BMI, mean (SD)	24.6 (4.7)	24.4 (4.2)
Smoking (%)		
Current smokers	1 (13%)	5 (42%)
Former smokers	4 (50%)	4 (33%)
Never smoked	3 (38%)	3 (25%)
missing	0	2 (17%)
Medications other than study medication (%)		
Probiotics	6 (75%)	2 (17%)
Lipid-lowering drugs	1 (13%)	1 (8%)
Adverse events (%)		
Diarrhoea/gastroenteritis during 0–3 months' follow-up	2 (25%)	4 (33%)
Diarrhoea/gastroenteritis during 3–12 months' follow-up	3 (38%)	1 (8%)

by mapping them against the nucleotide\_fasta\_protein\_homolog\_model from the Comprehensive Antibiotic Resistance Database (CARD)<sup>41</sup> (v.3.0.9) antimicrobial resistance (AMR) database using Bowtie2 (v.2.3.4) with parameter *-very-sensitive-local*. Additionally, the ARGs that are not confidently expected to confer resistance based solely on a short-read marker were removed from further analysis as described by D'Souza et al.<sup>42</sup> For annotation of ARGs, a coverage threshold of 80% (100% nucleotide homology with reference across 80% of target gene) was used. The mapped reads were filtered from unmapped reads, sorted and indexed using Samtools<sup>43</sup> (v.1.9). The number of read counts mapped for each ARG was calculated using Bedtools<sup>44</sup> (v.2.27.1). The counts were then normalized for differences in both gene lengths and bacterial abundances by calculating reads per kilobase reference per million bacterial reads (RPKM) values for every sample. The relative abundance of ARGs for each sample was estimated by dividing the RPKM by the sum of the RPKM for each sample. In addition to gene level, we also summed the RPKMs to higher functional levels, as annotated in their respective databases.

### Functional metagenomics

Three functional metagenomic libraries were constructed from pooled stool DNA of amoxicillin-treated patients, one for each time point, according to the published protocol,<sup>45</sup> with minor modifications. In brief, each metagenomic library was cloned

into pZE21 (kanamycin-resistant) and screened individually on Mueller–Hinton agar plates against 14 antibiotics (amoxicillin–clavulanate, aztreonam, carbenicillin, cefepime, cefoxitin, ceftazidime, colistin, meropenem, penicillin G, piperacillin, tigecycline, amoxicillin, ciprofloxacin and gentamicin), using the selective concentrations listed in Supplementary Table S2. All plates were incubated aerobically at 37°C for 24 h. In addition, a negative control plate of *E. coli* clones transformed with unmodified pZE21 (without metagenomic insert) was plated for each antibiotic selection to confirm that the concentration of antibiotic used completely inhibited the growth of clones with only pZE21. The functionally selected colonies were subsequently collected from the agar plate using an L-shaped cell-spreader and stored at – 80°C. To isolate the metagenomic inserts, the frozen stocks were then thawed, subjected to cell lysis, and pelleted by centrifugation at 15,000 × g for 2 min. The supernatants were then collected and used as a template for PCR amplification of metagenomic DNA fragments. These were further purified using the MinElute PCR Purification Kit (Qiagen) and prepared for sequencing using the NEXTFLEX® Rapid DNA-Seq Kit 2.0 (PerkinElmer).

The samples were sequenced on an Illumina NovaSeq platform (2 × 150 bp reads), and the quality of sequenced reads was accessed using FastQC. Similar to shotgun metagenomic data, quality filtering of sequenced reads was also performed using Trimmomatic with a slight change in the MINLEN:60 parameter. These quality-trimmed



reads for each selection were assembled into contigs using metaSPAdes<sup>46</sup> (v.3.13.0). Quality assessment of assembled contigs was done using MetaQUAST from QUAST<sup>47</sup> (v.5.0.2) with (-m 1000) parameter. Notably, the selections were excluded from the analysis if the number of contigs assembled was more than tenfold of the number of colonies on the selection plate. Contigs with a length shorter than 500 bp were also filtered out. Finally, the open reading frames (ORFs) were predicted from the remaining contigs using the Prodigal<sup>48</sup> (v.2.6.3) software. These ORFs were annotated hierarchically by searching them first against the BLAST-based (ResFinder,<sup>49</sup> CARD,<sup>41</sup> AMRFinder-Prot<sup>50</sup>) ARG databases, and then the residual ORFs were annotated using HMM-based (AMRFinder-hmm,<sup>50</sup> Resfams<sup>51</sup>) ARG databases using publicly available *resAnnotator.py* pipeline (<https://github.com/dantaslab/resAnnotator>). The objective for such sequential annotation of ARGs was to first identify perfect hits to known resistance determinants through BLAST-based databases (with a high percent identity (>95%) and coverage (>95%) threshold) and afterward detect the variants of known resistance determinants using HMM-based databases. Further, FASTA files of all the annotated ARG sequences from all antibiotic selections were concatenated and perfectly identical ARGs were merged into one sequence using CD-HIT<sup>52</sup> (v.4.8.1) with options: -c 1.0 -aS 1.0 -g 1 -d 0. The percent identity of all ARGs was determined using BlastP query against both a custom database that combined all ARG proteins from CARD and AMRFinder, and the database provided by NCBI for non-redundant proteins (nr database, retrieved January 25, 2022). The best local hit identified was then used for a global alignment using the needle program with the following non-default parameters: -gapopen-10 -gapextend = 0.5. The detailed protocol on functional validation of novel resistance gene along with its amino acid and nucleotide sequence is presented in Additional file S2.

The putative MGEs were detected from functional metagenomic assemblies through string searches to one of the following keywords in their Pfam and TIGRFAMs annotations: 'transposon', 'plasmid', 'transposase', 'integron', 'conjugative', 'integrase', 'recombinase', 'conjugal', 'mobilization' or 'recombination', as previously described.<sup>53</sup>

### Statistical analysis and data visualization

Statistical analysis was accomplished in R (v.3.6.4). The graphical illustrations were mainly made using ggplot2 (v.3.3.5) R package with post-editing in Adobe Illustrator (v.16.0.0). Diversity analysis ( $\alpha$ - and  $\beta$ -diversity) were performed using the vegan (v.2.5.7) and phyloseq (v.1.34.0) packages. The changes in  $\alpha$ -diversity over time points in groups were statistically evaluated using linear mixed-effects (LME) model and Tukey's HSD post hoc test using R nlme package.  $\beta$ -Diversity was conducted on centered log-ratio (CLR) transformed species and ARG relative abundance data, and principal component analysis (PCA) ordination technique was used to further understand which metadata variables were associated with the dispersion of samples microbiome and resistome compositions, respectively. Permutational multivariate analysis of variance (PERMANOVA) test by adonis function (vegan R package) was used to determine the statistical significance of  $\beta$ -diversity. The homogeneity of multivariate dispersion between treatment groups and time points was evaluated prior by permutational analysis of multivariate dispersion (PERMDISP) test with function betadisper (vegan). The dispersion of the data was homogenous at all time points within the treatment groups. Patient\_ID was used as a random effect or strata in the models to account for inter-individual variation present in the dataset. The compositional shifts were quantified and compared across individuals or pairs using Bray-Curtis and Jaccard distance metrics. The Jaccard distance accounts for the presence/absence of features (species or ARGs) and is thus more sensitive to changes driven by rare features. While Bray-Curtis distance accounts for differences in the relative abundance of features. For comparisons of group differences, Wilcoxon test using wilcox.test function was applied with paired = T/F, as appropriate. Differentially abundant features between time points within treatment groups for all the datasets were identified using LEfSe<sup>54</sup> algorithm. Enterotypes were identified using the Dirichlet multinomial mixture (DMM) models approach

using the DirichletMultinomial (v.1.38.0) package. Procrustes analysis was performed to determine the association or effect of underlying bacteriome on the resistome. The symmetric Procrustes correlation coefficients and significance were retrieved through the 'protest' function from the vegan package with 999 permutations. Heatmaps and synteny maps were generated using the Pheatmap (v.1.0.12) and Gggenes (v.0.4.1) R package, respectively. The *P*-values below or equal to 0.05 were considered significant and corrected for false discovery rate (FDR) where appropriate using Benjamini–Hochberg (BH) method.

## Results

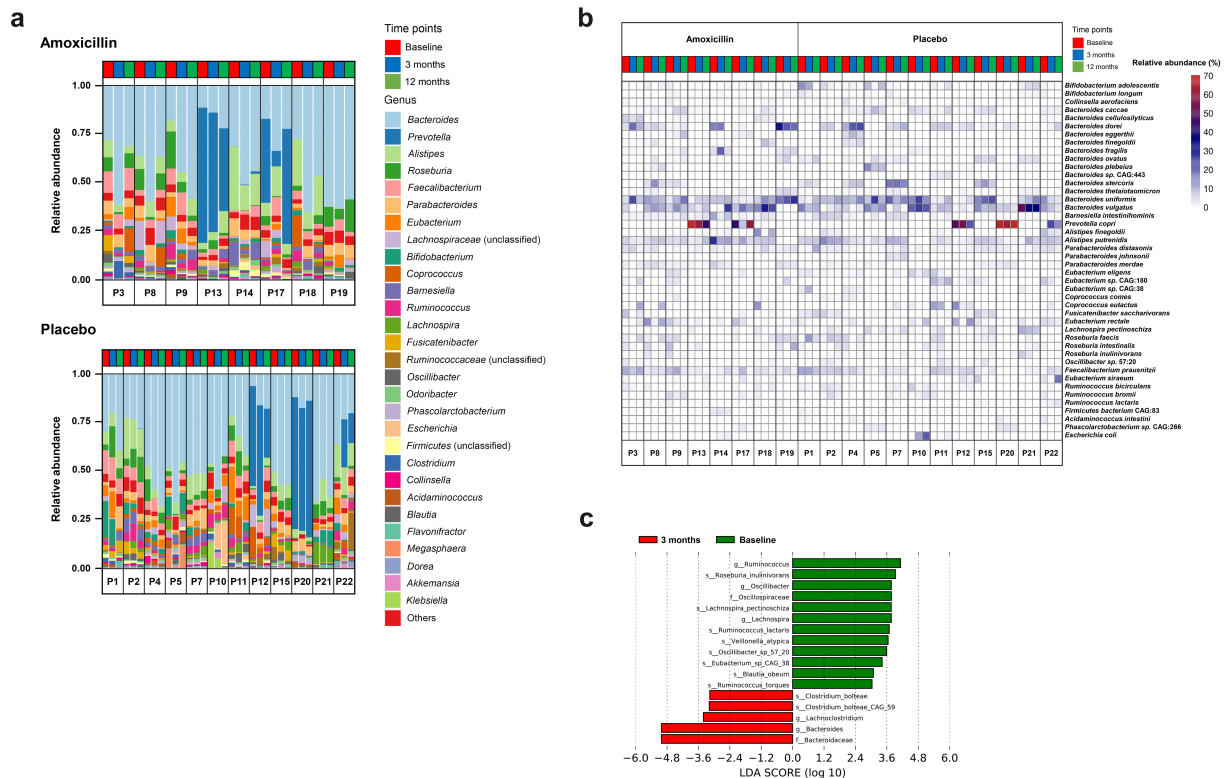
### **Detrimental short-term alterations in the diversity and composition of the human gut microbiome upon prolonged amoxicillin therapy**

To examine the post-treatment and long-term taxonomic response to prolonged amoxicillin exposure, total DNA was extracted from fecal samples and sequenced on Illumina platform, resulting in ~200 Gb of raw sequencing data. From 60 fecal samples, we obtained a total of ~1.5 billion raw sequences with a number of reads ranging from 11.3 to 41.6 million (M) per sample with an average of 25.1 M reads. Overall, 49.5% (0.72 billion in total) of high-quality clean sequences with a number of reads ranging from 5.07 to 23.2 M per sample with an average of 12 M reads were classified as bacterial (Supplementary Table S1). In most of the samples, Bacteroidetes (mean 65.31%; SD 15.75%) and Firmicutes (mean 29.53%; SD 13.94%) were the most abundant phyla, followed by Actinobacteria (mean 2.94%; SD 3.57%) and Proteobacteria (mean 1.72%; SD 4.22%). These four phyla had a combined relative abundance of 99.5% across samples. At genus level, the microbiome composition was primarily dominated by *Bacteroides* (mean 39.06%; SD 17.68%), with *Prevotella* (mean 11.04%; SD 22.40%) and *Alistipes* (mean 7.87%; SD 6.37%) being the second and third most abundant genera. *Parabacteroides*, *Eubacterium*, *Blautia*, *Roseburia*

and *Faecalibacterium* were the other most prevalent genera identified in all the patients at all time points. In total, 118 different genera (range: 41–71 per patient) were identified, with 16 common genera present across all patients.

The overall individual microbial profiles did not differ substantially following antibiotic exposure, and most changes observed varied from individual to individual (Figure 1a & Supplementary Table S6). However, a common response to amoxicillin was observed in the treatment group. Indeed, the genera *Ruminococcus*, *Oscillibacter* and *Lachnospira* were decreased considerably in their relative abundances after amoxicillin intake, while the proportion of *Bacteroides* and *Lachnospira* increased significantly in majority of the individuals. At higher taxonomic levels, we also observed changes in the proportion of two bacterial families, i.e., Bacteroidaceae and Oscillospiraceae, upon amoxicillin treatment (Supplementary Figure 1). In contrast, no substantial changes at genus and higher taxonomic levels were identified in the overall gut microbiome composition between any time points in the placebo group, suggesting the relative stability of the gut microbiome without antibiotic perturbations. Despite differences, enterotyping analysis using Dirichlet multinomial mixture (DMM) model dichotomized all the fecal samples into *Bacteroides* and *Prevotella* dominant genera based on the similarities in the composition of gut microbiome (Supplementary Figure 2).

The dominant species detected in the gut microbiome were typically associated with 1 of the 15 species belonging to the *Bacteroides*, *Faecalibacterium*, *Prevotella*, *Ruminococcus*, *Eubacterium*, *Roseburia*, *Alistipes*, *Coprococcus* and *Parabacteroides* genera (Figure 1b). *Faecalibacterium prausnitzii*, *Bacteroides vulgatus*, *Bacteroides uniformis*, *Alistipes putredinis* and *Parabacteroides distasonis* were among the most common and abundant species found across most of the patients. At the same time, some of the other bacterial species such as *Prevotella copri*, *Bacteroides dorei* and *Bacteroides stercoris* were predominantly present in some fecal samples. These findings were coherent with previously published studies, indicating the existence of a global core gut microbiome.<sup>55</sup> Nevertheless, the changes in the microbiome composition at the species level were mostly specific for each



**Figure 1.** Composition of bacterial communities in all patients before exposure (baseline), 3 months later and 9 months after cessation of treatment (i.e., 12 months). (a) Bar plot representing the relative abundance of top 30 most abundant bacterial genera in amoxicillin (top) and placebo-exposed (bottom) participants. Genera with lower abundance were merged into “Others” category. Source Data: Supplementary Table S6. (b) Heat map representing the relative abundance of bacterial species that constituted >1% of the community in at least two samples. P precedes the identification number of the patient. (c) Histogram of LDA scores of bacterial taxa that were significantly different in abundance between baseline and 3 months in amoxicillin group identified using linear discriminant analysis (LDA) combined with effect size (LEfSe) algorithm, are shown, with a cutoff value of LDA score (log<sub>10</sub>) above 2.0. Bacterial taxa enriched in baseline are indicated with a positive LDA score (green) and taxa enriched immediately after the cessation of treatment (3 months) are represented by a negative score (red).

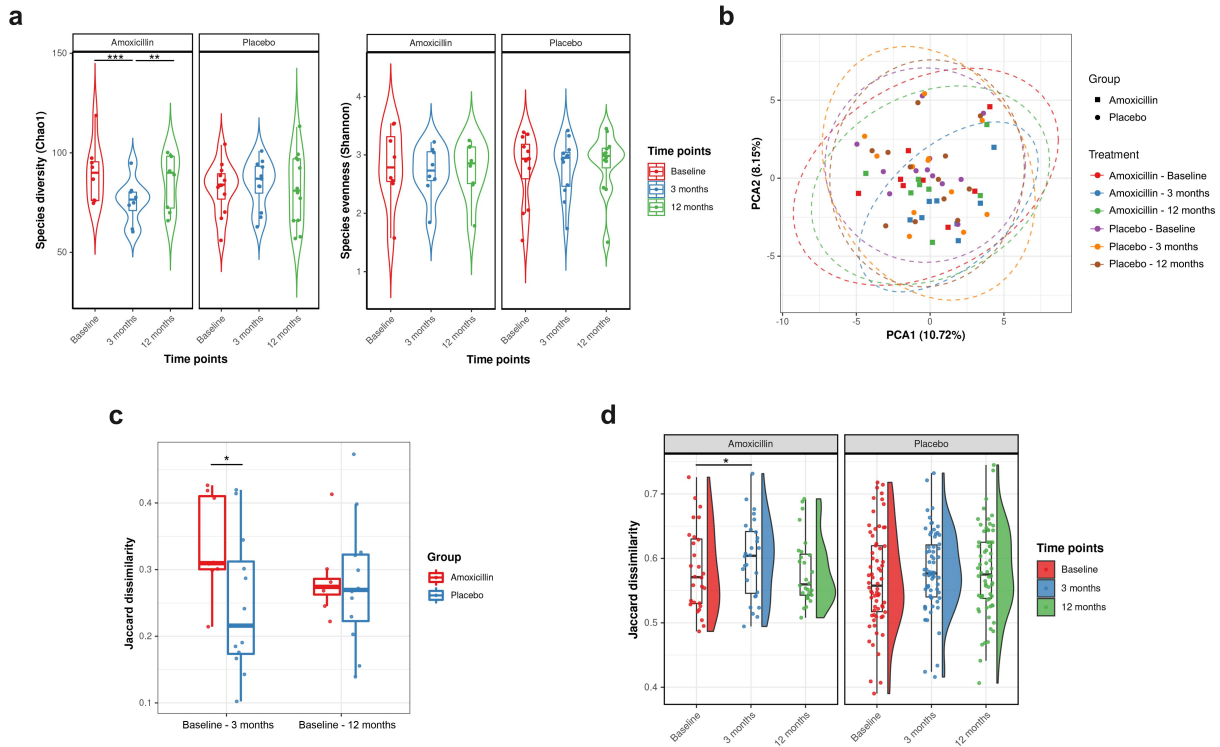
individual, and no major changes in the dominant species were observed after exposure to antibiotics.

Next, we characterized the alteration in  $\alpha$ -diversity using richness (Chao1) and evenness (Shannon) measures. Amoxicillin administration led to a significant reduction in species richness immediately after treatment cessation (3 months) as compared to placebo (LME: adj  $P = 1.94e-05$ ), whereas evenness was not significantly reduced (adj  $P = .20$ ). The species richness returned to baseline levels after 9 months post-treatment (Figure 2a). Contrarily, no significant alteration in Chao1 and Shannon diversity index was observed in the placebo group at all time points, revealing the overall stability of  $\alpha$ -diversity over time.

Then, to assess the overall changes in the microbial community composition and structure,  $\beta$ -diversity analysis using principal component analysis (PCA) ordinate with Euclidean distance was

conducted. PCA plot reveals a significant shift in the microbiome composition immediately after cessation of amoxicillin intervention (3 months) (PERMANOVA:  $R^2 = 0.070$ ;  $P = .007$ ) (Figure 2b). However, the microbial shifts were not discernible after 9 months post-treatment (PERMANOVA:  $R^2 = 0.02$ ;  $P = .25$ ). In addition, PERMANOVA analysis showed that the inter-individual differences in microbiome profile is greater than the effect of antibiotics as the samples from same individual tend to primarily cluster together (Supplementary Figure 3A, PERMANOVA:  $R^2 = 0.79$ ;  $P = .0001$ ;  $nperm = 999$ ). No significant alterations in the overall bacterial community composition were observed at any time point in the placebo group.

Microbial compositional shifts between the baseline and each of the other two time-point samples within each individual were quantified using



**Figure 2.** Impact of amoxicillin on fecal microbiota diversity and composition. (a) Violin plots reporting the species  $\alpha$ -diversity measured and compared over time points in amoxicillin and placebo groups using Chao1 (richness; left) and Shannon (evenness; right) index, respectively. Each point corresponds to a given sample, and each box span from the first to third quartiles with a horizontal line inside the boxes representing the median. Adjusted  $P$  values were computed using the LME mixed effect model and the Tukey's HSD post hoc test. Adjusted  $P$  values ( $P$ ): \*\*\* $P < .001$ , \*\* $P < .01$ ; \* $P < .05$ . (b) Principal component analysis (PCA) based on the centered-log ratio (clr) transformed species abundance matrix. Each point represents the bacterial microbiome of an individual sample. Different symbols indicate different treatment groups; colors indicate different time points in different treatment groups. Ellipses represent 95% confidence intervals (CI) around the group clustered centroid. (c) Dissimilarity in microbiome composition between the baseline and the other time points, i.e., 3 and 12 months. Each point corresponds to Jaccard dissimilarity calculated between baseline and each of the other time point samples of the respective individual. Black horizontal line on the top connects statistically significantly different groups within each visit pair (\* $P < .05$ ; Wilcoxon rank-sum test). (d)  $\beta$ -Diversity boxplots showing the distribution of the Jaccard dissimilarity in microbiome profile between individuals at the same time point within each treatment group. Each point is a comparison between two samples within the same time point group. The distributions are displayed to the right of the points, and boxplots showing the median and interquartile ranges are superimposed on top of the points (Statistical significance: paired Wilcoxon test ( $P$  values ( $P$ ): \* $P < .05$ , \*\* $P < .01$  and \*\*\* $P < .001$ ).

Jaccard distance to determine the extent of alteration upon amoxicillin intervention. We observed that the average compositional differences between baseline and 3-month samples were significantly larger in the amoxicillin than in the placebo group (0.33 vs. 0.24, respectively,  $P = .042$ , Wilcoxon rank-sum test) (Figure 2c), though the composition in the amoxicillin group did not differ significantly from the placebo group at 12 months (0.28 vs. 0.27, respectively,  $P = .73$ , Wilcoxon rank-sum test).

We also investigated whether the gut microbiome presents a more common than an individualized response to amoxicillin treatment by targeting similar sets of species and converging the

microbiome profile across individuals. In such case, we would expect the microbiome profiles to become more similar after amoxicillin treatment across individuals as a result of the selective stress imposed by antibiotics. However, the results from our analysis revealed that the average dissimilarity of the microbiome composition measured by the Jaccard distance between individuals increased significantly immediately after the amoxicillin treatment (0.57 at baseline vs. 0.61 at 3 months,  $P = .047$ , Wilcoxon signed rank exact test), but 9 months after the cessation of the amoxicillin, microbiome profiles converged to baseline levels (Figure 2d). No significant increase in microbiome divergence was



observed in the placebo group over time. In general, these findings indicate that the overall composition of the microbiome diversified due to prolonged amoxicillin exposure, and that such effect persisted only for a short term.

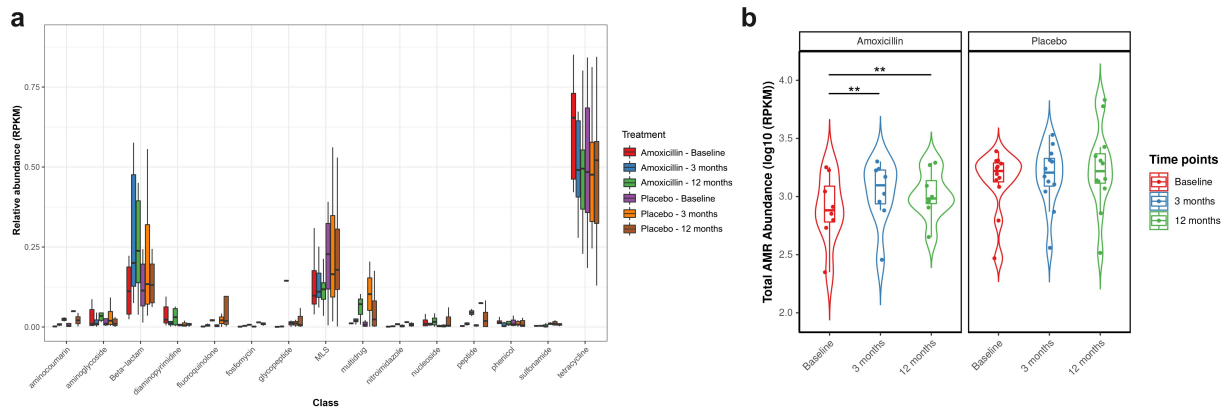
Lastly, we explored the effect of treatment on individual microbial taxa. In concordance with the results regarding the alteration in richness and composition described above, significant differences were observed among species with low abundance. At the species level, well-known short-chain fatty acids (including butyrate) producers, such as *Roseburia inulinivorans*, *Ruminococcus torques*, *Lachnospira pectinoschiza*, *Ruminococcus lactaris*, *Veillonella atypica* and *Blautia obeum* were significantly reduced immediately after exposure to amoxicillin as compared to baseline (LEfSe). More recently discovered, “healthy gut” marker bacteria, i.e., *Oscillibacter* sp. 57\_20, was also found to be differentially decreased in abundance after cessation of treatment. Most of these species belong to phylum Firmicutes and class Clostridia (except *Veillonella atypica*) and are part of the normal gut microbiome. They play a crucial role in maintaining homeostasis, inflammation and oxidative stress in the human gut. However, all these species except *Lachnospira pectinoschiza* almost returned to their baseline abundance levels at 9 months post-treatment. In contrast, two species associated with overall poor gut health, i.e., *Clostridium bolteae* and *Clostridium bolteae* CAG 59,<sup>56</sup> were increased in most individuals exposed to amoxicillin and returned to baseline at the 12 months sampling time point (Figure 1c, Supplementary Figure 1 & Table S3). However, for *C. bolteae* CAG 59, but not *C. bolteae*, we also observed an increase in its proportion for some of the patients (4 out of 12) in the placebo group. In general, the microbiome composition in the placebo group remained stable, as no other microbial taxa were significantly increased or decreased over time.

### **Long-term changes in the abundance and diversity of the gut resistome upon prolonged amoxicillin therapy**

In total, we detected 147 unique ARGs conferring resistance to 15 classes of antibiotics via 5 resistance mechanisms in the fecal microbiome of all the

studied participants using shotgun metagenomic data (Supplementary Table S7). Tetracycline (mean 64.59%; SD 18.47%), beta-lactam (mean 14.10%; SD 13.45%), macrolide-lincosamide-streptogramin (MLS) (mean 12.48%; SD 13.40%), aminoglycoside (mean 1.78%; SD 6.35%) and multidrug (mean 2.36%; SD 6.35%) were the most abundant class of ARGs across samples, with the *tetQ*, *tetW*, *tetO*, *cfxA*, *mef(En2)* and *ermF* genes comprising the majority of alignments within these classes (Figure 3a & Supplementary Figure 4). The total AMR abundance measured as the total reads per kilobase million bacterial reads (RPKM) of all ARGs in the gut was lower in the amoxicillin than in the placebo group at baseline. Notably, the total AMR abundance significantly increased over time after amoxicillin treatment, suggesting that antibiotic treatment may enrich the microbiome for AMR determinants (Figure 3b). This increase in abundance corresponded with the increase in  $\alpha$ -diversity evenness (Shannon) of ARGs in the resistome at 3-month timepoint (LME: adj  $P = .02$ ), which remained significantly higher after 9 months post-treatment (LME: adj  $P = .01$ ). A similar increase in ARG richness (Chao1) was observed after amoxicillin exposure, though the trend was not statistically significant. In contrast, no significant differences in the total ARG abundance and  $\alpha$ -diversity measures were observed over time in the placebo group (Figure 4a).

Similar to the microbiome, the  $\beta$ -diversity PCA plot also showed that variation between different individuals (inter-individual variation) is the most dominant factor affecting the resistome composition (PERMANOVA:  $R^2 = 0.82$ ,  $P = .0001$ ;  $nperm = 999$ ; Supplementary Figure 3B). Moreover, the inter-individual variability of resistome was even higher than in microbiome composition, with most ARGs observed in a limited number of participants. At the same time, no visible pattern of resistome profile changes over time was observed in both placebo and treated groups (Figure 4b). However, PERMANOVA analysis with Bray-Curtis distance demonstrated a significant shift in the overall resistome composition after amoxicillin exposure at 3 months (PERMANOVA: Bray:  $R^2 = 0.072$ ;  $P = .046$ ; strata = Patient)

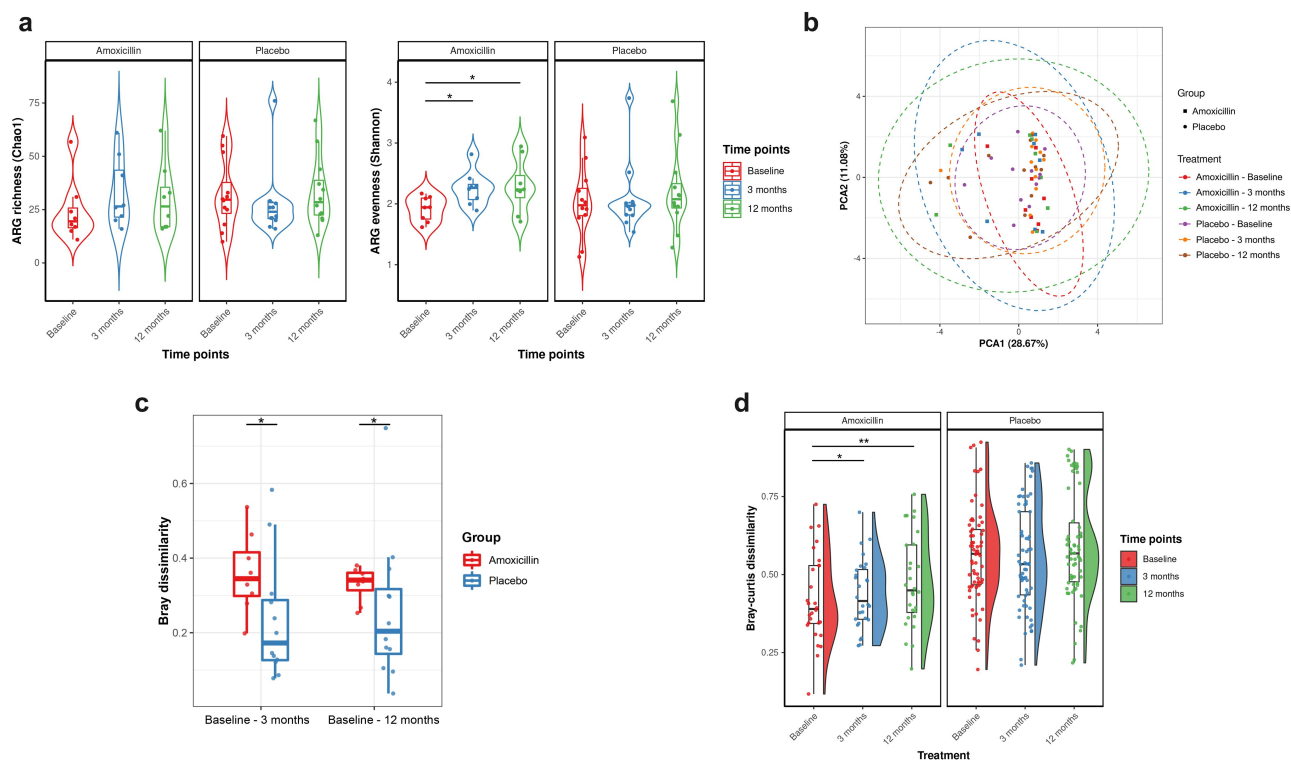


**Figure 3.** Beta-lactam and total AMR gene abundance increases after amoxicillin exposure. (a) Box plots showing the relative abundance measured as reads per kilobase million (RPKM) of antibiotic resistance gene (ARG) classes across all samples, stratified by treatment group and time points. The center horizontal line of box is median, box limits are upper and lower quartiles, whiskers are  $1.5\times$  interquartile ranges. (b) The total abundance, measured as the total reads per kilobase million (RPKM): Violin plots showing the total AMR genes abundance level per sample (represented as point), stratified by treatment and time points. The horizontal box lines represent the first quartile, the median and the third quartile. Whiskers denote the range of points within the first quartile  $-1.5\times$  the interquartile range and the third quartile  $+1.5\times$  the interquartile range. The black horizontal line on the top of plot connects statistically significantly different time points within each treatment group (Adjusted  $P$  values ( $P$ ):  $***P < .001$ ,  $**P < .01$ ;  $*P < .05$ ; one-way ANOVA with repeated measures followed by post hoc pairwise  $t$ -test is used to check the statistical significance).

and 12 months (PERMANOVA: Bray:  $R^2 = 0.074$ ;  $P = .01$ ; strata = Patient) when taking significant interindividual variation into account. Further, the compositional differences (Bray) observed upon amoxicillin exposure were higher and long-lasting, which remained statistically significant compared to the placebo group up to 12 months (Figure 4c; Wilcoxon rank-sum test), unveiling persistent diversification of resistome composition after antibiotic treatment. We also found that amoxicillin exposure was associated with prolonged resistome diversification through an individualized selection of ARGs as the dissimilarity of the resistome composition among the individuals significantly increased over time (Figure 4d; Wilcoxon signed rank exact test). When considering ARGs presence/absence rather than the abundance, similar resistome divergence was observed in the amoxicillin treatment group. In contrast, we observed that the resistome composition significantly converged at 3 months based on Jaccard distance in the placebo group (Supplementary Figure 5; Wilcoxon signed rank exact test). No significant divergence or convergence based on Bray-Curtis distance was detected in the resistome composition over time among placebo-treated individuals. Thus, these long-term

changes observed in the resistome profiles may have resulted from a drug-specific selection rather than from short-term shifts discerned in the microbiome profiles.

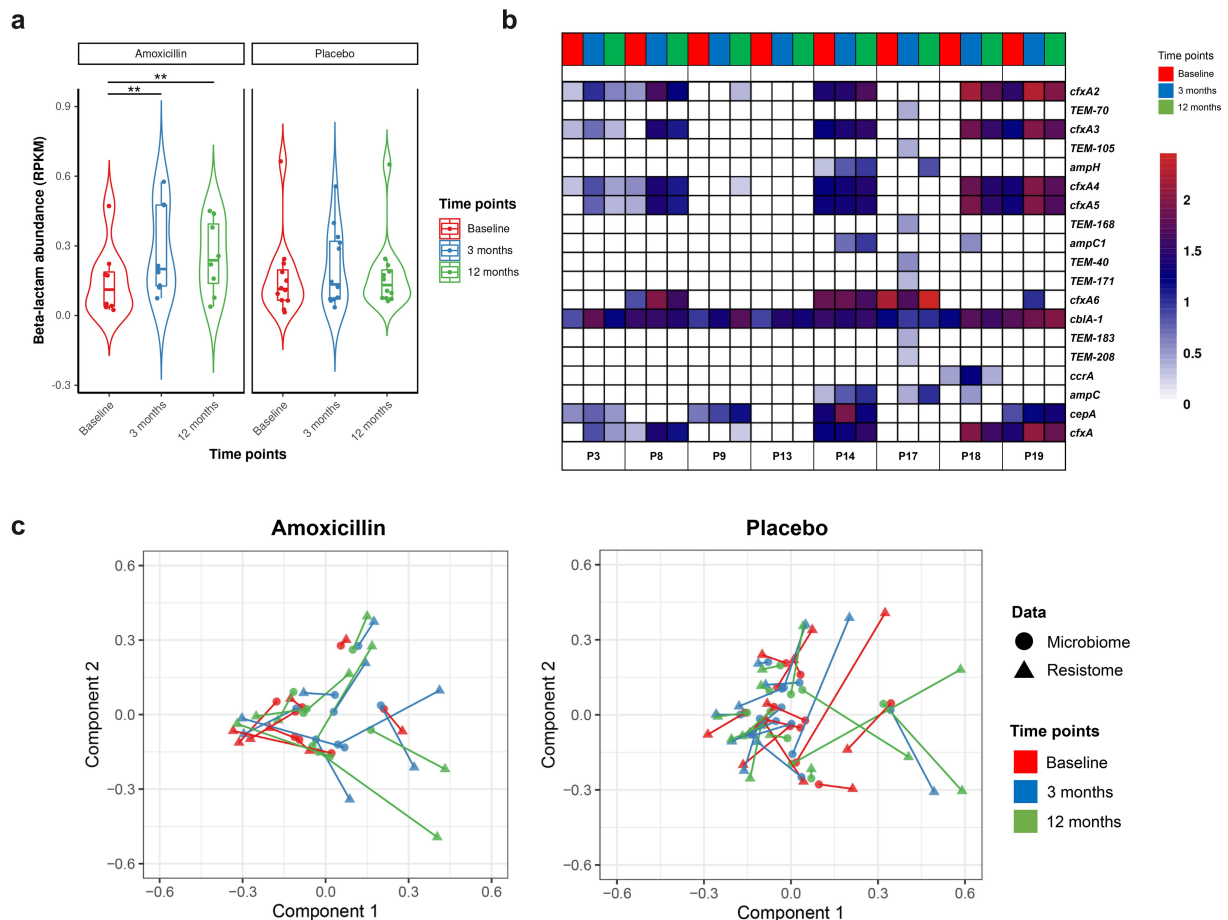
Additional investigation of AMR determinants into the gut resistome revealed a common targeted response toward the amoxicillin treatment at the ARG-Class level. The class which directly corresponded to the treatment, i.e., beta-lactam, significantly increased in abundance in the amoxicillin group. Such effect persisted even after 9 months completion of the treatment (LEfSe) (Figure 5a). Also, the mean relative abundance of multidrug class increased by more than fourfold ( $\log_2 FC > 2$ ) compared to baseline after amoxicillin treatment, while several other ARG classes with a threefold increase in relative abundances were observed over time in both the amoxicillin and placebo groups. Thus, such changes could not be linked to antibiotic treatment (Supplementary Figure 6). At the gene level, there was a high inter-individual variation of increase in abundances or emergence of beta-lactamase ARGs noticed after amoxicillin intervention (Figure 5b). The most consistent effect of the antibiotic was a threefold increase in abundance of *cfxA* beta-lactamase and its gene variants (*cfxA2*, *cfxA3*, *cfxA4* and *cfxA6*) at 3 months in five out of



**Figure 4.** Impact of amoxicillin on fecal ARG diversity and composition. (a) Violin plots showing the ARG  $\alpha$ -diversity measured and compared over time points in amoxicillin and placebo groups using Chao1 (richness; left) and Shannon (evenness; right) index, respectively. Each point corresponds to a given sample, and each box span from the first to third quartiles with a horizontal line inside the boxes representing the median. Adjusted  $P$  values were computed using the LME mixed effect model and Tukey's HSD post hoc test. Adjusted  $P$  values ( $P$ ): \*\*\* $P < .001$ , \*\* $P < .01$ , \* $P < .05$ . (b)  $\beta$ -diversity was analyzed using principal component analysis (PCA) based on the clr-transformed ARG count abundance matrix. Each point represents a single sample, shape indicate treatment group, colored according to different time points and groups. Ellipses indicate 95% confidence intervals (CI). (c) ARG  $\beta$ -diversity measured by Bray–Curtis dissimilarity is compared between the amoxicillin (red) and placebo (blue) samples at baseline and the other time points, i.e., 3 and 12 months. Each point represents the dissimilarity in one individual's resistome at baseline compared each of the other time point samples. Center line is median, box limits are upper and lower quartiles, whiskers are 1.5 $\times$  interquartile ranges and points beyond whiskers are outliers ( $P < .05$ ; Wilcoxon rank sum test). (d) Distribution of the Bray–Curtis distance of resistome (ARGs) between patients at the same time point within treatment groups. The statistical difference between the timepoints in both groups was tested by the paired Wilcoxon test, and the significance is marked with \*  $P < .05$ , \*\* $P < .01$  and \*\*\* $P < .001$ .

eight amoxicillin-exposed participants. The abundance of these genes remained higher at 9 months post-treatment compared to their baseline levels. These *cfxA* beta-lactamase family resistance genes are commonly found in multiple bacterial genera, including *Bacteroides*, and are associated with penicillin and cephalosporin resistance. Notably, the relative abundance of *Bacteroides* species was significantly increased in the amoxicillin patients, indicating that *Bacteroides* may have harbored the *cfxA* genes. Similarly, another clinically relevant beta-lactamase ARG, i.e., *cepA* encoding resistance to cephalosporins enriched immediately following antibiotic intake in patients, when present above

detection levels at baseline (four out of eight patients). Additionally, we observed a post-antibiotic increase in abundances or induction of *ampC*-type beta-lactamase ARGs (*ampC*, *ampC1* and *ampH*) in three patients. Interestingly, their relative abundance pattern is congruent with *Escherichia coli*, a known microbial host harboring these ARGs. Further, we detected several of the TEM beta-lactamase ARGs (*TEM-70*, *TEM-168*, *TEM-105*, *TEM-171*, *TEM-183*, *TEM-205*) in one participant (P17) that were initially undetectable and selected toward higher abundance immediately after antibiotic exposure (3 months) and returned to lower levels following antibiotic withdrawal.



**Figure 5.** Impact of amoxicillin on abundance of beta-lactam ARGs and association between microbiome and resistome compositions. (a) Violin plots showing the relative abundance of class Beta-lactam measured as log<sub>10</sub> of reads per kilobase million (RPKM) of Beta-lactam ARGs, stratified by treatment group and then time points. The changes in abundance were statistically evaluated using one-way ANOVA with repeated measures followed by post hoc pairwise t-test. The horizontal box lines represent the first quartile, the median and the third quartile. Whiskers denote the range of points within the first quartile – 1.5× the interquartile and the third quartile + 1.5× the interquartile. (Adjusted *P* values (*P*): \*\*\**P* < .001, \*\**P* < .01; \**P* < .05 one-way ANOVA with repeated measures followed by post-hoc pairwise t-test). (b) Heat map displaying the log-transformed RPKM relative abundance of beta-lactam ARGs identified in amoxicillin-exposed patients. The x-axis displays all the samples at different time points which are stratified by the participants. (c) Procrustes analysis of resistome composition (filled triangles) and species composition (filled circles) for amoxicillin (left) and placebo (right)-treated groups using Hellinger transformation and PCoA ordination. The points are colored based on sampling time points in both groups. The length of line connecting two points indicates the degree of dissimilarity or distance between microbiome and resistome composition of the same sample.

These TEM-genes are potentially clinically relevant ARGs because they confer resistance toward commonly used antibiotics, including penicillins and cephalosporins, and because they are known to be precursors for beta-lactamases with an extended spectrum of resistance (ESBLs). Also, a fivefold increase in abundance of another clinically significant metallo beta-lactamase ARG (*ccrA*) encoding resistance to last-resort antibiotic (i.e., carbapenem) was observed for patient P18 at 3 months (Figure 5b).

Finally, the Procrustes analysis revealed that the resistome composition (ARG level) was significantly correlated with the microbiome (species level) in both placebo (Protest: sum of squares ( $m^2$ ) = 0.21;  $R^2$  = 0.88;  $P$  = .001; permutations = 999) and amoxicillin (Protest: sum of squares ( $m^2$ ) = 0.27;  $R^2$  = 0.85;  $P$  = .001; permutations = 999) groups, indicating that microbial community composition shapes the ARG distribution in the human gut. However, the overall correlation was more robust in placebo. More importantly, the baseline



has a tighter correlation than post-treatment samples in the amoxicillin group, supporting the hypothesis that additional factors, such as antibiotic exposure, may have influenced the resistome changes independently of the microbiome (Figure 5c & Supplementary Table S5).

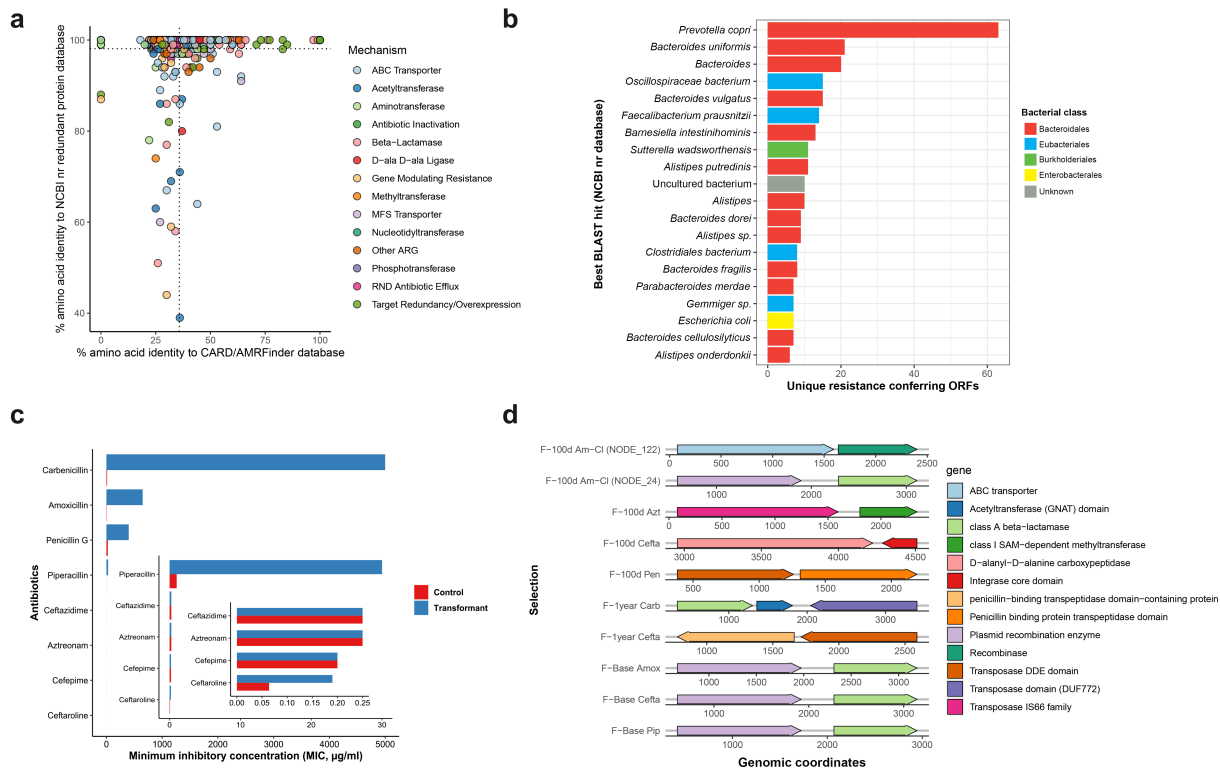
### **Functional characterization of the antibiotic resistance reservoir in human gut microbiota**

To functionally investigate the resistome, including the diversity of ARGs that remain undetected by conventional AMR databases in the gut microbiome, we next complemented shotgun metagenomics with functional metagenomic analysis. The constructed functional libraries encompassed a total of ~14.7 gigabase (Gb) pairs of metagenomic DNA with library sizes of 6Gb (baseline), 4.4Gb (3 months), and 4.3Gb (12 months), respectively, and an average insert size of 1.5–kb. Resistance was identified against 8 out of the 14 antibiotics screened. Importantly, we did not recover any resistant transformants for all the last-resort antibiotics tested, including meropenem, colistin and tigecycline. These antibiotics are considered as the last line of defense against infection caused by multidrug-resistant organisms (MDROs). In addition, none of the libraries constructed from feces of antibiotic-treated adult individuals retrieved ciprofloxacin, ceftioxin or gentamicin resistance (Supplementary Figure 7). Of the 42 selections performed, 24 yielded antibiotic-resistant *E. coli* transformants, of which 19 resulted in successfully sequenced and assembled libraries. On average, we assembled 3082 contigs with lengths greater than 500 bp with 4083 ORFs predicted in these contigs (Supplementary Table S2). In total, 1089 complete sequences were assigned to antibiotic resistance functions after hierarchical annotation of ORFs using multiple AMR databases, representing 599 unique sequences. These functionally discovered ARGs included multiple beta-lactamase classes (mainly class A and no class C), transcriptional regulators, multidrug efflux pumps, transporter proteins, acetyltransferases, aminotransferases, resistance modulators, aminoglycoside-modifying enzyme, tetracycline resistance protein, among others. Most of the functional ARGs (89.31%) were annotated through the Resfams-full

database that contains additional HMM profiles of specific AMR genes that can contribute to antibiotic resistance, such as acetyltransferases, AraC transcriptional regulators, and ATP-binding cassette (ABC) efflux pumps.

The mean identity of these functionally selected ARGs to the CARD and AMRFinder-Protein database was 35.8% (median: 33%), whereas their mean identity to NCBI non-redundant protein entries was 98.1% (median 100%) (Figure 6a). This indicates that while most of the functionally identified ARGs in our selections are similar to previously sequenced proteins, these ARGs are remarkably infrequent in the current AMR databases. Interestingly, we also found that several of the functionally identified genes have already been annotated as ARGs in the NCBI database but are not functionally characterized with similar approach (HMM-based), across various habitats and environments in large-scale studies.<sup>51,53,57</sup> The predicted microbial hosts of these functional ARGs (determined by the best BLAST hit to the NCBI non-redundant protein database) were predominantly commensal bacteria from the order Bacteroidales (Figure 6b). The detection of these commensal Bacteroidales as the potential hosts of ARGs in the human gut microbiome is not surprising and is in concordance with the current understanding of Bacteroidales as traffickers and prolific hosts of ARGs.<sup>12,58–60</sup> Additionally, we identified uncultured bacterium as one of the top potential hosts of these ARGs (11 ORFs), suggesting that even not well-represented uncultured organisms can be important sources of ARGs in the gut microbiome.

We next investigated the influence of amoxicillin intake on the prevalence of functionally discovered ARGs over time. To avoid false positives and over-inflation of ARGs abundance levels in the sequenced metagenomes, we only mapped ARGs identified using the Resfams-core database to the shotgun data.<sup>42,61</sup> We noticed a similar pattern as for the shotgun metagenomics of high inter-individual variation in the abundance of functional ARGs in the gut microbiome of the treated subjects (Supplementary Figure 8A). These ARGs were found both in the treated and in the placebo subjects. The abundance of functional beta-lactamases slightly increased after treatment in the



**Figure 6.** Functional metagenomic screening using *E. coli* as a surrogate host reveals an enriched gut resistome in amoxicillin exposed patients. (a) Amino acid identity between all of the functionally identified ARGs and their top hit in CARD/AMRFinder versus their top hit in the NCBI non-redundant protein database, colored by their mechanism of action. The vertical and horizontal dotted lines indicate the mean percentage identity of functional ARGs in the CARD/AMRFinder and NCBI database, respectively. (b) Horizontal bar plot showing the top 20 most commonly predicted hosts of functionally selected ARGs on the basis of maximum identity BLAST hits in the NCBI non-redundant protein database. Different colored bars represent different bacterial species orders. (c) Minimum inhibitory concentrations (MICs) of putatively novel ARG for *Escherichia coli* pZE21 transformants producing class A beta-lactamase and *Escherichia coli* pZE21 without the insert. This novel ARG has 100% amino acid identity to serine hydrolase (NCBI GenBank) and 38% of amino acid identity to known class A beta-lactamase (CARB-11) gene identified using BLASTP against the Comprehensive Antibiotic Resistance Database (CARD). Bars represent the average value from three independent experiments each. (d) Synteny of functionally selected ARGs with mobile genetic elements. Putative mobile genetic elements were identified syntenic to ARGs recovered from amoxicillin-clavulanate (a,b), aztreonam (c), ceftazidime (d, g, i), penicillin G (e), carbenicillin (f), amoxicillin (h) and piperacillin (j) selections. The x-axis shows the metagenomic coordinates.

amoxicillin-treated subjects, though, unlike shotgun metagenomic data, these changes were not statistically significant (Supplementary Figure 8B). In comparison, no such trend was noticed in the placebo group.

We also discovered at least 13 putatively novel ARGs in the gut microbiome of antibiotic-treated patients. No sequence homologs were detected for eight of these ARGs in known AMR sequence databases. The list of putatively novel ARGs annotated based on AMR (CARD, AMRFinder) and NCBI non-redundant protein database is presented in Table 2. We further experimentally validated a putative class A beta-lactamase resistance protein with 100% amino

acid identity to serine hydrolase and 38% identity to known beta-lactamase resistance gene (CARB-11) by expressing it in *E. coli*. The results showed that this novel ARG conferred resistance toward carbenicillin, as observed by the increase in the minimal inhibitory concentration (MIC) by more than 400-fold compared to the control. For amoxicillin, we found approximately a 108-fold (650 µg/ml) increase in the MIC values. It also demonstrated reduced susceptibility, ranging from 16- to 30-fold, to other penicillins such as penicillin G and piperacillin. These are typical characteristics of class A carbenicillin-hydrolyzing beta-lactamases (CARB-type) (Figure 6c and Additional file S4).

**Table 2. Putative novel functional ARGs.**

ORF Info	CARD+AMRFINDER Database					NCBI Nr Protein Database					Global alignment of best top local hit			
	Selection	Length	Source	Protein type	Protein ID	Identity (%)	Coverage (%)	Description	Protein ID	Identity (%)	Coverage (%)	Description	Identity (%)	Note
NODE_23_length_2861_cov_41484676_2_#_339_#_1_#_ID = 23_2; partial = 0;start_type = ATGrbs_motif = None;rbs_spacer = None; gC_cont = 0.569	F-100d Pip, F-100d Anox	303	Resfams-core	Class A beta-lactamase	WP_063857838.1	38	82	CARB-11	HJAO9995.1	99.01	100	serine hydrolase	99	Functional validated
NODE_34_length_2830_cov_744602162_2_#_338_#_1_#_ID = 34_2; partial = 0;start_type = ATGrbs_motif = None;rbs_spacer = None; gC_cont = 0.569	F-1 year Cefta	646	Resfams-full	Penicillin binding protein transpeptidase domain	WP_000138348.1	31	84	pbp2	OLA80974.1	99.36	96	hypothetical protein BHW58_06090 [Azospirillum sp. 51_20] (624)	96	
NODE_17_length_2193_cov_247850327_2_#_125_#_2065_#_1_#_ID = 17_2; partial = 0;start_type = ATGrbs_motif = None;rbs_spacer = None; gC_cont = 0.567	F-1 year Cefta	303	Resfams-full	Beta-lactamase superfamily domain	AWN09461.1	22	65	metallo-beta-lactamase PNGM-1	WP_092366234.1	100	100	ribonuclease Z	100	
NODE_524_length_1351_cov_123915895_2_#_247_#_1_#_ID = 524_2; partial = 0;start_type = ATGrbs_motif = GGAG/GAGG; rbs_spacer = 5-10bp;gC_cont = 0.622	F-1 year Am-CI	306	Resfams-full	Beta-lactamase superfamily domain	AWN09461.1	24	45	metallo-beta-lactamase PNGM-1	WP_206813282.1	100	99	ribonuclease Z	99.3	
NODE_3401_length_519_cov_1668103_2_#_72_#_509_#_1_#_ID = 3401_2; partial = 0;start_type = ATGrbs_motif = None;rbs_spacer = None; gC_cont = 0.562	F-100d Am-CI	145	Resfams-full	Acetyltransferase (GNAT) domain	WP_046928126.1	31	80	aacA	OLA82746.1	99	98.61	hypothetical protein BHW58_00500 [Azospirillum sp. 51_20]	77.2	
NODE_295_length_1136_cov_5179463_1_#_86_#_979_#_1_#_ID = 295_1; partial = 0;start_type = ATGrbs_motif = None;rbs_spacer = None; gC_cont = 0.511	F-Base Pen	297	Resfams-full	Aminotransferase class I and II	No hit	-	-	-	WP_006846964.1	100	100	pyridoxal phosphate-dependent aminotransferase	73.7	
NODE_336_length_1177_cov_8866809_1_#_22_#_528_#_1_#_ID = 336_1; partial = 0;start_type = ATGrbs_motif = TATAArbs_spacer = 3bp; gC_cont = 0.491	F-1 year Pen	168	Resfams-full	Aminotransferase class I and II	No hit	-	-	-	GKH14396.1	100	89	hypothetical protein CE91512_26060	84	
NODE_97_length_1542_cov_21211163_2_#_292_#_1443_#_1_#_ID = 97_2; partial = 0;start_type = ATGrbs_motif = AAAAArbs_spacer = 8bp; gC_cont = 0.443	F-Base Pip	383	Resfams-full	Aminotransferase class I and II	No hit	-	-	-	ALK85074.1	100	100	8-amino-7-oxononanoate synthase	97.7	
NODE_35_length_2999_cov_27911685_1_#_292_#_1488_#_1_#_ID = 35_1; partial = 0;start_type = ATGrbs_motif = None;rbs_spacer = None; gC_cont = 0.596	F-100d Am-CI	398	Resfams-full	Aminotransferase class I and II	No hit	-	-	-	WP_014775721.1	100	100	pyridoxal phosphate-dependent aminotransferase	100	
NODE_88_length_1596_cov_2811811_2_#_162_#_1244_#_1_#_ID = 88_2; partial = 0;start_type = ATGrbs_motif = None;rbs_spacer = None; gC_cont = 0.494	F-100d Cefta	360	Resfams-full	Aminotransferase class I and II	No hit	-	-	-	MBN2914949.1	99	100	aminotransferase PatB family C-5 lyase	90.2	
NODE_90_length_1589_cov_8839635_2_#_275_#_1294_#_1_#_ID = 90_2; partial = 0;start_type = ATGrbs_motif = TAAArbs_spacer = 11bp; gC_cont = 0.496	F-100d Cefta	339	Resfams-full	Aminotransferase class IV	No hit	-	-	-	WP_005833410.1	100	100	branched-chain amino acid aminotransferase	100	

(Continued)





To underline the potential of horizontal ARG transfer within the gut microbiome, we explored the presence of mobile genetic elements in the assembled contigs (in synteny with ARGs). 4.43% of the total contigs encoded a putative mobile genetic element. These were observed in all selections against beta-lactam antibiotics, commonly appearing in penicillin, amoxicillin and piperacillin selections, but also in lower numbers in, the ceftazidime selections (Supplementary Figure 9B). Nevertheless, antibiotic treatment differences did appear. We noticed a significantly higher load of putative MGEs in post-antibiotic treatment selections than in baseline selections in our functional metagenomic data (Supplementary Figure 9A) (pairwise Wilcoxon test). Interestingly, these post-treatment selections had the most putative MGE-associated annotations despite lower sequencing reads and lower input libraries. MGEs colocalized with AMR genes are of serious concern due to the ease of their spread. In particular, we recovered a functionally identified class A beta-lactamase ARG colocalized with a plasmid element selected in penicillins (amoxicillin, piperacillin, amoxicillin-clavulanate) and cephalosporin (ceftazidime). Such ARGs with typical characteristics of extended-spectrum beta-lactamases (ESBLs) can play a crucial role in resistance dissemination (Figure 6d). All these findings emphasize the usefulness of functional metagenomics as a culture-independent and sequence-unbiased approach for resistome characterization.

## Discussion

In this study, we explored the ecological side effects of prolonged amoxicillin exposure on the human gut microbiome and resistome. We found that long-term amoxicillin therapy in a relatively healthy population had more pronounced yet transient effects on microbiome composition. This was in contrast with the effects on the resistome, for which changes persisted for long-term after treatment cessation. Enrichment of antibiotic resistance genes was mostly specific to beta-lactam antibiotics, including instances showing enrichment of beta-lactamase genes associated with extended spectrum of activity (ESBL) and with action against carbapenems.

Perturbations in gut microbiome balance by short-course amoxicillin interventions have reported minor or not significant effects on microbiome composition.<sup>30–32,62</sup> Our results indicate that in contrast to short-course, prolonged use of amoxicillin can have severe side-effects. We found that amoxicillin had a common and reproducible impact on the microbiome with most noticeable changes observed in the species with lower abundance. Among these were known short-chain fatty acid (particularly butyrate) producers, significantly depleted upon antibiotic therapy completion. Contrarily, bacterial genera *Bacteroides*, well-known for carrying resistance to beta-lactams, systematically increased upon amoxicillin treatment.<sup>60,63</sup> In addition, we validated previous findings where the potential marker of beta-lactam or antibiotic-induced dysbiosis, i.e., *L. bolteae*, significantly enriched upon amoxicillin treatment.<sup>18,64</sup> Upon treatment, we also observed depletion of a recently discovered health-associated species, i.e., *Oscillibacter* sp. 57\_20.<sup>56</sup> The effects on microbiome composition were short-lived as the microbiome returned to pre-treatment levels at 9 months after amoxicillin treatment, which can be explained by community resilience.<sup>33,65</sup> Despite their transience, such microbial disturbances can also have negative health-related consequences such as increased susceptibility to recurrent *Clostridioides difficile* infections, which can be more detrimental for patients with an already dysbiotic microbiome (diseased).<sup>21,66</sup>

Further, our work also highlights that amoxicillin exposure led to a short-term but significant increase in microbiome dissimilarity between baseline and post-treatment samples. The possible explanation behind such diversification would be considerable variations in baseline microbiome composition across patients, which may result in different levels of severity and vulnerability to antibiotic perturbation. The influence of inter-individual variability on microbiome responses was particularly evident on the resistome results, with a higher inter-individual variability in response to amoxicillin for the resistome than the microbiome. Such difference may reflect the fact that most of the ARGs are specific to individuals, while microbial taxa are more conserved across patients. In line with our results, individualized responses to antibiotic interventions have also been reported for resistome studies evaluating

short-course antibiotic therapies.<sup>18,67,68</sup> Notably, no significant changes were observed in the microbiome and resistome diversity and composition over time in patients who did not take amoxicillin, thus confirming the stability, resilience and robustness of the healthy, adult human gut microbiome.<sup>69,70</sup>

Complementation of whole metagenomic sequencing with functional metagenomics enables broader characterization of resistance determinants in metagenomes.<sup>42,45</sup> Functional metagenomics provides not only functional information on the ARGs present in microbial communities but also enables the discovery of novel ARGs. The technique utilizes a heterologous host, e.g. *E. coli*, for gene expression via library cloning which poses as a strong advantage given that it allows for the identification of ARGs from uncultured bacteria. We found that one of the top potential hosts of functionally identified ARGs included an uncultured bacterium, along with the well-known and predominant (identified through shotgun metagenomics) *Bacteroides* and *Prevotella* species. Notably, although we did not identify resistance against last resort antibiotics in the functional screening, some were detected by whole metagenomic sequencing, thus highlighting the complementarity of the two methods. We have also discovered several putatively novel ARGs (1 validated functionally) in the gut microbiome of the amoxicillin-treated patients, with most of them detected in selections from post-treatment libraries. Lastly, detection of MGEs mediating the spread of ARGs (colocalized) in our functional metagenomic data highlights the likelihood of a mobilizable resistome within the human gut microbiome. Interestingly, most putative MGE-associated annotations were in the post-treatment samples, suggesting an additional possible collateral effect of prolonged amoxicillin therapy.

The small treatment group size limits the generalization of our conclusions and makes it difficult to discern significant relationships at the population level, mainly due to high inter-individual variability and personalized response to antibiotics perceived in our data. Nonetheless, we rigorously controlled for such variability by using individuals as their controls in pre- and post-treatment comparisons. Moreover, additional sampling time

points before and during antibiotic administration, would have provided a better understanding of the dynamics and natural variability of the human gut microbiome overtime. Functional metagenomics was employed to complement the resistome information from whole metagenomic analysis. While it provides precious information based on phenotypic characterization, this approach is still limited in scope due to definite classes of antibiotics and the lack of efficient cloning systems for expression in Gram-positives.<sup>71</sup> Additionally, the isolation of mobile genetic elements colocalized with ARGs is technically constrained, as little information regarding the genetic background of the functional ARGs is available because of small DNA fragments (insert) size in our functional selections. Still, functional metagenomics is the only high-throughput metagenomics method that enables identification of completely novel ARGs. Future multicenter, longitudinal sequencing-based studies with increased sampling density, complemented with functional metagenomics assays using multi-host-systems would have the potential to provide solid evidence of relevance for antibiotic stewardship practices. However, large cohort prospective studies to investigate the impact of prolonged use of antibiotics in humans would need special ethical considerations regarding benefits and risks, which only in rare cases would be justifiable. Such considerations highlight the critical relevance of the high-throughput microbiome and resistome analysis of unique data from our randomized trial of relatively healthy individuals, despite limitations in sample size.

In conclusion, our results provide compelling evidence that the human adult gut microbiome and resistome respond differently to prolonged amoxicillin treatment. The persistent increase in AMR abundance and diversity along with enrichment of beta-lactam resistance genes long after antibiotic use termination, including resistance genes to last resort antibiotics detected by whole metagenomic sequencing, highlights the risks associated with prolonged antibiotic exposure. Such risks need to be considered in view of potential benefits. Patients with chronic low back pain and Modic changes included in the current study were part of a multicenter study in Norway that showed a lack of significant effects of prolonged amoxicillin

use in self-reported measures of disability, pain intensity or quality of life.<sup>35</sup> Thus, our study adds an additional layer of support against the general recommendation of antibiotic therapy for chronic low back pain and Modic changes and emphasizes the importance of considering such risks in relation to other conditions with no or minor benefits from prolonged antibiotic therapy.

## Acknowledgments

The sequencing service was provided by the Norwegian Sequencing Centre ([www.sequencing.uio.no](http://www.sequencing.uio.no)), a national technology platform hosted by the University of Oslo and supported by the “Functional Genomics” and “Infrastructure” programs of the Research Council of Norway and the Southeastern Regional Health Authorities. The sequencing data analysis and storage were performed on SAGA & NIRD resources provided by Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway. Also, we would like to thank Marius Trøseid, Martin Kummen and Johannes ER Hov for invaluable advice with planning collection of fecal samples and Knut Morten Huneide, Marianne Thorsø, Anne Julsrud Haugen and Lars Grøvlø for help with sample collection.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by the Norwegian Research Council (NFR) (project number: 273833, 274867), Olav Thon Foundation (project number: 421258), by governmental organizations Helse Sør-Øst RHF (grant number: 2015090) and KLINBEFORSK (grant number: 2017201).

## ORCID

Roger Junges  <http://orcid.org/0000-0002-5538-1088>  
Fernanda Cristina Petersen  <http://orcid.org/0000-0003-4649-247X>

## Data availability

Clean shotgun metagenomic reads after removal of human DNA and assembled functional metagenomic contigs are available at NCBI SRA under BioProject ID: PRJNA894204. Additional participants metadata (deidentified) will be available upon request addressed to [kjersti.storheim@medisin.uio.no](mailto:kjersti.storheim@medisin.uio.no), in accordance with local registration and ethical approval.

## Contributions

FCP conceived the study design, experiments and analyses. LCHB collected the fecal samples and clinical characteristics from participants. K Stu., RJ, and GS were involved in metagenomic DNA extraction from feces and prepared shotgun metagenomic sequencing libraries. HA created functional metagenomic libraries, performed functional selections and prepared functional metagenomic sequencing libraries. AB and GS also contributed to the functional metagenomics. K Sto. and JAZ planned and performed the original AIM trial. DB contributed with the concept and idea of collecting fecal samples. AD performed the computational analyses of shotgun metagenomic sequencing and functional metagenomic data. AD interpreted the results and drafted the manuscript and figures with critical revision performed by FCP. RJ and AD edited the final figures in the manuscript. All authors made a substantial contribution to the revision of the manuscript and approved its final version.

## References

1. United IACG. No time to wait—securing the future from drug-resistant infections. 2019.
2. World Health Organization. 2014. Antimicrobial resistance: global report on surveillance. World Health Organization. [https://scholar.google.com/scholar?hl=en&as\\_sdt=0%2C5&q=Antimicrobial+resistance%3A+global+report+on+surveillance&btnG=#d=gs\\_cit&t=1671178189963&u=%2Fscholar%3Fq%3Dinfo%3Aani4nP-tGNAsJ%3Ascholar.google.com%2F%26out%3Dcite%26scirp%3D0%26hl%3Den](https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=Antimicrobial+resistance%3A+global+report+on+surveillance&btnG=#d=gs_cit&t=1671178189963&u=%2Fscholar%3Fq%3Dinfo%3Aani4nP-tGNAsJ%3Ascholar.google.com%2F%26out%3Dcite%26scirp%3D0%26hl%3Den)
3. Centers for Disease Control and Prevention. Office of infectious disease antibiotic resistance threats in the United States, 2013. Atlanta: Centers for Disease Control Prevention; 2013.
4. Ventola CL. The antibiotic resistance crisis: part 1: causes and threats. *Pharmacy and Therapeutics*. 2015;40:277.
5. Viswanathan V. Off-label abuse of antibiotics by bacteria. *Gut Microbes*. 2014;5(1):3–4. doi:10.4161/gmic.28027.
6. Kährström CT. Entering a post-antibiotic era? *Nature Reviews. Microbiology*. 2013;11(3):146. doi:10.1038/nrmicro2978.
7. Lown M, McKeown S, Stuart B, Francis N, Santer M, Lewith G, Su F, Moore M, Little P. Prescribing of long-term antibiotics to adolescents in primary care: a retrospective cohort study. *British Journal of General Practice*. 2021;71(713):e887–e94. doi:10.3399/BJGP.2021.0332.
8. Guarner F, Malagelada J-R. Gut flora in health and disease. *The Lancet*. 2003;361(9356):512–519. doi:10.1016/S0140-6736(03)12489-0.
9. Ramirez J, Guarner F, Bustos Fernandez L, Maruy A, Sdepanian VL, Cohen H. Antibiotics as major

- disruptors of gut microbiota. *Frontiers in cellular infection microbiology*. *Frontiers in Cellular and Infection Microbiology*. 2020;10:572912. doi:10.3389/fcimb.2020.572912.
10. Coyne MJ, Zitomersky NL, McGuire AM, Earl AM, Comstock LE, Mekalanos J. Evidence of extensive DNA transfer between bacteroidales species within the human gut. *MBio*. 2014;5(3):e01305–14. doi:10.1128/mBio.01305-14.
  11. Salyers AA. Antibiotic resistance transfer in the mammalian intestinal tract: implications for human health, food safety and biotechnology. Berlin, Heidelberg: Springer; 1995. Jointly published with R.G. Landes, Biomedical Publishers, Austin, USA.
  12. Sommer MO, Dantas G, Church GM. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science*. 2009;325(5944):1128–1131. doi:10.1126/science.1176950.
  13. Van Schaik W. The human gut resistome. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2015;370(1670):20140087. doi:10.1098/rstb.2014.0087.
  14. Dethlefsen L, Relman DA. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences*. 2011;108:4554–4561. doi:10.1073/pnas.1000087107.
  15. Jernberg C, Löfmark S, Edlund C, Jansson JK. Long-term impacts of antibiotic exposure on the human intestinal microbiota. *Microbiology*. 2010;156(11):3216–3223. doi:10.1099/mic.0.040618-0.
  16. Korry BJ, Cabral DJ, Belenky P. Metatranscriptomics reveals antibiotic-induced resistance gene expression in the murine gut microbiota. *Front Microbiol*. 2020;11:322. doi:10.3389/fmicb.2020.00322.
  17. Pérez-Cobas AE, Artacho A, Knecht H, Ferrús ML, Friedrichs A, Ott SJ, Moya A, Latorre A, Gosalbes MJ. Differential effects of antibiotic therapy on the structure and function of human gut microbiota. *PloS one*. 2013;8(11):e80201. doi:10.1371/journal.pone.0080201.
  18. Raymond F, Ouameur AA, Déraspe M, Iqbal N, Gingras H, Dridi B, Leprohon P, Plante PL, Giroux R, Bérubé È. The initial state of the human gut microbiome determines its reshaping by antibiotics. *ISME J*. 2016;10(3):707–720. doi:10.1038/ismej.2015.148.
  19. Modi SR, Collins JJ, Relman DA. Antibiotics and the gut microbiota. *J Clin Invest*. 2014;124(10):4212–4218. doi:10.1172/JCI72333.
  20. Panda S, El Khader I, Casellas F, Lopez Vivancos J, Garcia Cors M, Santiago A, Cuenca S, Guarner F, Manichanh C. Short-term effect of antibiotics on human gut microbiota. *PloS one*. 2014;9(4):e95476. doi:10.1371/journal.pone.0095476.
  21. Anthony WE, Wang B, Sukhum KV, D'Souza AW, Hink T, Cass C, Seiler S, Reske KA, Coon C, Dubberke ER, et al. Acute and persistent effects of commonly used antibiotics on the gut microbiome and resistome in healthy adults. *Cell Rep*. 2022;39(2):110649. doi:10.1016/j.celrep.2022.110649.
  22. Palleja A, Mikkelsen KH, Forslund SK, Kashani A, Allin KH, Nielsen T, Hansen TH, Liang S, Feng Q, Zhang C, et al. Recovery of gut microbiota of healthy adults following antibiotic exposure. *Nature Microbiology*. 2018;3(11):1255–1265. doi:10.1038/s41564-018-0257-9.
  23. Doan T, Worden L, Hinterwirth A, Arzika AM, Maliki R, Abdou A, Zhong L, Chen C, Cook C, Lebas E, et al. Macrolide and nonmacrolide resistance with mass azithromycin distribution. *New England Journal of Medicine*. 2020;383(20):1941–1950. doi:10.1056/NEJMoa2002606.
  24. Murray AK, Zhang L, Snape J, Gaze WH. Comparing the selective and co-selective effects of different antimicrobials in bacterial communities. *Int J Antimicrob Agents*. 2019;53(6):767–773. doi:10.1016/j.ijantimicag.2019.03.001.
  25. Lopatkin AJ, Sysoeva TA, You L. Dissecting the effects of antibiotics on horizontal gene transfer: analysis suggests a critical role of selection dynamics. *Bioessays*. 2016;38(12):1283–1292. doi:10.1002/bies.201600133.
  26. Jakobsson HE, Jernberg C, Andersson AF, Sjölund-Karlsson M, Jansson JK, Engstrand L, Ratner AJ. Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. *PloS one*. 2010;5(3):e9836. doi:10.1371/journal.pone.0009836.
  27. Koo H, Hakim JA, Crossman DK, Kumar R, Lefkowitz EJ, Morrow CD. Individualized recovery of gut microbial strains post antibiotics. *NPJ Biofilms and Microbiomes*. 2019;5(1):1–6. doi:10.1038/s41522-019-0103-8.
  28. Lavelle A, Hoffmann TW, Pham H-P, Langella P, Guédon E, Sokol H. Baseline microbiota composition modulates antibiotic-mediated effects on the gut microbiota and host. *Microbiome*. 2019;7(1):1–13. doi:10.1186/s40168-019-0725-3.
  29. Centers for Disease Control and Prevention. Centers for Disease Control and Prevention. Atlanta (GA, USA): Outpatient antibiotic prescriptions—United States, 2014. CDC; 2018.
  30. Elvers KT, Wilson VJ, Hammond A, Duncan L, Huntley AL, Hay AD, Van Der Werf, ET. Antibiotic-induced changes in the human gut microbiota for the most commonly prescribed antibiotics in primary care in the UK: a systematic review. *BMJ open*. 2020;10(9):e035677. doi:10.1136/bmjopen-2019-035677.
  31. Pallav K, Dowd SE, Villafuerte J, Yang X, Kabbani T, Hansen J, Dennis M, Leffler DA, Newburg DS, Kelly CP, et al. Effects of polysaccharopeptide from *Trametes Versicolor* and amoxicillin on the gut microbiome of healthy volunteers. *Gut Microbes*. 2014;5(4):458–467. doi:10.4161/gmic.29558.
  32. Zaura E, Brandt BW, Teixeira de Mattos MJ, Buijs MJ, Caspers MP, Rashid MU, Weintraub A, Nord CE, Savell A, Hu Y. Same exposure but two radically different



- responses to antibiotics: resilience of the salivary microbiome versus long-term microbial shifts in feces. *MBio*. 2015;6(6):e01693–15. doi:10.1128/mBio.01693-15.
33. Dethlefsen L, Huse S, Sogin ML, Relman DAJ, Eisen JA. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol*. 2008;6(11):e280. doi:10.1371/journal.pbio.0060280.
  34. Albert HB, Kjaer P, Jensen TS, Sorensen JS, Bendix T, Manniche C. Modic changes, possible causes and relation to low back pain. *Med Hypotheses*. 2008;70(2):361–368. doi:10.1016/j.mehy.2007.05.014.
  35. Bråten LCH, Rolfsen MP, Espeland A, Wigemyr M, Aßmus J, Froholdt A, Haugen AJ, Marchand GH, Kristoffersen PM, Lutro O, et al. Efficacy of antibiotic treatment in patients with chronic low back pain and Modic changes (the AIM study): double blind, randomised, placebo controlled, multicentre trial. *BMJ*. 2019;367. doi:10.1136/bmj.l5654
  36. Storheim K, Espeland A, Grøvle L, Skouen JS, Aßmus J, Anke A, Froholdt A, Pedersen LM, Haugen AJ, Fors T, et al. Antibiotic treatment In patients with chronic low back pain and Modic changes (the AIM study): study protocol for a randomised controlled trial. *Trials*. 2017;18(1):1–11. doi:10.1186/s13063-017-2306-8.
  37. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30:2114–2120.
  38. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357–359. doi:10.1038/nmeth.1923.
  39. Andrews S. FastQC: a quality control tool for high throughput sequence data. Cambridge (United Kingdom): Babraham Bioinformatics, Babraham Institute; 2010.
  40. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, Mailyan A, Manghi P, Scholz M, Thomas AM, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife*. 2021;10:e65088. doi:10.7554/eLife.65088.
  41. Alcock BP, Raphenya AR, Lau TT, Tsang KK, Bouchard M, Edalatmand A, Huynh W, Nguyen ALV, Cheng AA, Liu S, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res*. 2020;48(D1):D517–D25. doi:10.1093/nar/gkz935.
  42. D'Souza AW, Boolchandani M, Patel S, Galazzo G, van Hattem JM, Arcilla MS, Melles DC, de Jong MD, Schultsz C, Bootsma MCJ, et al. Destination shapes antibiotic resistance gene acquisitions, abundance increases, and diversity changes in Dutch travelers. *Genome Med*. 2021;13(1):1–21. doi:10.1186/s13073-021-00893-z.
  43. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–2079. doi:10.1093/bioinformatics/btp352.
  44. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841–842. doi:10.1093/bioinformatics/btq033.
  45. Boolchandani M, Patel S, Dantas G. Functional metagenomics to study antibiotic resistance. *Antibiotics*: Springer; 2017. 307–329.
  46. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. *Genome Res*. 2017;27(5):824–834. doi:10.1101/gr.213959.116.
  47. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072–1075. doi:10.1093/bioinformatics/btt086.
  48. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform*. 2010;11(1):1–11. doi:10.1186/1471-2105-11-119.
  49. Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, Philippon A, Allesoe RL, Rebelo AR, Florensa AF, et al. ResFinder 4.0 for predictions of phenotypes from genotypes. *Journal of Antimicrobial Chemotherapy*. 2020;75(12):3491–3500. doi:10.1093/jac/dkaa345.
  50. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, Tyson GH, Zhao S, Hsu C-H, McDermott PF, et al. Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrobial Agents Chemotherapy*. 2019;63(11):e00483–19. doi:10.1128/AAC.00483-19.
  51. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J*. 2015;9(1):207–216. doi:10.1038/ismej.2014.106.
  52. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 2012;28(23):3150–3152. doi:10.1093/bioinformatics/bts565.
  53. Gasparrini AJ, Wang B, Sun X, Kennedy EA, Hernandez-Leyva A, Ndao IM, Tarr PI, Warner BB, Dantas G. Persistent metagenomic signatures of early-life hospitalization and antibiotic treatment in the infant gut microbiota and resistome. *Nature Microbiology*. 2019;4(12):2285–2297. doi:10.1038/s41564-019-0550-2.
  54. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. Metagenomic biomarker discovery and explanation. *Genome Biol*. 2011;12(6):1–18. doi:10.1186/gb-2011-12-6-r60.
  55. Pulipati P, Sarkar P, Jakkampudi A, Kaila V, Sarkar S, Unnisa M, Reddy DN, Khan M, Talukdar R. The Indian

- gut microbiota—Is it unique? *Indian Journal of Gastroenterology*. 2020;39(2):133–140. doi:10.1007/s12664-020-01037-8.
56. Asnicar F, Berry SE, Valdes AM, Nguyen LH, Piccinno G, Drew DA, Leeming E, Gibson R, Le Roy C, Khatib HA. Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat Med*. 2021;27(2):321–332. doi:10.1038/s41591-020-01183-8.
57. Xie F, Jin W, Si H, Yuan Y, Tao Y, Liu J, Wang X, Yang C, Li Q, Yan X. An integrated gene catalog and over 10,000 metagenome-assembled genomes from the gastrointestinal microbiome of ruminants. *Microbiome*. 2021;9(1):1–20. doi:10.1186/s40168-021-01078-x.
58. Crits-Christoph A, Hallowell HA, Koutouvalis K, Suez J. Good microbes, bad genes? The dissemination of antimicrobial resistance in the human microbiome. *Gut Microbes*. 2022;14(1):2055944. doi:10.1080/19490976.2022.2055944.
59. Shoemaker N, Vlamakis H, Hayes K, Salyers A. Evidence for extensive resistance gene transfer among bacteroides spp. and among bacteroides and other genera in the human colon. *Applied Environmental Microbiology*. 2001;67(2):561–568. doi:10.1128/AEM.67.2.561-568.2001.
60. Yan W, Hall AB, Jiang X. Bacteroidales species in the human gut are a reservoir of antibiotic resistance genes regulated by invertible promoters. *npj Biofilms and Microbiomes*. NPJ Biofilms and Microbiomes. 2022;8(1):1–9. doi:10.1038/s41522-021-00260-1.
61. Forsberg KJ, Patel S, Gibson MK, Lauber CL, Knight R, Fierer N, Dantas G. Bacterial phylogeny structures soil resistomes across habitats. *Nature*. 2014;509(7502):612–616. doi:10.1038/nature13377.
62. Ladirat SE, Schoterman MH, Rahaoui H, Mars M, Schuren FHJ, Gruppen H, Nauta A, Schols HA . Exploring the effects of galacto-oligosaccharides on the gut microbiota of healthy adults receiving amoxicillin treatment. *British Journal of Nutrition*. 2014;112(4):536–546. doi:10.1017/S0007114514001135.
63. Wexler HM. Bacteroides: the good, the bad, and the nitty-gritty. *Clin Microbiol Rev*. 2007;20(4):593–621. doi:10.1128/CMR.00008-07.
64. MacPherson CW, Mathieu O, Tremblay J, Champagne J, Nantel A, Girard SA, Tompkins TA . Gut bacterial microbiota and its resistome rapidly recover to basal state levels after short-term amoxicillin-clavulanic acid treatment in healthy adults. *Sci Rep*. 2018;8(1):1–14. doi:10.1038/s41598-018-29229-5.
65. Ley RE, Peterson DA, Gordon JI. Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell*. 2006;124(4):837–848. doi:10.1016/j.cell.2006.02.017.
66. Mosca A, Leclerc M, Hugot JP. Gut microbiota diversity and human diseases: should we reintroduce key predators in our ecosystem? *Frontiers in Microbiology*. *Frontiers in Microbiology*. 2016;7:455. doi:10.3389/fmicb.2016.00455.
67. Forslund K, Sunagawa S, Kultima JR, Mende DR, Arumugam M, Typas A, Bork P. Country-specific antibiotic use practices impact the human gut resistome. *Genome Res*. 2013;23(7):1163–1169. doi:10.1101/gr.155465.113.
68. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, et al. A core gut microbiome in obese and lean twins. *Nature*. 2009;457(7228):480–484. doi:10.1038/nature07540.
69. Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, Clemente JC, Knight R, Heath AC, Leibel RL, et al. The long-term stability of the human gut microbiota. *Science*. 2013;341(6141):1237439. doi:10.1126/science.1237439.
70. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. *Nature*. 2012;489(7415):220–230. doi:10.1038/nature11550.
71. Mullany P. Functional metagenomics for the investigation of antibiotic resistance. *Virulence*. 2014;5(3):443–447. doi:10.4161/viru.28196.







1 **Unravelling the landscape of antibiotic resistance determinants in the**  
2 **nasopharynx and the impact of antibiotics: a longitudinal study of preterm**  
3 **infants**

4 Achal Dhariwal†, Polona Rajar†, Gabriela Salvadori, Heidi Aarø Åmdal, Dag Berild, Ola Didrik  
5 Saugstad, Drude Fugelseth, Gorm Greisen, Ulf Dahle, Kirsti Haaland, Fernanda Cristina Petersen\*

6

7 †Shared First authorship

8 \*Correspondence: Fernanda Cristina Petersen ([f.c.petersen@odont.uio.no](mailto:f.c.petersen@odont.uio.no))

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31 **ABSTRACT**

32 Respiratory infections caused by common pathogens that colonize the respiratory tract, including  
33 the nasopharynx, are a leading global cause of deaths associated with antimicrobial resistance. Yet,  
34 little is known about the antibiotic resistance determinants in the microbial communities of the  
35 respiratory tract. In this study, we examined the development of the resistome (collection of  
36 antibiotic resistance genes) in the nasopharynx of preterm infants and the impact of early  
37 ampicillin plus gentamicin exposure in those requiring broad-spectrum antibiotic treatment for  
38 suspected early-onset neonatal sepsis. We analyzed the resistome of 181 nasopharyngeal samples  
39 collected longitudinally from 36 preterm infants using deep shotgun metagenomic sequencing.  
40 Antibiotic resistance genes were found in most samples (~95%) of both naive and antibiotic-  
41 treated infants. We observed a significant increase in the diversity and total abundance of antibiotic  
42 resistance genes primarily immediately after early antibiotic therapy. The nasopharyngeal  
43 resistome in preterm infants was mainly driven by the microbiome ( $r = 0.85$ ,  $p$ -value = 0.001),  
44 with inter-individual differences explaining most of the variation observed in the resistome  
45 composition ( $R^2 = 30.03\%$ , adjusted  $p$ -value = 0.001). We also noted a discernible impact of  
46 prenatal maternal antibiotic exposure on the overall resistome composition ( $R^2 = 3.6\%$ , adjusted  
47  $p$ -value = 0.001), which was observed as a subgroup within the antibiotic-treated group. Overall,  
48 these data suggest that the nasopharynx in preterm infants is characterized by a rich resistome with  
49 a trajectory that is negatively affected by antibiotics. Although the effects appear to be transient,  
50 any adverse changes in resistome during the first days of life bear risks, particularly those selecting  
51 for drug resistant colonizers with the potential to spread and cause invasive infections.

52

53 **Keywords:** respiratory resistome, metagenomics, nasopharynx, microbiome, preterm infants,  
54 antibiotics, antimicrobial resistance

55

56

57

58

59

60

## 61 INTRODUCTION

62 Antimicrobial resistance (AMR) is one of the biggest global health threats, estimated to be  
63 associated with approximately 5 million deaths in 2019. Among these, respiratory infections are  
64 the leading AMR infection burden, accounting for more than 1.5 million deaths [1]. These are  
65 often caused by pathogens such as *Streptococcus pneumoniae*, *Staphylococcus aureus*, *Klebsiella*  
66 *pneumoniae*, and *Haemophilus influenzae*, which are part of the normal human microbiome found  
67 in the respiratory tract, including the nasopharynx. Recent studies have shown that for these  
68 pathogens, more than 90% of antibiotic exposure represents bystander events, in which they are  
69 not the infective agent [2]. This highlights the importance of the potential off-target impact of  
70 antibiotics in the selection and development of antibiotic resistance by human pathogens. These  
71 pathogens share common niches with commensals, which provide colonization resistance by  
72 preventing pathogen overgrowth. However, commensals can also serve as reservoirs of antibiotic  
73 resistance genes (ARGs) that can readily be transferred between different strains and species of  
74 bacteria in microbiomes [3-5]. Together, commensals and pathogens interact to provide additional  
75 benefits, such as immune system education and development [6, 7]. Such interactions highlight  
76 the importance of an ecological perspective when considering AMR and the impact of antibiotics  
77 on microbiomes.

78 Recent studies have contributed to an increased understanding of the landscape of ARGs, known  
79 as resistomes, and the impact of antibiotics on the gut microbiome and resistome [8-12]. Of critical  
80 relevance are studies focusing on vulnerable groups, in particular preterm infants, who are highly  
81 exposed to antibiotic interventions early in life [9, 13-15]. Sepsis is a serious complication for  
82 preterm infants and one of the main reasons why many receive empiric antibiotics in the first days  
83 or weeks after birth [16, 17]. However, the risks associated with suspected early onset sepsis  
84 (sEONS) vary, and over-prescription of antibiotics in low sepsis risk preterm infants may result in  
85 unnecessary microbiome and resistome disturbances with potential short- and long-term health  
86 implications [18]. Early and prolonged antibiotic use in this group has been associated with an  
87 increased risk of various adverse outcomes, including bronchopulmonary dysplasia, late-onset  
88 sepsis, invasive fungal infections, and mortality, with conflicting results regarding necrotizing  
89 enterocolitis [19-26]. Furthermore, antibiotic use in early life may have unwanted consequences  
90 beyond the neonatal period, such as an increased risk of asthma and obesity [27-29]. In relation to  
91 antibiotic resistance, studies focusing on gut resistomes indicate that infants exhibit higher

92 resistome burden than their adult mothers, and that antibiotic exposure amplifies this burden [9-  
93 11, 14, 15, 30, 31].

94

95 Although most of the associations between microbial disturbances and the risk of common  
96 morbidities are well-acknowledged for the gut microbiome in infants, growing evidence suggests  
97 that respiratory health is also significantly influenced by microbial communities in the upper  
98 respiratory tract. This includes the nasopharynx, which serves as a critical reservoir of respiratory  
99 pathogens and a gatekeeper of respiratory health [5-7, 32, 33]. Recent studies have highlighted the  
100 potential negative impact of antibiotic treatment on the nasopharynx microbiome and its  
101 association with an increased risk of developing asthma and respiratory infections in infants [10,  
102 33, 34]. There is currently a significant gap in our understanding of how the infant respiratory  
103 resistome develops, including the trajectory, dynamics, and factors that can influence its  
104 establishment in preterm infants.

105

106 We recently developed an optimization strategy for DNA extraction and library preparation that  
107 helped to circumvent the challenges posed by low biomass and high host DNA content in  
108 nasopharynx samples of preterm infants [35]. In this study, we applied the optimized protocol to  
109 nasopharyngeal aspirate samples to characterize the development of the resistome in 36 preterm  
110 infants from birth up to 6 months corrected age. By analyzing ~1.85 terabytes (Tb) of metagenomic  
111 DNA data generated from deep shotgun sequencing, we obtained the most comprehensive  
112 information to date on the characteristics and dynamics of resistome development in the respiratory  
113 tract of preterm infants. Our hypothesis was that early exposure to broad-spectrum antibiotics for  
114 treatment of sEONS in preterm infants, may promote the enrichment of ARGs and compositional  
115 alterations in the respiratory resistome, which would persist until 6 months corrected age, as  
116 compared to antibiotic-naive preterm infants. Additionally, we explored the extensive metadata  
117 collected to gain insights into additional clinical covariates that may be associated with resistome  
118 development.

## 119 **MATERIALS AND METHODS**

### 120 **Ethics statement**

121 Premature infant metadata and all nasopharynx aspirate samples used in this study were collected  
122 at the Neonatal Intensive Care Unit (NICU) at Oslo University Hospital, Ullevål, Norway, as part

123 of the Born in the Twilight of Antibiotic project. The study received approval from the Regional  
124 Committee for Medical and Health Research Ethics – South East, Norway (2018/1381 REKD),  
125 and followed the principles of the Declaration of Helsinki guidelines. Infants were enrolled in the  
126 study after obtaining written informed consent from their parents.

127

### 128 **Study design and sample collection**

129 The study is a prospective observational cohort study. Between July 2019 and January 2021, we  
130 included preterm infants born (or transferred within 48 hours) at the NICU (Ullevål) who were  
131 between 28+0 and 31+6 weeks of gestational age (GA). In total, 66 preterm infants were enrolled  
132 in the study, and followed from birth until six months corrected age (i.e., adjusted for prematurity)  
133 during their stay in the NICU and after discharge. Five of the infants were lost to follow-up.  
134 Extensive metadata and covariates from infants and mothers, including mode of delivery,  
135 gestational age (GA) at birth, postmenstrual age (PMA), gender, birth weight (BW), postnatal  
136 infant antibiotics usage (duration, type, days), maternal antibiotic usage, feeding regimen during  
137 hospitalization and after discharge, and other demographic characteristics, were collected through  
138 medical records. This data is stored in secured server at the Services for Sensitive Data (TSD),  
139 University of Oslo.

140 In our cohort of 66 infants, 33 infants with sEONS received broad-spectrum antibiotics, consisting  
141 of a combination of ampicillin (a beta-lactam antibiotic) and gentamicin (an aminoglycoside  
142 antibiotic), within 24 hours after birth. Due to the circumstances of premature birth and the  
143 instability of the infants, baseline samples could not be collected before antibiotic therapy  
144 initiation. Instead, the first nasopharyngeal aspirate samples (baseline) were obtained within 48  
145 hours after the initiation of early antibiotic therapy. To investigate temporal changes in the  
146 resistome in response to antibiotics and control for potential baseline differences as confounding  
147 factors, this study included only infants who had nasopharyngeal samples collected on the same  
148 day as the initiation of early antibiotic treatment ( $n = 15$ ) and antibiotic-naive infants as controls  
149 ( $n = 21$ ). The baseline characteristics of the infants stratified by antibiotic exposure are presented  
150 in Table 1.

151 Nasopharyngeal aspirate samples were collected from infants at six distinct time points: days of  
152 life (DOL) 0, 7, 14, 28, 56, and at 6 months corrected age. Additional samples were taken if the

153 antibiotic therapy was initiated or terminated more than 48 hours before/after a predefined  
154 sampling time. In cases where the sample could not be collected at the scheduled time points (such  
155 as with an unstable infant), a new sample was obtained as soon as possible. Infants who were  
156 transferred to another hospital before discharged to home adhered to the study protocol. When the  
157 infants reached six months of corrected age, they were either invited to an outpatient visit at Ullevål  
158 or their local hospital or visited at home for sample collection. Nasopharynx aspirates were  
159 collected as previously described [35] and stored at -80°C until further processing.

### 160 **Metagenomic DNA extraction and sequencing**

161 Metagenomic DNA was extracted by first thawing the aspirate samples (2 ml) on ice, pelleting  
162 them by centrifugation at 10,000 g for 10 minutes at 4°C, and then using MolYsis™ for host DNA  
163 depletion (Molzym, Bremen, Germany), following the manufacturer's protocol with some  
164 modifications, as previously described [35]. After depletion of human DNA, fresh pellets were  
165 spiked with 20 µl of ZymoBIOMICS Spike-in Control II (catalog number: D6321 & D6321-10),  
166 and bacterial DNA was extracted using the MasterPure™ Gram Positive DNA Purification Kit  
167 (Epicentre, Madison, WI, USA), according to the manufacturer's recommendations. The bacterial  
168 DNA was eluted in 35 µl of TE buffer and stored at -80°C until further use. Final DNA  
169 concentration was quantified using Qubit dsDNA HS assay kits in a Qubit 4.0 Fluorometer  
170 (Invitrogen, Thermo Fisher Scientific, USA). Metagenomic sequencing libraries were constructed  
171 using the Nextera DNA Flex library preparation kit (Illumina Inc., CA, USA). All samples were  
172 subjected to paired-end sequencing at 2 × 150 bp on an Illumina NovaSeq S4 high-output platform  
173 at the Norwegian Sequencing Centre (Oslo, Norway).

174

### 175 **Controls for low microbial biomass**

176 The study followed recent recommendations for low microbial biomass microbiome research [36].  
177 This involved the following steps: (1) collecting samples from the included infants in the two  
178 groups within the same time period, (2) using standard protection measures during sample  
179 collection (such as clean suits, disposable gloves, and face masks), (3) including sampling and  
180 DNA extraction blank controls (negative controls), (4) conducting DNA extraction and library  
181 preparation while wearing laboratory coats and disposable gloves in a carefully cleaned  
182 environment, and (5) using unique barcodes in library preparation. The negative controls had no

183 or extremely low DNA concentrations to proceed to library preparation as determined by RT-PCR  
184 using the Femto Bacterial DNA Quantification Kit from Zymo and were therefore excluded from  
185 further analysis ( $n = 10$ ; mean = 0.00043 ng per  $\mu\text{L}$ ). Additionally, spike-in controls ( $n = 6$ ) were  
186 extracted and used as positive controls, and longitudinal samples from each preterm infant were  
187 extracted within the same run.

188

### 189 **Bioinformatics processing**

190 The raw metagenomic sequencing data was pre-processed using our in-house bioinformatics  
191 pipelines using a high-performance computing cluster inside a secured environment, i.e., TSD at  
192 the University of Oslo. In brief, the Nextera adaptor sequences and low-quality reads were filtered  
193 using Trim Galore (v.0.6.1) with default parameters [37]. Next, the human DNA contaminant  
194 sequences were identified and removed by mapping the quality-trimmed reads against the human  
195 reference genome (GRCh38) using Bowtie2 (v.2.3.4.2) [38] (non-default parameters:  $q -N 1 -k 1$   
196  $--fr --end-to-end --phred33 --very-sensitive -no-discordant$ ) along with SAMtools (v.1.9) [39] and  
197 BEDTools (v.2.27.1) [40]. The sequence quality reports on raw and processed reads were  
198 generated using FastQC (v.0.11.9) [41] and summarized using MultiQC (v.1.7) [42].

199

### 200 **Resistome profiling**

201 To characterize the presence of ARGs, cleaned high-quality reads for each sample were mapped  
202 against the *nucleotide\_fasta\_protein\_homolog\_model* from the Comprehensive Antibiotic  
203 Resistance Database (CARD) (v.3.2.2) [43] using Bowtie2 under parameter  $-very-sensitive-local$ .  
204 The mapped reads from each sample were then filtered, sorted, and indexed using SAMtools. The  
205 number of reads mapped to each ARG was calculated using SAMtools *idxstats* and BEDTools.  
206 ARGs with a coverage of at least 80% from the alignment were retained for further downstream  
207 analyses. The mapped read counts were normalized for bacterial sequence abundances and gene  
208 lengths by calculating reads per kilobase of reference gene per million bacterial reads (RPKM) for  
209 each CARD reference sequence. The relative abundance of ARGs for each sample was computed  
210 by dividing the RPKM by the total sum of the RPKM for each sample. For keeping the acyclic  
211 hierarchical annotation structure for accurate resistome profiling at higher functional levels, ARGs  
212 were manually re-annotated based on the drug class to which they confer resistance. ARGs  
213 belonging to macrolides, lincosamides and streptogramins were congregated into the MLS class.

214 While ARGs belonging to carbapenem, cephamycin, cephalosporin, penem, penam and  
215 monobactam were congregated into the Beta-lactam class. Rarefaction analysis was performed  
216 using Rarefaction Analyzer [44] to determine the saturation of samples at various sequencing  
217 depths for recovery of ARGs.

218

### 219 **Statistical analysis and data visualization**

220 All statistical analyses were conducted in R (v.4.2.1) within RStudio (v.2022.07.2+576) [45, 46].  
221 The ARG abundance, annotation table, and metadata file were compiled into a single data object  
222 using the phyloseq (v.1.40.0) package [47]. Figures, unless stated otherwise, were created using  
223 the ggplot2 (v.3.4.0) package [48] and further edited using Adobe Illustrator (v.16.0.0).  $\alpha$ - and  $\beta$ -  
224 diversity analysis was performed using the vegan (v.2.6.2) [49] and phyloseq R packages.  $\alpha$ -  
225 diversity was calculated using the Shannon diversity metric, and differences over time points in  
226 groups were assessed using a linear mixed-effects (LME) model with infants set as a random effect  
227 while correcting for age. The overall resistome composition of nasopharyngeal samples was  
228 visualized using principal component analysis (PCA) ordination plots. The ordinates were based  
229 on the Euclidean dissimilarity matrix of centered log-ratio (CLR) transformed relative abundance  
230 data (i.e., Aitchison distance). Associations of clinical covariates with the dispersion of sample  
231 resistome compositions were statistically evaluated with the Permutational multivariate analysis  
232 of variance (PERMANOVA) test using the *adonis2* function (vegan package) with 999  
233 permutations. To test the homogeneity of multivariate dispersion, the permutational analysis of  
234 multivariate dispersion (PERMDISP) test was used. In case of heterogeneity, the analysis of  
235 similarities (ANOSIM) test using the *anosim* function (vegan package) was further used to  
236 statistically re-assess the association of variables with the overall resistome composition. In the  
237 multivariable, temporal analyses, only the variables that individually showed a significant  
238 association with resistome composition were included.

239

240 To identify the resistotypes, the Dirichlet multinomial mixture (DMM) model approach was  
241 utilized using the DirichletMultinomial (v.1.38.0) R package [50]. Procrustes analysis was  
242 performed to examine the effect of underlying microbiota on the resistome. The ARG RPKM  
243 abundance matrix and the species-level count per million (CPM) abundance matrix were  
244 Hellinger-transformed and ordinated using PCoA (Principal coordinates analysis) on Bray–Curtis



245 dissimilarities. The symmetric Procrustes correlation coefficients between the ordinations and P-  
246 values were obtained through the *protest* function from the *vegan* package. Spearman's correlation  
247 was applied to resistome and microbiome relative abundance profiles for paired samples. To avoid  
248 the potential bias introduced by Spearman's rank when ranking zero values, we removed ARGs  
249 and species that were present in less than half of the samples. The correlation coefficient and  
250 corresponding adjusted p-values were calculated using the *associate* function from the *microbiome*  
251 (v.1.18.0) package [51]. Associations between individual taxa and overall resistome outcomes  
252 were calculated using repeated measure correlation analysis from the *rmcorr* (v.0.5.4) package  
253 while controlling for inter-individual variation [52]. Heatmaps were created using the *pheatmap*  
254 function from the *pheatmap* R package (v.1.0.12) [53]. For simple, independent comparisons of  
255 group differences, we used a one-way analysis of variance test, Wilcoxon rank-sum test, Kruskal-  
256 Wallis test, or chi-square test as appropriate, and we considered  $p < 0.05$  to be significant. For all  
257 analyses involving multiple comparisons, we applied the Benjamini-Hochberg (BH) method to  
258 correct for multiple testing. The statistical analysis scheme used in the study is shown in  
259 Supplementary Fig. 1.

260

## 261 **RESULTS**

### 262 **Study cohort and sample characteristics**

263 Infants exposed to early antibiotics ( $n = 15$ ) had significantly lower GA and BW, and their mothers  
264 had a more prolonged period of ruptured membranes before delivery compared to antibiotic-naive  
265 infants ( $n = 21$ ). None of the mothers of naive infants received antibiotics during pregnancy  
266 (prenatal antibiotics) before onset of delivery (Table 1). The average early antibiotic treatment  
267 duration was 5 days (SD: 2 days). The sampling time point T2 corresponds to the first available  
268 sample collected (except in one infant) after the cessation of early antibiotic treatment. In addition  
269 to early antibiotics, one of the included infants received subsequent antibiotics initiated  $> 72$  hours  
270 after birth and before discharge from NICU and two other infants received antibiotic therapy  
271 between discharge from NICU and 6 months corrected age. A timeline of sample collection and  
272 antibiotic administration is represented in Fig. 1.

273

274 In total, we collected and extracted metagenomic DNA from 198 nasopharyngeal samples obtained  
275 from 36 preterm infants, from day of birth until six months corrected age. The DOL at which

276 samples were obtained for each time point varied: T1 (DOL: 0-3), T2 (DOL: 4-9), T3 (DOL: 12-  
277 19), T4 (DOL: 20-31), T5 (DOL: 32-76) and 6 months corrected age (DOL: 229-288). Four (2%)  
278 samples were excluded after DNA extraction due to negative 16S rRNA values (qRT-PCR) after  
279 subtracting spike-in background. The remaining 194 samples proceeded to library preparation and  
280 whole metagenomic sequencing (WMS). Detailed information on sample exclusion and inclusion  
281 statistics is described in Supplementary Table 1.

282

### 283 **Overall Sequencing results**

284 A total of 6.14 billion raw sequencing reads were obtained with a mean value of sample reads  
285 counts of 33.97 million (M) ranging from 4.19 to 69.25 M reads. The mean Phred quality score of  
286 raw reads for samples ranged from 32 to 37. More than 99% of the total reads generated across all  
287 samples passed the sequence quality filtering and trimming. On average 46.71% (range: 1.09 -  
288 97.69%) of total quality-filtered reads were identified as belonging to the human genome. After  
289 removal of the human-associated metagenomic reads, a total of 2.98 billion remaining high-quality  
290 clean reads across all samples with number of reads ranging from 0.76 to 59.9 M per sample, and  
291 with median of 16.4 M reads were subjected to resistome profiling. In general, 0.31% (range: 0 -  
292 2.8%; 9.32 million in total) of the clean reads were classified as ARGs. Results from rarefaction  
293 analysis suggested that most of the samples had sufficient sequencing depth for ARG  
294 characterization (Supplementary Fig. 2). More detailed information regarding the overall  
295 sequencing results is presented in Supplementary Table 2. Ten samples were excluded from  
296 downstream analysis as no reads were assigned to ARGs. In two infants there was more than one  
297 sample available for the defined time points (collected due to initiation or discontinuation of  
298 antibiotics more than 48 hours before/after scheduled time point). Therefore, to conduct time point-  
299 wise analysis, only one sample was considered for each of the time points (Supplementary Table  
300 1 & Fig. 1). Collectively, we assessed the resistome of the nasopharyngeal microbiome in a total  
301 of 181 samples collected from 36 preterm infants.

302

### 303 **Distribution and characterization of ARGs in the preterm infant nasopharyngeal** 304 **microbiome**

305 In total, we found 373 ARGs belonging to 15 ARG classes conferring resistance via 5 distinct  
306 mechanisms. Multidrug efflux mediated ARGs comprised the majority of the resistome with an

307 average of 27% at T1 to T5 and 17.8% at T6 (six months corrected age) of the total relative  
308 abundance. There was a median of 18 (mean: 24) unique ARGs ranging from 1 to 157 per analyzed  
309 sample. On average, maximum number of ARGs were detected after the discontinuation of  
310 treatment (T2) (mean: 40) in most infants receiving early antibiotics treatment. In the non-treated  
311 group, the corresponding mean value at T2 was 19. Only 13 ARGs were both most prevalent  
312 (detected in > 40% of all samples) and highly abundant with mean relative abundance of 64.8%  
313 (SD: 32.06%) across all preterm infant nasopharyngeal resistome. These core set of resistance  
314 genes included multidrug resistance genes (*acrB*, *oqxB*, *mexI*), macrolide-lincosamide-  
315 streptogramin (MLS) resistance genes (*mel*, *pmrA*, *rlmA(II)*), two beta-lactam resistance genes  
316 (*blaZ*, *SST-1*), fluoroquinolone resistance genes (*patA*, *patB*), tetracycline resistance genes  
317 (*tet(41)*, *tetM*), and the aminoglycoside resistance gene *AAC(6')-Ic*.

318

319 In both antibiotic-treated and naive group, the most abundant class of ARG was multidrug (mean:  
320 27.85%; SD: 21.90%), followed by fluoroquinolone (mean: 17.41%; SD: 20.18%), beta-lactam  
321 (mean: 15.27%; SD 16.99%), tetracycline (mean: 14.70%; SD: 10.75%), MLS (mean: 13.06%;  
322 SD: 15.29%) and aminoglycoside (mean: 8.77%; SD: 12.48%) (Fig. 2A). In terms of prevalence,  
323 aminoglycoside (76/82) and beta-lactam (75/82) were the most common class present in the  
324 samples from antibiotic-treated group, while multidrug (92/99) and beta-lactam (91/99) were the  
325 most common class in the naive infants. When classified based on resistance mechanism, antibiotic  
326 efflux (mean: 59.56%; SD: 20.97%) was the most abundant category in both the groups across all  
327 time points (Supplementary Fig. 3). At baseline (T1), the ARG composition of antibiotic-treated  
328 group was found to be more heterogenous (higher inter-individual variability) and diverse  
329 compared to naive infants (Fig. 2B).

330

331 Homology to ARGs posing major current threat for public health, including those identified by  
332 recent omics-based framework [54], were identified in nasopharyngeal samples of our cohort.  
333 Most of these potential high-risk ARGs (rank I) were detected at multiple time points across infants  
334 in both antibiotic-treated and naive group (Supplementary Fig. 4). We also observed a few high  
335 risk and clinically relevant ARGs that appeared directly after the early antibiotic treatment at T2.  
336 For instance, high-risk ARGs specific to the antibiotics encoding for aminoglycoside-modifying  
337 enzymes (*AAC(3)-II*) along with extended-spectrum TEM beta-lactamase (*bla<sub>TEM-1</sub>*) were detected

338 in two infants. These ARGs co-occurred with other non-targeted high-risk ARGs such as *qnrS*,  
339 *floR* and *aadA* (rank II). Previous studies have shown the co-carriage of these ARGs on multidrug  
340 resistance plasmids in *Enterobacter* species [55]. Also, other clinically relevant Extended Spectrum  
341 Beta-Lactamase (ESBL) genes belonging to SHV beta-lactamase (rank II) family were detected in  
342 two other infants following antibiotic exposure. In one patient, the plasmid-mediated AmpC-type  
343  $\beta$ -lactamase (ACT-beta lactamase), known to confer resistance against all classes of beta-lactam  
344 antibiotics such as penicillins, cephalosporins and carbapenems was detected only after the  
345 treatment cessation (T2). These high-risk ARGs were observed mainly in early antibiotics exposed  
346 infants whose mothers had also received antibiotics during pregnancy (prenatal). Nonetheless, they  
347 did not persist in nasopharyngeal samples collected after discharge at the corrected age of 6  
348 months. Amongst ARGs that are found in Gram-positive pathogens, the *Staphylococcus aureus*  
349 methicillin-resistance gene i.e., *mecA* was detected in 44 samples from almost all (14/15)  
350 antibiotic-treated infants and 10 of the 21 naive infants.

351

### 352 **Impact of early antibiotics on preterm infant nasopharyngeal resistome**

353 Firstly, we investigated the total ARG abundance (quantified as the sum of RPKM) and ARG  $\alpha$ -  
354 diversity (quantified by Shannon index) across sampling time points. In both the antibiotic-naive  
355 and treated groups, we found that total ARG abundance and diversity increased from T1 to T3-T4,  
356 before descending at T5 until 6 month corrected age (T6). In treated infants, total ARG abundance  
357 and diversity increased more rapidly upon antibiotic treatment, whereas the changes were  
358 generally more gradual over time points in naive infants (Fig. 3A, B). To identify the direct effects  
359 of early antibiotic treatment on the total ARG abundance and diversity, we used a generalized  
360 linear mixed model with individual set as random effect while correcting for age (DOL). Early  
361 antibiotics were significantly associated with increase in both total ARG abundance (LME model:  
362  $p.adj = 0.008$ ) and Shannon diversity ( $p.adj = 0.008$ ) after treatment (T2) as compared to baseline  
363 (T1) samples in preterm infant nasopharyngeal resistome. Moreover, the significant effect of  
364 antibiotics on increase in total ARG abundance was observed until T3 ( $p.adj = 0.001$ ).  
365 Nonetheless, such effects were short-lived, as we did not identify significant differences at any of  
366 the following timepoints compared to baseline (T1) in antibiotic-treated group. In naive group of  
367 infants, no significant differences appeared in total ARG abundance and  $\alpha$ -diversity between any  
368 timepoints during the first 6 months of life while correcting for age.

369

370  $\beta$ -diversity analysis revealed that the overall resistome composition did significantly differ  
371 between the two groups (PERMANOVA:  $R^2 = 8.2\%$ ,  $p.adj = 0.02$ ) at baseline (T1). Infants that  
372 received early antibiotic treatment had more dissimilar (dispersed) resistome composition than the  
373 naive infants (Fig. 4A). Further, we observed that nasopharyngeal samples were clustering more  
374 effectively according to birth weight group ( $R^2 = 16.4\%$ ,  $p.adj = 0.012$ ; Supplementary Fig. 5A).  
375 No significant difference in baseline resistome composition between antibiotic-treated and naive  
376 infants ( $R^2 = 0.03$ ,  $p.adj = 0.319$ ) was found while correcting for birth weight group. However, a  
377 minor yet significant compositional difference in resistome was still observed between the two  
378 groups ( $R^2 = 5.9\%$ ,  $p.adj = 0.03$ ) directly after the cessation of early antibiotic treatment (T2) (Fig.  
379 4B). Moreover, the compositional changes due to early antibiotics at T2 were more discernible in  
380 infants whose mothers had received prenatal antibiotics during pregnancy (prenatal + early ( $n = 8$ )  
381 vs naive ( $n = 20$ ):  $R^2 = 10.09\%$ ,  $p.adj = 0.006$ ; Supplementary Fig. 5B). Following treatment,  
382 resistome normalized rapidly as no significant differences in composition between naive and  
383 treated group were detected from T3 until 6 months of corrected age (T6) (Fig. 4C).

384

### 385 **Clinical covariates associated with preterm infant nasal resistome composition development**

386 The associations between resistome composition and clinical factors were analyzed using  
387 PERMANOVA-tests (all  $p.adj \leq 0.001$ ) and principal component analysis (PCA) across all  
388 nasopharyngeal samples from preterm birth until 6 months corrected age, using centered log-ratio  
389 (CLR) transformed relative abundance data (Supplementary Fig. 1). The PCA was explained by  
390 two main principal components, describing 18.5% and 17.7% of the variation, respectively.  
391 Besides inter-individual variation (PERMANOVA (univariate):  $R^2 = 37.7\%$ ,  $p.adj = 0.001$ ), age  
392 was the other significant covariate found to be associated with overall resistome composition in  
393 the overarching cohort. A moderate yet statistically significant effect was observed with GA ( $R^2$   
394 4%,  $p.adj = 0.001$ ) and BW group ( $R^2$  4.1%,  $p.adj = 0.001$ ), which are highly correlated variables  
395 (as indicated by chi-squared test:  $p < 2.2e-16$ ). Mode of delivery, gender, and intrapartum  
396 antibiotic treatment did not appear to shape the overall nasopharyngeal resistome composition  
397 (Supplementary Table 3). Covariates that differed between the antibiotic-treated and naive infants  
398 at baseline and that also had a significant association with the overall resistome composition (GA

399 and prenatal antibiotics ( $R^2$  4.5%,  $p.adj = 0.001$ ) were used for correction in the subsequent  
400 multivariate analysis.

401  
402 Multivariate PERMANOVA analysis showed that the largest effect on overall resistome  
403 composition remained attributed to inter-individual variation ( $R^2$  30.03%), followed by PMA ( $R^2$   
404 9.7%) and chronological age group (sampling time points) ( $R^2$  8.4%, all  $p.adj = 0.001$ ). Although  
405 we observed a high degree of variation between samples with no obvious discrete spatial clustering  
406 in relation to these variables upon visual inspection (Fig. 5A-C). We found no significant influence  
407 of early antibiotics ( $R^2 = 0.6%$ ,  $p.adj = 0.31$ ) and its duration ( $R^2 = 1.2%$ ,  $p.adj = 0.009$ ) after  
408 adjusting for confounding variables. However, we observed a significant effect of prenatal  
409 antibiotics ( $R^2 = 3.6%$ ,  $p.adj = 0.001$ ) and its duration (categorized) ( $R^2 = 4.4%$ ,  $p.adj = 0.001$ ) in  
410 shaping the overall nasopharyngeal resistome composition in preterm birth. To account for the  
411 likelihood of prenatal antibiotics with early antibiotics exposure in our analyses (Chi-squared test,  
412  $p < 0.0003$ ), we further categorized our early antibiotic-treated group based on whether the mothers  
413 had also received antibiotics during pregnancy (prenatal + early,  $n = 9$ ) or not (only early,  $n = 6$ ).  
414 Post-hoc analysis revealed significant differences in the overall resistome composition of infants  
415 having exposure to both prenatal and early life antibiotics (prenatal + early) compared to infants  
416 receiving either only early or naive (no prenatal + no early) ( $p.adj = 0.001$ ; Fig. 5D). No difference  
417 in overall resistome composition was observed between the only early antibiotics and naive group  
418 ( $p.adj = 0.14$ ), underlining the relevance of prenatal antibiotic exposure in shaping the overall  
419 resistome composition of preterm infants during the first six months of life (corrected).

420  
421 **Unsupervised clustering classifies nasopharyngeal samples into distinct clusters based on**  
422 **composition differences**

423 Next, we determined whether the nasopharyngeal samples were organized into clusters (i.e.,  
424 resistotypes) according to their ARG abundance profiles using an unsupervised method, i.e., DMM  
425 models. This way, we identified three distinct clusters (or resistotypes) that were driven by specific  
426 ARGs. Majority of our samples (~90%) were classified into two main resistotypes, R1 and R2. No  
427 significant difference in the sequencing depth, was detected between the two Resistotypes  
428 (pairwise Wilcox test:  $p = 0.14$ ), suggesting that sequencing depth was not driving their separation.  
429 The first resistotype (R1) was characterized by a predominance of fluoroquinolone resistance

430 genes (*patA*, *patB*) and *rlmA(II)* (Supplementary Fig. 6). On the other hand, resistotype R2 was  
431 driven mainly by relatively higher abundance of the beta-lactam resistance gene *SST-1* and the  
432 tetracycline resistance gene *tet(41)*. While resistotype R3 was enriched with the blaZ  $\beta$ -lactamase  
433 resistance gene. The identified resistotypes did vary in their overall ARG composition ( $\beta$ -  
434 diversity), as shown in a PCA plot (Fig. 6A). Further, PERMANOVA analysis showed that  
435 resistotypes were more strongly (larger effect size) and significantly associated with the resistome  
436 than other clinical covariates or early-life factors in preterm infants during the first 6 months of  
437 life ( $R^2 = 19.05\%$ ,  $p_{adj} = 0.001$ ).

438

### 439 **Nasopharyngeal resistome composition of preterm infants is linked to microbial composition**

440 To investigate the impact of microbial composition on resistome composition, we taxonomically  
441 profiled the infant nasopharyngeal samples using MetaPhlAn3 (manuscript under preparation,  
442 Rajar et al. 2023). We noticed that microbial community types, as determined by DMM models,  
443 were highly linked with the resistotypes (chi-square test:  $p < 0.0001$ ). Resistotype R2 was very  
444 specific to the *Serratia* community type. Resistotype R1 was associated with the community types  
445 mainly driven by *Streptococcus* and *Gemella* and Resistotype R3 was enriched in community types  
446 predominated by *Staphylococcus* and *Cutibacterium* (Fig. 6B). These community types were also  
447 significantly associated to the overall composition of the resistome (PERMANOVA:  $R^2 = 13.01\%$ ,  
448  $p_{adj} = 0.001$ ; Fig. 6C). Additionally, result from Procrustes analysis showed a strong and  
449 significant association between the microbial taxa abundance and ARG abundance profiles  
450 (correlation coefficient ( $r$ ) = 0.85,  $p = 0.001$ ,  $m^2 = 0.26$ ), underlining that bacterial community  
451 composition structured the resistome composition in the nasopharyngeal microbiome of preterm  
452 infants. However, the correlation was more robust in the antibiotic-naive infants ( $r = 0.94$ ,  $p =$   
453  $0.001$ ,  $m^2 = 0.11$ ) as compared to antibiotic-treated infants ( $r = 0.85$ ,  $p = 0.001$ ,  $m^2 = 0.27$ ; Fig.  
454 7A, B).

455

456 Next, we performed pairwise Spearman's correlation analysis between ARG and microbial species  
457 abundances to predict the origin or potential microbial host of ARGs in nasopharyngeal samples.  
458 As expected, the strongest positive correlation ( $r \geq 0.8$ ,  $p_{adj} < 0.05$ ) was found between the  
459 Resistotype 2 enriched ARGs i.e., *SST-1*, *AAC(6')-Ic*, *tet(41)*, *MexI* and *Serratia*  
460 *marcescens/nematodiphila*. The highly prevalent and Resistotype R1 associated ARGs like *patA*,



461 *patB* and *rlmA(II)* were found to have strong co-occurrence with *Streptococcus mitis/oralis* and  
462 *Gemella haemolysans/sanguinis* (Fig. 7C). While *Staphylococcus aureus* and *epidermidis*  
463 correlated only with *blaZ* beta-lactam ARG. Interestingly, these associations are in accordance  
464 with information present in the literature and AMR gene databases on the known microbial host  
465 of these ARGs, highlighting the potential of such analysis approach to predict the origin of ARG  
466 in metagenomes. However, this analysis is only applicable on the features (ARGs or taxa) that are  
467 present in high prevalence across samples to reduce the bias due to false ranking of features with  
468 many zero values.

469

470 Lastly, we also examined whether some of these strongly associated bacterial species with ARGs  
471 would also directly correlate with our overall resistome outcomes. We plotted the relative  
472 abundance of each species versus the total ARG abundance (sum of RPKM) and the number of  
473 unique ARGs from all samples. Our analysis revealed that only the relative abundance of  
474 *Streptococcus mitis* (rmcorr correlation coefficient ( $r^m = 0.29$ ,  $p.adj = 0.01$ ) showed a moderately  
475 linear association with the total ARG abundance while adjusting for inter-individual variability  
476 (Supplementary Fig. 7). However, we did not observe a significant correlation between the relative  
477 abundance of *S. mitis* and the number of unique ARGs ( $r^m = 0.19$ ,  $p.adj = 0.12$ ). To statistically  
478 assess whether *S. mitis* relative abundance explains the dispersion of resistomes across samples,  
479 we conducted PERMANOVA test and found that *Streptococcus mitis* significantly explains  
480 11.69% ( $p.adj = 0.001$ ) of the variation in the overall resistome composition. Only *Serratia*  
481 *marcescens* ( $R^2 = 11.8\%$ ,  $p.adj = 0.001$ ) can describe rather similar variation than *Streptococcus*  
482 *mitis* in the resistome, though it was not found to be positively associated with any overall  
483 resistome outcomes in our data.

484

## 485 **DISCUSSION**

486 In this study, we (1) examined the dynamics of nasopharynx resistome development in preterm  
487 infants, (2) assessed the impact of early antibiotics on its trajectory, and (3) explored the potential  
488 associations between clinical variables and the overall resistome composition in preterm infants.  
489 We detected ARGs in nearly all (~95%) samples from both the antibiotic-treated and naive groups,  
490 including some potentially high-risk ARGs. Our analysis revealed that early antibiotic exposure  
491 was transiently associated with an increase in both the diversity and total abundance of ARGs, as

492 well as changes in the overall resistome composition. The resistome composition was strongly  
493 correlated with the microbiome composition, with inter-individual variation and age explaining  
494 most of the compositional variation in the resistomes. Among the clinical variables, prenatal  
495 maternal antibiotic use, which was a subgroup found only in the antibiotic group of preterm infants,  
496 appeared to be one of the most significant factors influencing the overall nasopharyngeal resistome  
497 composition in preterm infants.

498  
499 The nasopharynx represents a highly accessible microbial community and also serves as a crucial  
500 diagnostic window in combating respiratory infections and AMR [56]. It is also the primary  
501 ecological niche for common respiratory pathogens [6]. To date, one of the closest attempts for an  
502 in-depth investigation of ARGs in the respiratory tract using a culture-free approach identified  
503 ARGs in 86% of the samples. However, this was based on the microbiomes of the anterior nares  
504 and from an adult population (Human Microbiome Project) [57], which are different from the  
505 groups and collection sample sites used in our study. In a recent study focusing on infants, ARGs  
506 were detected in 64% of nasopharynx samples [58]. Samples in this study were enriched for  
507 streptococci, thus precluding general conclusions on the extension of ARG presence in the natural  
508 microbiome. In our study, ARGs were present in 95% of the preterm infant samples from the  
509 nasopharynx, regardless of whether they were exposed to antibiotics or not. Although our study  
510 differs in several aspects from the ones above, the combined findings indicate that the respiratory  
511 tract is likely an important reservoir of ARGs and an integral component of human respiratory  
512 microbiomes. For preterm infants, the detection of ARGs in the nasopharynx during the first few  
513 hours of life might occur due to microbiome colonization from multiple sources such as  
514 transgenerational transfer from mothers and the environment [31, 59]. In addition, preterm infants  
515 usually have longer stay in the hospital than term infants after birth. This most likely contributes  
516 to an increased risk for respiratory tract colonization with drug resistant microorganisms, as  
517 recently reported in relation to the gut resistome [9].

518 We found that ARGs conferring multidrug resistance via efflux pumps encompassed a large  
519 proportion of the nasopharyngeal resistome in the preterm infants. Similar findings have been  
520 previously reported in the gut and oral resistome of preterm and term infants by other metagenomic  
521 studies [10, 59-61]. In addition, we detected many of the same aminoglycoside-modifying  
522 enzymes, tetracycline protection proteins, and beta-lactamases ARGs identified previously in the

523 nasopharynx of South African infants, despite the fact that the samples in the South African study  
524 were culture-enriched for streptococci and were from a country with higher AMR challenges than  
525 the origin of samples used in the present study. Such similarities in resistome composition may, at  
526 least in part, be explained by the large proportion of streptococci found in the natural microbial  
527 communities of the infant nasopharynx. Together, these findings suggest the existence of a core  
528 infant nasopharyngeal resistome.

529 *Staphylococcus aureus* is one of the leading pathogens causing neonatal sepsis, and one of the  
530 priority drug resistant pathogens by the WHO [62]. We found a strong correlation between the  
531 abundance of *Staphylococcus aureus/epidermidis* and the PC1-beta-lactamase (*blaZ*) gene. This is  
532 a gene found in the chromosome and mobile plasmids of staphylococci, thus representing high risk  
533 ARGs. Another strong correlation was found between *Serratia* species and specific resistance  
534 genes such as *SST-1*, which confers resistance against cephalosporins. Other resistance genes  
535 correlated with this group included aminoglycoside and tetracycline resistance genes. This is of  
536 relevance, since *Serratia spp.* are associated with hospital infection outbreaks, especially in NICUs  
537 [63]. However, it is uncertain whether the identified *Serratia spp.* in the study were acquired from  
538 the hospital environment or not, as this group of bacteria has been also found as part of the nasal  
539 microbiome in healthy adults [64].

540 Our results showed a significant effect of early broad-spectrum antibiotic regimen  
541 (ampicillin + gentamicin) on the total abundance, diversity, and composition of ARGs in the  
542 nasopharynx of preterm infants. These effects on the resistome were short-lived and only observed  
543 directly after the cessation of antibiotic treatment (T2). In a recent study investigating the impact  
544 of antibiotics used in the first week of life on intestinal resistome, Reyman et al. reported that three  
545 other antibiotic combinations (penicillin + gentamicin, co-amoxiclav + gentamicin, and  
546 amoxicillin + cefotaxime) also had a transient impact on the infant resistome [11]. As the effects  
547 of antibiotics on microbiome and resistome composition are unique to each individual and  
548 dependent on the antibiotics used and the body site, caution should be taken when extending the  
549 interpretation of such general findings. Interestingly, the transient effects of  
550 ampicillin + gentamicin on the nasopharyngeal resistome were more discernible in preterm infants  
551 whose mothers were exposed to antibiotics during pregnancy before delivery (prenatal antibiotics).  
552 Moreover, we also observed a more profound effect of prenatal antibiotic exposure on the

553 nasopharyngeal resistome of preterm infants at T2 (after cessation of early antibiotics), which  
554 persisted until the last sampling time point before discharge from the NICU (T4-T5), but not at 6  
555 months corrected age. Although very few studies have focused on the association between prenatal  
556 antibiotic exposure and the resistome, a recent study based on the gut indicates a possible  
557 correlation of prenatal antibiotics with increased abundance of ARGs, at least when used 5 to 6  
558 weeks before birth [10]. Additionally, in concordance with other studies on the infant gut [15, 59],  
559 our data showed no significant impact of intrapartum antibiotic therapy on the overall  
560 nasopharyngeal resistome composition. The nasopharyngeal samples of preterm infants at 6  
561 months corrected age were still clustered according to the birth weight group (similar to T1,  
562 Supplementary Fig. 5C), suggesting a strong but transient influence of other environmental factors,  
563 such as antibiotic exposures. Nonetheless, there is a paucity of relevant studies on nasopharynx  
564 that can corroborate our findings, as they are mainly restricted to the infant gut resistome [9, 11,  
565 14, 15, 31, 59].

566 One of the strengths of this study is that it presents the most comprehensive information to date on  
567 the characteristics and dynamics of resistome development in the nasopharynx, including results  
568 for the first six months of life (corrected) – a critical period of airway development in infants. In  
569 addition, the naive group was enrolled at the same time as our treated group from the same hospital,  
570 thus removing potential confounding differences in enrollment years or sampling duration. On the  
571 other hand, the small number of samples at each time point and inter-individual response to  
572 antibiotic exposure may be insufficient to unravel significant associations between antibiotic  
573 regimen and resistome development. Moreover, the baseline samples for the antibiotic-treated  
574 group ideally would represent samples taken before treatment. However, antibiotic treatment  
575 initiation for sEONS and clinical stabilization of the preterm infants were a high priority that  
576 needed to precede nasopharynx sample collection. To partially address this issue, we included only  
577 preterm infants with samples collected no more than 24 hours after antibiotic initiation. Even so,  
578 this choice may have resulted in the selection of more clinically stable infants, which could  
579 possibly limit the generalizability of our results. Despite this approach, the differences in resistome  
580 changes between baseline and after treatment found in our study may underestimate the true impact  
581 of antibiotics on the resistome. We rigorously adjusted for such confounding factors and inter-  
582 individual variation using multivariable methods or matched analysis to ensure the reliability of  
583 our observations as a consequence of early antibiotics. Nevertheless, we cannot isolate early

584 antibiotic effects from other adverse pre-life events that coincide with premature birth, such as  
585 prenatal and antepartum maternally administered antibiotics. Furthermore, for low biomass  
586 samples such as those used in our study, sensitive approaches such as deep sequencing are  
587 necessary to obtain meaningful microbiome information. This increases the risk of detecting  
588 potential DNA contaminants. Although it is not possible to completely eliminate the potential  
589 identification of contaminants, we adopted various control procedures, starting from sample  
590 collection to analysis, to ensure that the data presented in our study is a reliable representation of  
591 the ARGs found in the nasopharynx.

592

593 In conclusion, our findings suggest that early-life treatment with broad-spectrum antibiotics for  
594 sEONS can cause acute perturbations in the nasopharyngeal resistome of preterm infants.  
595 However, these perturbations were transient, as we did not observe a significant alteration in  
596 diversity, abundance, composition, or carriage persistence of high-risk ARGs in treated preterm  
597 infants compared to controls at 6 months of corrected age. The individual variability in the uneven  
598 distribution of high-risk ARGs and the changes in the resistome following antibiotic treatment  
599 highlight the potential benefits of developing resistome profiling tools for point-of-care use, which  
600 could aid in informing and tailoring antibiotic treatment decisions [65]. Future multicenter studies  
601 with long-term follow-up are needed to assess the ecological side-effects of different antibiotic  
602 regimens, including alternatives to broad-spectrum antimicrobial therapy on infant respiratory  
603 resistome development. This study also emphasizes the need to further investigate the  
604 underappreciated effect of different antibiotic exposures (pre, peri, post-birth) during the critical  
605 period of the nasopharyngeal microbiome and resistome development in infants, and the  
606 subsequent development of respiratory infections later in life. Understanding the development of  
607 the respiratory resistome and the impact of antibiotic therapies is relevant for striking a balance  
608 between the conflicting goals of minimizing the risk of serious infectious complications to preterm  
609 birth on one side, and minimizing the selection of organisms with ARGs on the other. Such  
610 knowledge can also help to strengthen antibiotic stewardship programs. In general, the  
611 disproportionate contribution of respiratory pathogens to drug-resistant deaths at all ages  
612 highlights the importance of future studies focusing on this yet rather unexplored reservoir of  
613 pathogens and ARGs.

614

## 615 REFERENCES

- 616 1. Murray, C., et al., & Naghavi, M.(2022). *Global burden of bacterial antimicrobial*  
617 *resistance in 2019: a systematic analysis*. The Lancet. **399**(10325): p. 629-655.
- 618 2. Tedijanto, C., et al., *Estimating the proportion of bystander selection for antibiotic*  
619 *resistance among potentially pathogenic bacterial flora*. Proceedings of the National  
620 Academy of Sciences, 2018. **115**(51): p. E11988-E11995.
- 621 3. Von Wintersdorff, C.J., et al., *Dissemination of antimicrobial resistance in microbial*  
622 *ecosystems through horizontal gene transfer*. Frontiers in microbiology, 2016. **7**: p. 173.
- 623 4. Kilian, M., et al., *Parallel evolution of Streptococcus pneumoniae and Streptococcus mitis*  
624 *to pathogenic and mutualistic lifestyles*. MBio, 2014. **5**(4): p. e01490-14.
- 625 5. Pailhoriès, H., et al., *Antibiotic resistance in chronic respiratory diseases: from*  
626 *susceptibility testing to the resistome*. European Respiratory Review, 2022. **31**(164).
- 627 6. Man, W.H., W.A. de Steenhuijsen Piters, and D. Bogaert, *The microbiota of the respiratory*  
628 *tract: gatekeeper to respiratory health*. Nature Reviews Microbiology, 2017. **15**(5): p. 259-  
629 270.
- 630 7. Huang, Y.J., *Nasopharyngeal microbiota: gatekeepers or fortune tellers of susceptibility*  
631 *to respiratory tract infections?* 2017, American Thoracic Society. p. 1504-1505.
- 632 8. Dhariwal, A., et al., *Differential response to prolonged amoxicillin treatment: long-term*  
633 *resilience of the microbiome versus long-lasting perturbations in the gut resistome*. Gut  
634 Microbes, 2023. **15**(1): p. 2157200.
- 635 9. Gasparrini, A.J., et al., *Persistent metagenomic signatures of early-life hospitalization and*  
636 *antibiotic treatment in the infant gut microbiota and resistome*. Nature microbiology, 2019.  
637 **4**(12): p. 2285-2297.
- 638 10. Li, X., et al., *The infant gut resistome associates with E. coli, environmental exposures, gut*  
639 *microbiome maturity, and asthma-associated bacterial composition*. Cell Host & Microbe,  
640 2021. **29**(6): p. 975-987. e4.
- 641 11. Reyman, M., et al., *Effects of early-life antibiotics on the developing infant gut microbiome*  
642 *and resistome: a randomized trial*. Nat Commun, 2022. **13**(1): p. 893.
- 643 12. Thänert, R., et al., *The resistance within: Antibiotic disruption of the gut microbiome and*  
644 *resistome dynamics in infancy*. Cell Host & Microbe, 2022. **30**(5): p. 675-683.
- 645 13. Gasparrini, A.J., et al., *Antibiotic perturbation of the preterm infant gut microbiome and*  
646 *resistome*. Gut microbes, 2016. **7**(5): p. 443-449.
- 647 14. Gibson, M.K., et al., *Developmental dynamics of the preterm infant gut microbiota and*  
648 *antibiotic resistome*. Nat Microbiol, 2016. **1**: p. 16024.
- 649 15. Lebeaux, R.M., et al., *The infant gut resistome is associated with E. coli and early-life*  
650 *exposures*. BMC Microbiol, 2021. **21**(1): p. 201.
- 651 16. Tripathi, N., C.M. Cotten, and P.B. Smith, *Antibiotic use and misuse in the neonatal*  
652 *intensive care unit*. Clinics in perinatology, 2012. **39**(1): p. 61-68.
- 653 17. Mukhopadhyay, S., E.C. Eichenwald, and K.M. Puopolo, *Neonatal early-onset sepsis*  
654 *evaluations among well-appearing infants: projected impact of changes in CDC GBS*  
655 *guidelines*. Journal of perinatology, 2013. **33**(3): p. 198-205.
- 656 18. Rajar, P., et al., *Antibiotic stewardship in premature infants: a systematic review*.  
657 Neonatology, 2020. **117**(6): p. 673-686.
- 658 19. Kuppala, V.S., et al., *Prolonged initial empirical antibiotic treatment is associated with*  
659 *adverse outcomes in premature infants*. The Journal of pediatrics, 2011. **159**(5): p. 720-  
660 725.

- 661 20. Novitsky, A., et al., *Prolonged early antibiotic use and bronchopulmonary dysplasia in*  
662 *very low birth weight infants*. American journal of perinatology, 2015. **32**(01): p. 043-048.
- 663 21. Ting, J.Y., et al., *Association between antibiotic use and neonatal mortality and*  
664 *morbidities in very low-birth-weight infants without culture-proven sepsis or necrotizing*  
665 *enterocolitis*. JAMA pediatrics, 2016. **170**(12): p. 1181-1187.
- 666 22. Ting, J.Y., et al., *Duration of initial empirical antibiotic therapy and outcomes in very low*  
667 *birth weight infants*. Pediatrics, 2019. **143**(3).
- 668 23. Cotten, C.M., et al., *Prolonged duration of initial empirical antibiotic treatment is*  
669 *associated with increased rates of necrotizing enterocolitis and death for extremely low*  
670 *birth weight infants*. Pediatrics, 2009. **123**(1): p. 58-66.
- 671 24. Esaiassen, E., et al., *Antibiotic exposure in neonates and early adverse outcomes: a*  
672 *systematic review and meta-analysis*. Journal of antimicrobial chemotherapy, 2017. **72**(7):  
673 p. 1858-1870.
- 674 25. Li, Y., et al., *Early use of antibiotics is associated with a lower incidence of necrotizing*  
675 *enterocolitis in preterm, very low birth weight infants: the NEOMUNE-NeoNutriNet*  
676 *cohort study*. The Journal of pediatrics, 2020. **227**: p. 128-134. e2.
- 677 26. Vatne, A., et al., *Early empirical antibiotics and adverse clinical outcomes in infants born*  
678 *very preterm: a population-based cohort*. The Journal of Pediatrics, 2023. **253**: p. 107-114.  
679 e5.
- 680 27. Tamburini, S., et al., *The microbiome in early life: implications for health outcomes*. Nature  
681 medicine, 2016. **22**(7): p. 713-722.
- 682 28. Bailey, L.C., et al., *Association of antibiotics in infancy with early childhood obesity*.  
683 JAMA pediatrics, 2014. **168**(11): p. 1063-1069.
- 684 29. Arrieta, M.-C., et al., *Early infancy microbial and metabolic alterations affect risk of*  
685 *childhood asthma*. Science translational medicine, 2015. **7**(307): p. 307ra152-307ra152.
- 686 30. Pärnänen, K., et al., *Maternal gut and breast milk microbiota affect infant gut antibiotic*  
687 *resistome and mobile genetic elements*. Nature Communications, 2018. **9**(1): p. 3891.
- 688 31. Yassour, M., et al., *Natural history of the infant gut microbiome and impact of antibiotic*  
689 *treatment on bacterial strain diversity and stability*. Science translational medicine, 2016.  
690 **8**(343): p. 343ra81-343ra81.
- 691 32. Kang, H.M. and J.H. Kang, *Effects of nasopharyngeal microbiota in respiratory infections*  
692 *and allergies*. Clinical and Experimental Pediatrics, 2021. **64**(11): p. 543.
- 693 33. Teo, S.M., et al., *The infant nasopharyngeal microbiome impacts severity of lower*  
694 *respiratory infection and risk of asthma development*. Cell host & microbe, 2015. **17**(5): p.  
695 704-715.
- 696 34. Toivonen, L., et al., *Antibiotic treatments during infancy, changes in nasal microbiota, and*  
697 *asthma development: population-based cohort study*. Clinical Infectious Diseases, 2021.  
698 **72**(9): p. 1546-1554.
- 699 35. Rajar, P., et al., *Microbial DNA extraction of high-host content and low biomass samples:*  
700 *Optimized protocol for nasopharynx metagenomic studies*. Frontiers in Microbiology,  
701 2022. **13**.
- 702 36. Eisenhofer, R., et al., *Contamination in low microbial biomass microbiome studies: issues*  
703 *and recommendations*. Trends in microbiology, 2019. **27**(2): p. 105-117.
- 704 37. Krueger, F., *Trim Galore: a wrapper tool around Cutadapt and FastQC*. Trim Galore,  
705 2012.



- 706 38. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2*. Nature  
707 methods, 2012. **9**(4): p. 357-359.
- 708 39. Li, H., et al., *The sequence alignment/map format and SAMtools*. bioinformatics, 2009.  
709 **25**(16): p. 2078-2079.
- 710 40. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic  
711 features*. Bioinformatics, 2010. **26**(6): p. 841-842.
- 712 41. Andrews, S., *FastQC: a quality control tool for high throughput sequence data*. 2010,  
713 Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.
- 714 42. Ewels, P., et al., *MultiQC: summarize analysis results for multiple tools and samples in a  
715 single report*. Bioinformatics, 2016. **32**(19): p. 3047-3048.
- 716 43. Alcock, B.P., et al., *CARD 2020: antibiotic resistome surveillance with the comprehensive  
717 antibiotic resistance database*. Nucleic acids research, 2020. **48**(D1): p. D517-D525.
- 718 44. Doster, E., et al., *MEGARes 2.0: a database for classification of antimicrobial drug,  
719 biocide and metal resistance determinants in metagenomic sequence data*. Nucleic acids  
720 research, 2020. **48**(D1): p. D561-D569.
- 721 45. Allaire, J., *RStudio: integrated development environment for R*. Boston, MA, 2012.  
722 **770**(394): p. 165-171.
- 723 46. R Core Team, R. and R.C. Team, *R: a language and environment for statistical computing*.  
724 *R Foundation for Statistical Computing; 2020*. 2021.
- 725 47. McMurdie, P.J. and S. Holmes, *phyloseq: an R package for reproducible interactive  
726 analysis and graphics of microbiome census data*. PloS one, 2013. **8**(4): p. e61217.
- 727 48. Wickham, H., *ggplot2*. Wiley interdisciplinary reviews: computational statistics, 2011.  
728 **3**(2): p. 180-185.
- 729 49. Oksanen, J., et al., *Vegan: Community Ecology Package (R Package Version 2.6-2)*. 2022.
- 730 50. Holmes, I., K. Harris, and C. Quince, *Dirichlet multinomial mixtures: generative models  
731 for microbial metagenomics*. PloS one, 2012. **7**(2): p. e30126.
- 732 51. Lahti, L. and S. Shetty, *Introduction to the microbiome R package*. 2018.
- 733 52. Bakdash, J.Z. and L.R. Marusich, *Repeated measures correlation*. Frontiers in psychology,  
734 2017. **8**: p. 456.
- 735 53. Kolde, R. and M.R. Kolde, *Package 'pheatmap'*. R package, 2018. **1**(10).
- 736 54. Zhang, A.-N., et al., *An omics-based framework for assessing the health risk of  
737 antimicrobial resistance genes*. Nature communications, 2021. **12**(1): p. 4765.
- 738 55. Long, Y., et al., *High Carriage Rate of the Multiple Resistant Plasmids Harboring  
739 Quinolone Resistance Genes in Enterobacter spp. Isolated from Healthy Individuals*.  
740 *Antibiotics*, 2022. **11**(1): p. 15.
- 741 56. Flynn, M. and J. Dooley, *The microbiome of the nasopharynx*. Journal of Medical  
742 Microbiology, 2021. **70**(6).
- 743 57. Consortium, H.M.P., *Structure, function and diversity of the healthy human microbiome*.  
744 *Nature*, 2012. **486**(7402): p. 207-214.
- 745 58. Manenzhe, R.I., et al., *Longitudinal changes in the nasopharyngeal resistome of South  
746 African infants using shotgun metagenomic sequencing*. PloS one, 2020. **15**(4): p.  
747 e0231887.
- 748 59. Parnanen, K., et al., *Maternal gut and breast milk microbiota affect infant gut antibiotic  
749 resistome and mobile genetic elements*. Nat Commun, 2018. **9**(1): p. 3891.

750 60. Rahman, S.F., et al., *Machine learning leveraging genomes from metagenomes identifies*  
751 *influential antibiotic resistance genes in the infant gut microbiome*. *MSystems*, 2018. **3**(1):  
752 p. e00123-17.

753 61. Sukumar, S., et al., *The development of the oral resistome during the first decade of life*.  
754 2022.

755 62. Tacconelli, E., *Global Priority List of Antibiotic-Resistant Bacteria to Guide Research,*  
756 *Discovery, and Development*. 2017.

757 63. Acar, J.F., *Serratia marcescens infections*. *Infection Control & Hospital Epidemiology*,  
758 1986. **7**(5): p. 273-280.

759 64. Liu, C.M., et al., *Staphylococcus aureus and the ecology of the nasal microbiome*. *Sci Adv*,  
760 2015. **1**(5): p. e1400216.

761 65. Leggett, R.M., et al., *Rapid MinION profiling of preterm microbiota and antimicrobial-*  
762 *resistant pathogens*. *Nature Microbiology*, 2020. **5**(3): p. 430-442.

763

764

765

766

767

768

769

770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

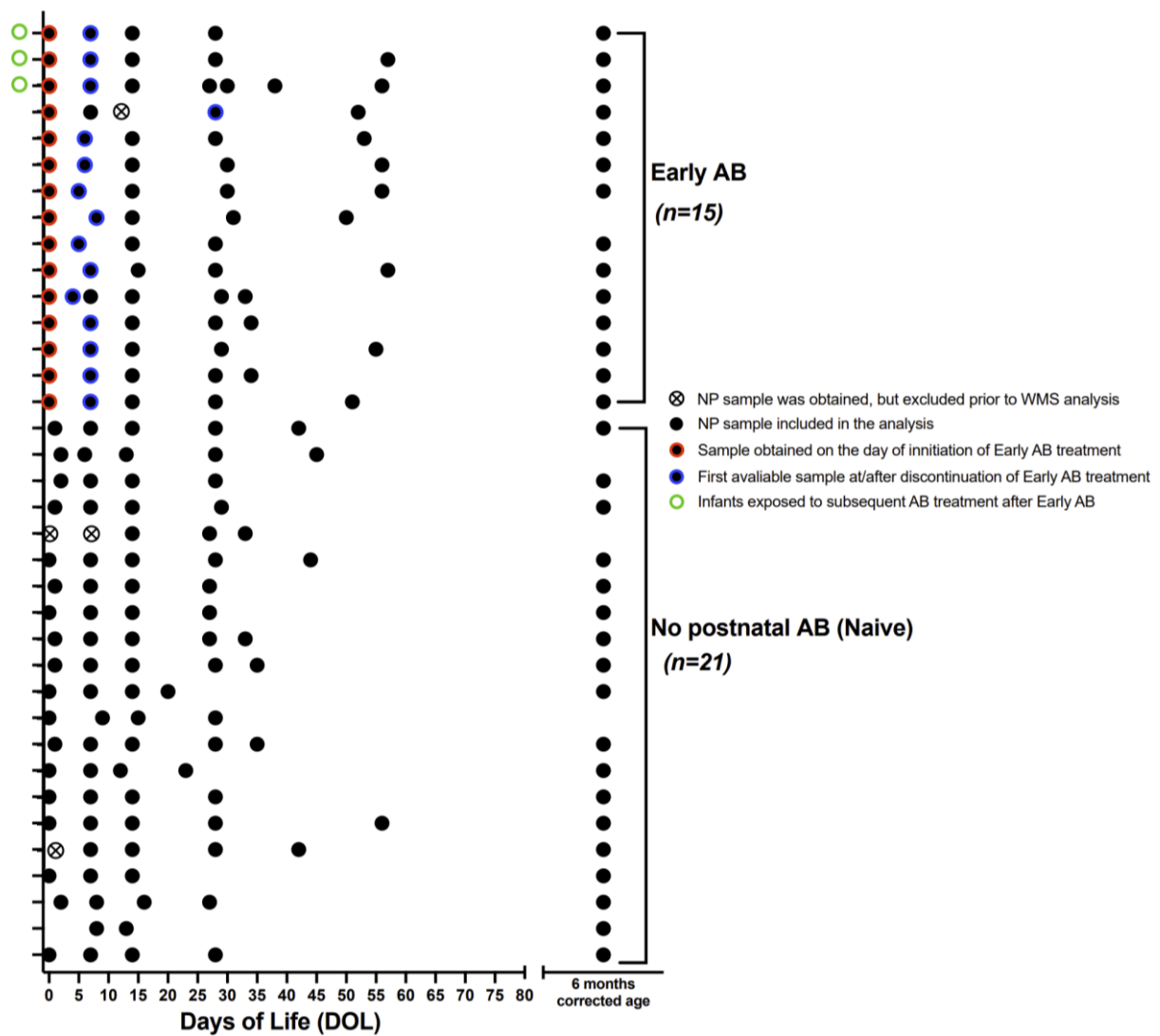
785

786 **ACKNOWLEDGEMENTS**

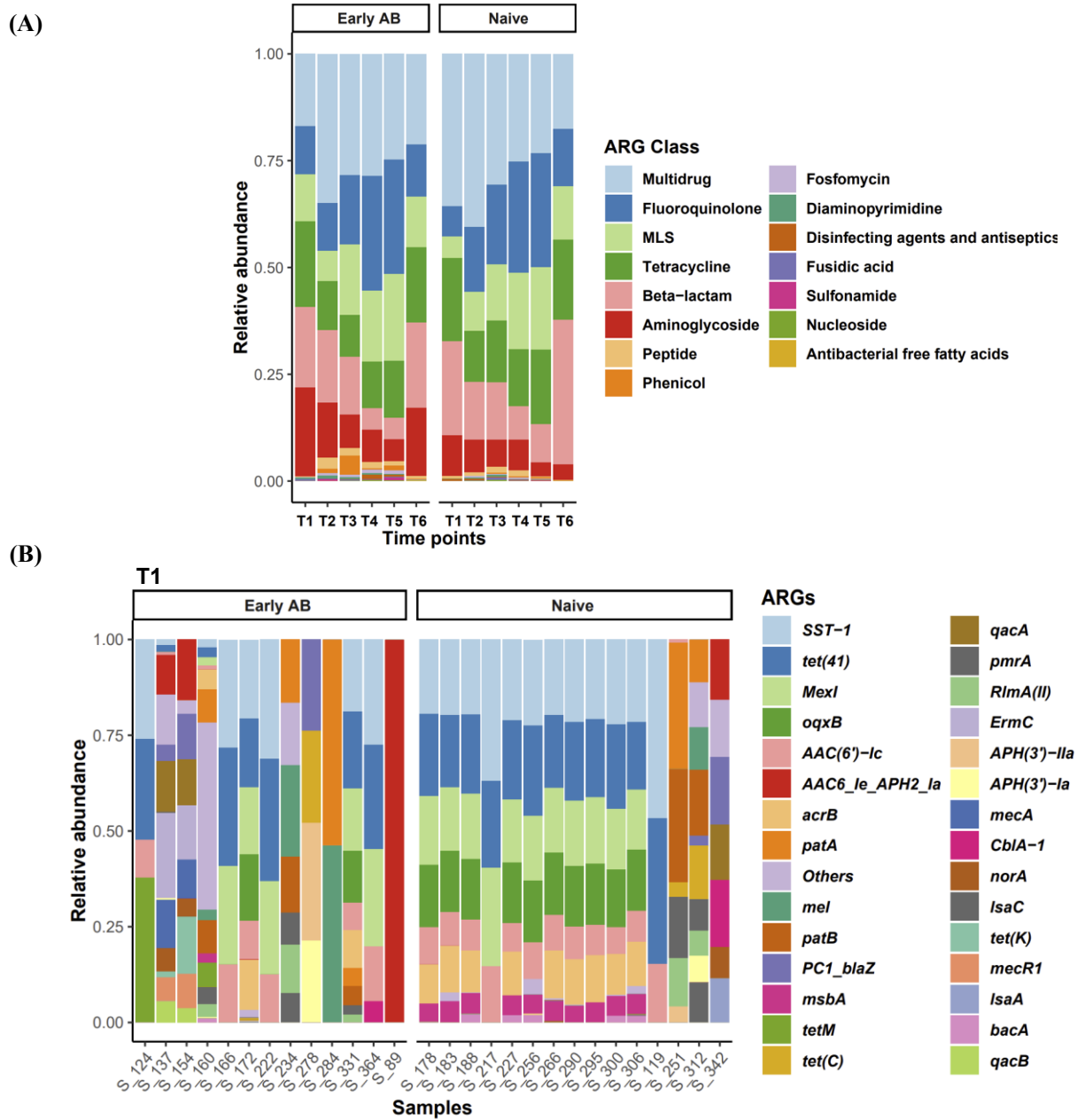
787 We would like to thank the parents of participating infants, and the clinical staff at the Neonatal  
788 Intensive Care units at Ullevål hospital for their assistance with sample collection. The sequencing  
789 service for this work was provided by the Norwegian Sequencing Centre  
790 ([www.sequencing.uio.no](http://www.sequencing.uio.no)) and the computations were performed on resources provided by Sigma2  
791 - the National Infrastructure for High Performance Computing and Data Storage in Norway.

792 **FUNDING**

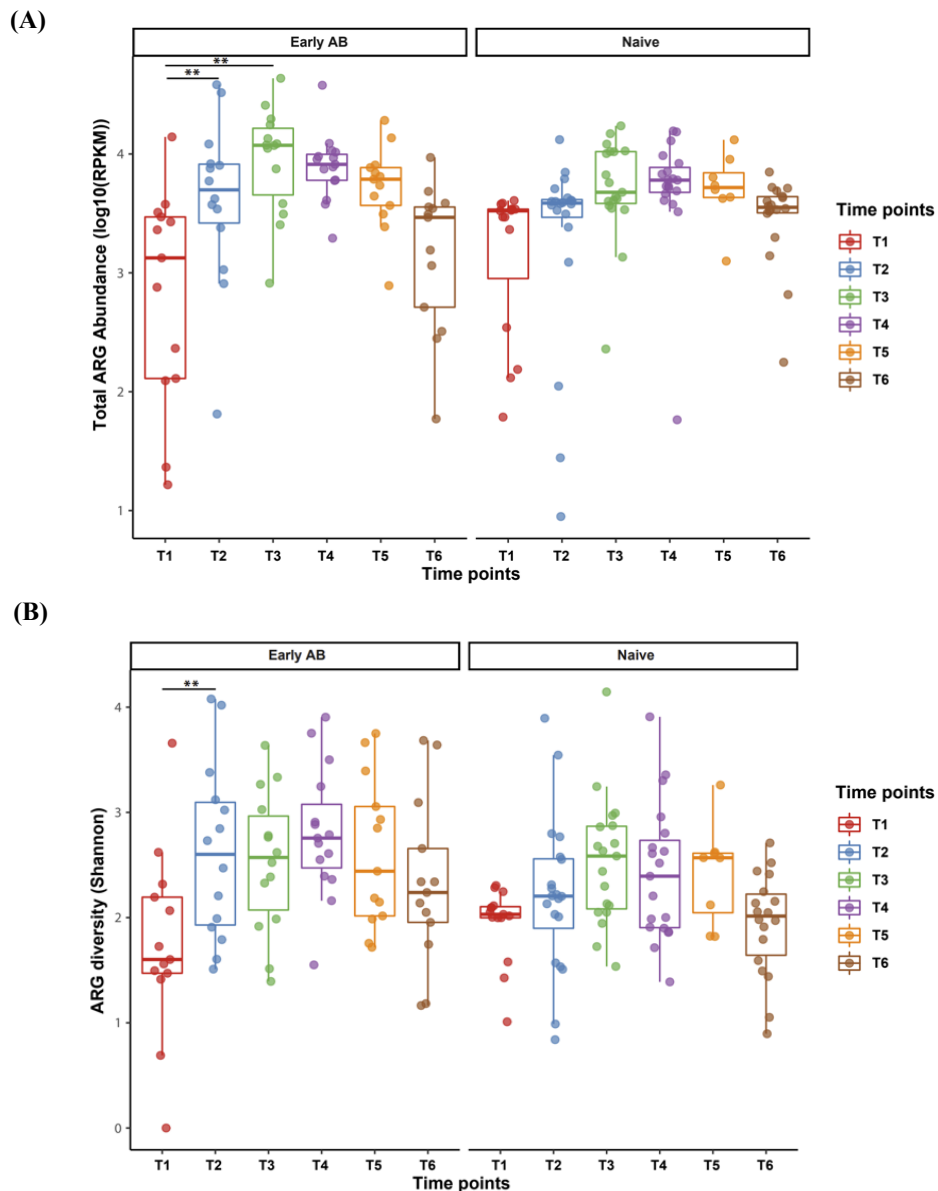
793 This work was financed by the Norwegian Research Council (NFR) (project number: 273833), the  
794 Olav Thon Foundation (project number: 421258), and the Faculty of Dentistry at the University of  
795 Oslo.



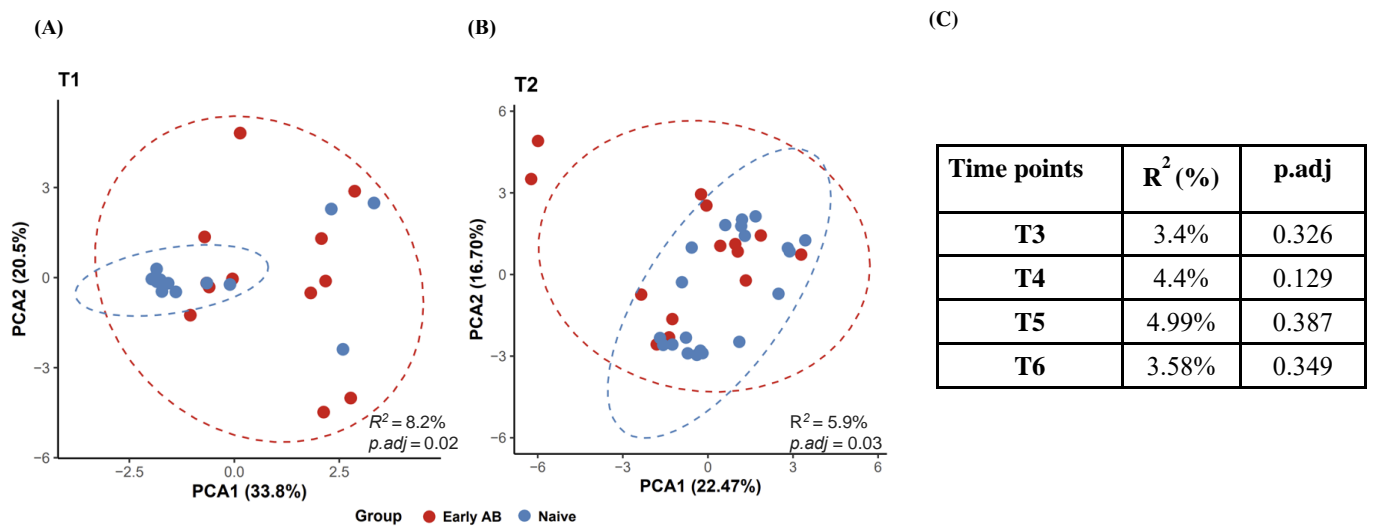
**Fig. 1: Sample collection and antibiotic exposure for preterm infants in this study.** Sampling timeline of 198 nasopharyngeal samples collected from 36 preterm infants over first six months (corrected age), together with information regarding the time of early antibiotic exposure (red and blue colored circles), exposure to subsequent antibiotic treatment (green colored circles) and included/excluded samples. The x-axis represents the days of life (i.e., age) and the y-axis represents the infants. NP = nasopharyngeal; AB = antibiotics.



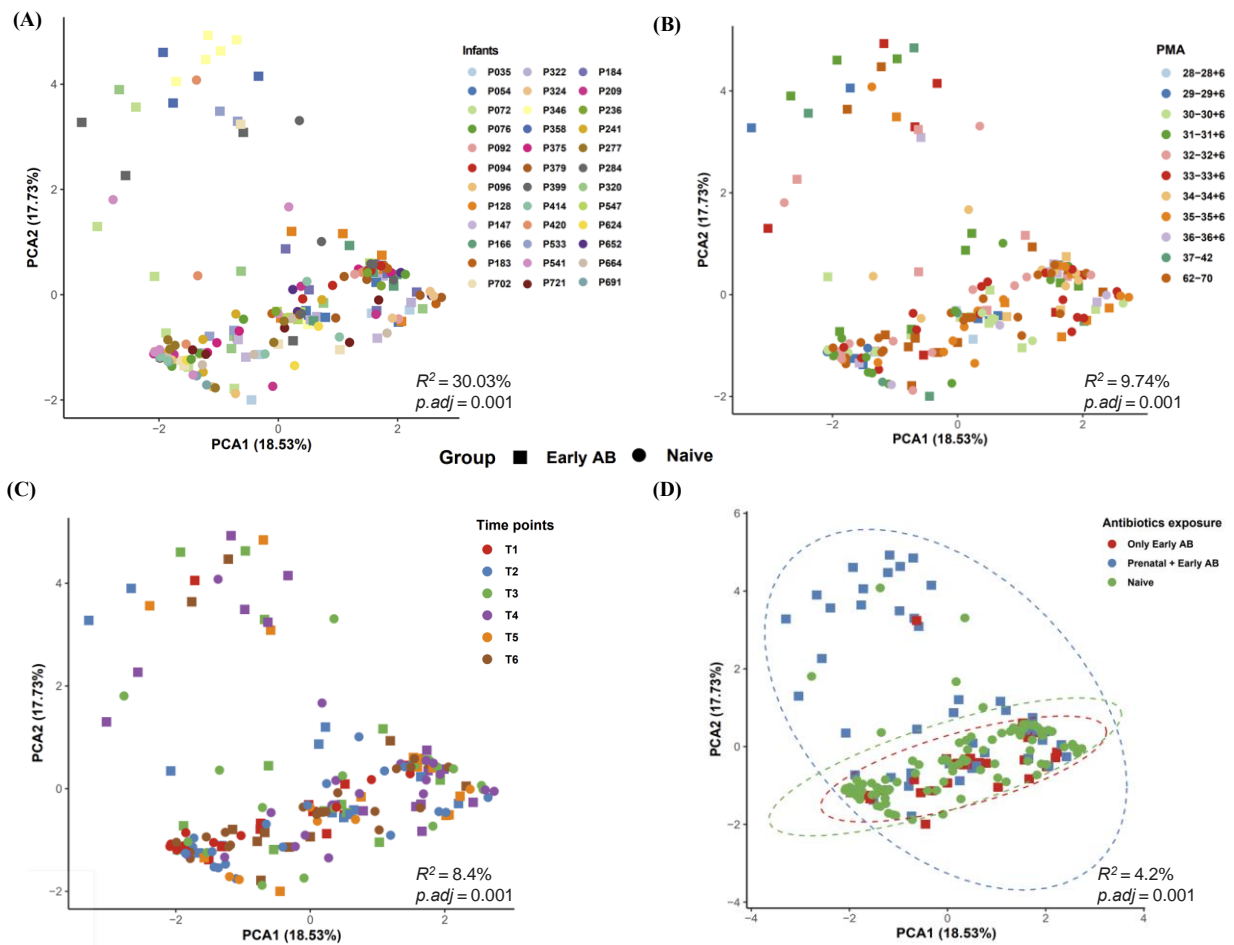
**Fig. 2: Resistome composition in the nasopharynx microbiome of preterm infants.** Stacked bar plot showing the relative abundance of (A) antibiotic resistance genes (ARGs) clustered at the Class level for the naive and antibiotic-treated groups over time; (B) top 30 most abundant ARGs identified across samples at baseline (T1) in early antibiotics treated (left) and naive (right) groups. All other ARGs are grouped into the “Other” category. MLS = macrolide-lincosamide-streptogramin.



**Fig. 3: Impact of early antibiotic treatment on nasopharynx total ARG abundance and ARG  $\alpha$ -diversity** Boxplots showing the (A) total ARG abundance (sum of RPKM) and (B) ARG Shannon diversity per sample, stratified by antibiotic treatment and time points. The lower and upper horizontal box lines correspond to the first and third quartiles, with the line in the middle of the boxes representing the median. Whiskers extending from the boxes represent the range of data points within 1.5 times the interquartile range below the first quartile and within 1.5 times the interquartile range above the third quartile. (Statistical significance: LME mixed effect model. Adjusted p-values (p): \*p < .05, \*\*p < .01 and \*\*\*p < .001). RPKM = reads per kilobase of reference gene per million bacterial reads.

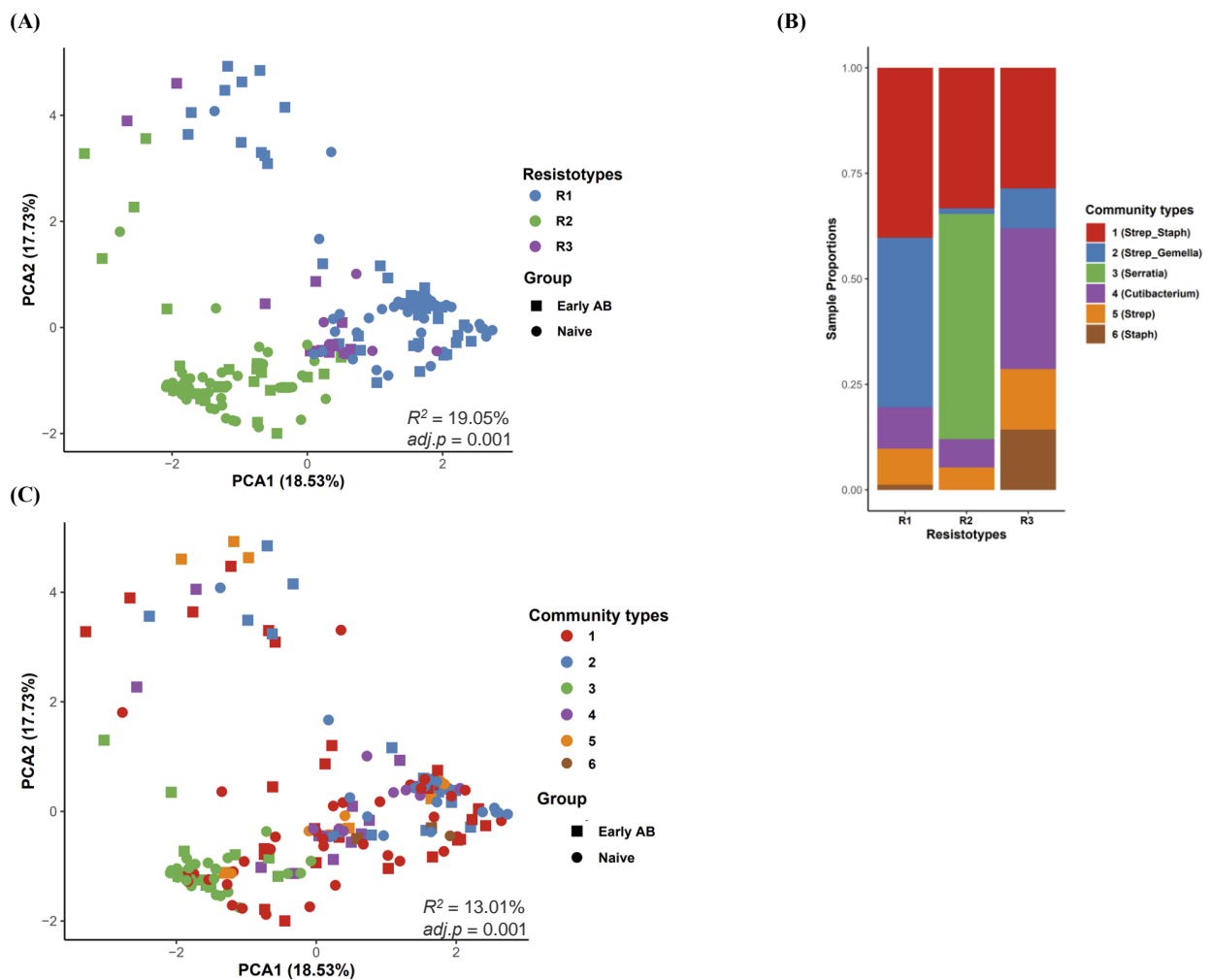


**Fig. 4:  $\beta$ -diversity PCA plots of overall nasopharyngeal resistome composition stratified per time point.** Principal component analysis (PCA) based on Euclidean distances of CLR-transformed ARG abundances (*i.e.*, Aitchison dissimilarity) between samples, visualizing the overall nasopharyngeal resistome composition stratified for antibiotic-treated infants and naive at time point (A) T1 (baseline) and (B) T2 (immediately after cessation of early antibiotic therapy). Each data point represents the resistome composition of one sample. The ellipses represent the 95% confidence interval around the centroid of resistome composition for each group. Statistical significance (adjusted p-values) and effect size ( $R^2$ ) of the differences in beta diversity were assessed using PERMANOVA. In panel (C), the results of the PERMANOVA-tests for the later time points are presented. Data points of early antibiotic-treated infants are colored red and those of naive infants are blue. PERMANOVA = permutational multivariate analysis of variance; p.adj = adjusted p-values.

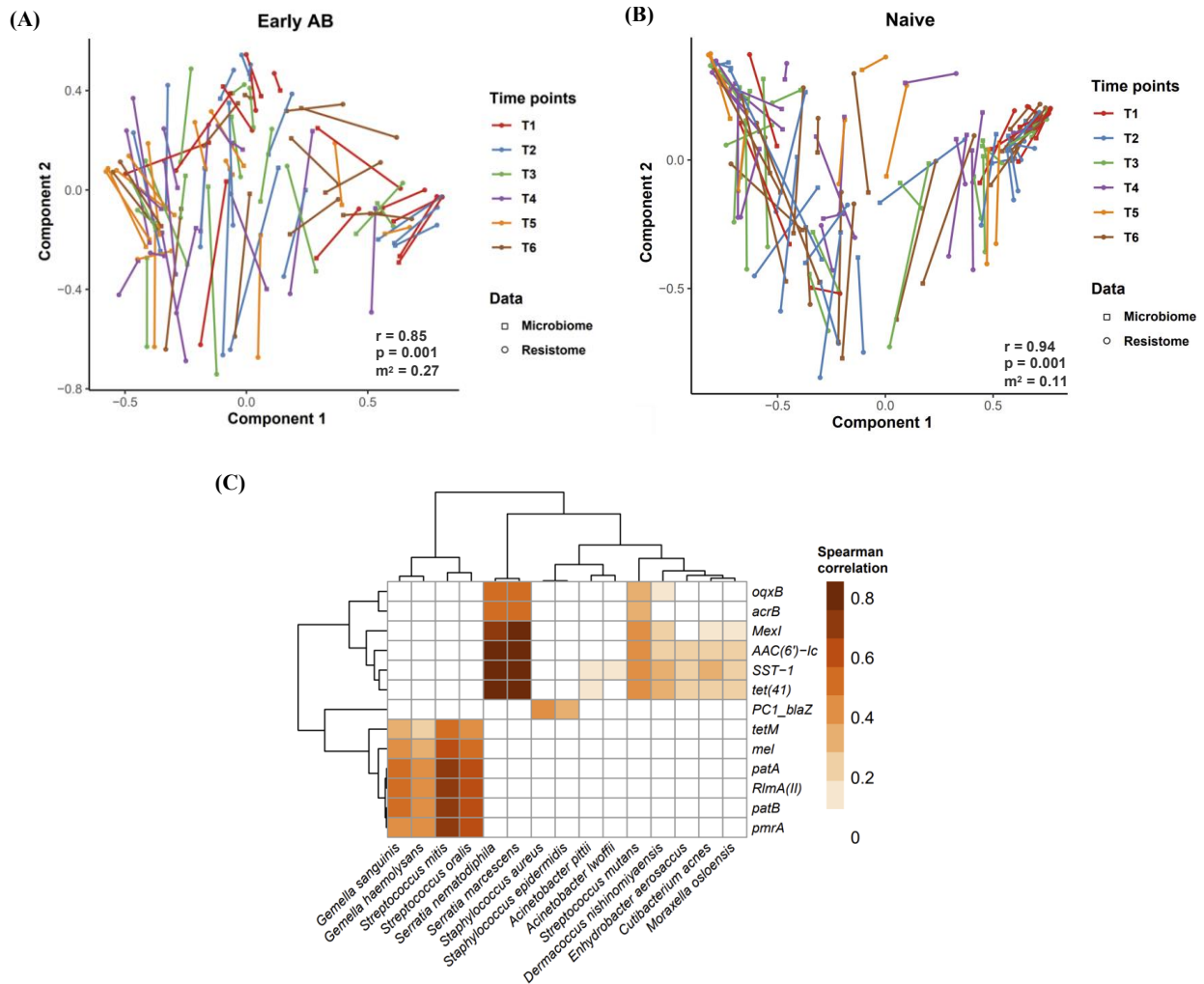


**Fig. 5: Factors influencing the overall nasopharyngeal resistome composition in the preterm infants.** PCA visualizations of Aitchison dissimilarities between all the nasopharyngeal samples from birth until 6 months corrected age at ARG level. Each data point corresponds to a sample, colored by (A) Infant, (B) PMA, (C) Time points and (D) Prenatal and post-natal antibiotic exposure and shaped according to early antibiotic exposure or naive groups. Effect sizes ( $R^2$ ) and corresponding adjusted p-values ( $p.adj$ , corrected using the Benjamini-Hochberg method) are calculated using PERMANOVA-tests (as shown in the bottom right of each plot). Percentage refers to the percentage of variance explained by the principal component. PMA = postmenstrual age; PC1 = principal component 1; PC2 = principal component 2; PERMANOVA = permutational multivariate analysis of variance;  $p.adj$  = adjusted p-values.

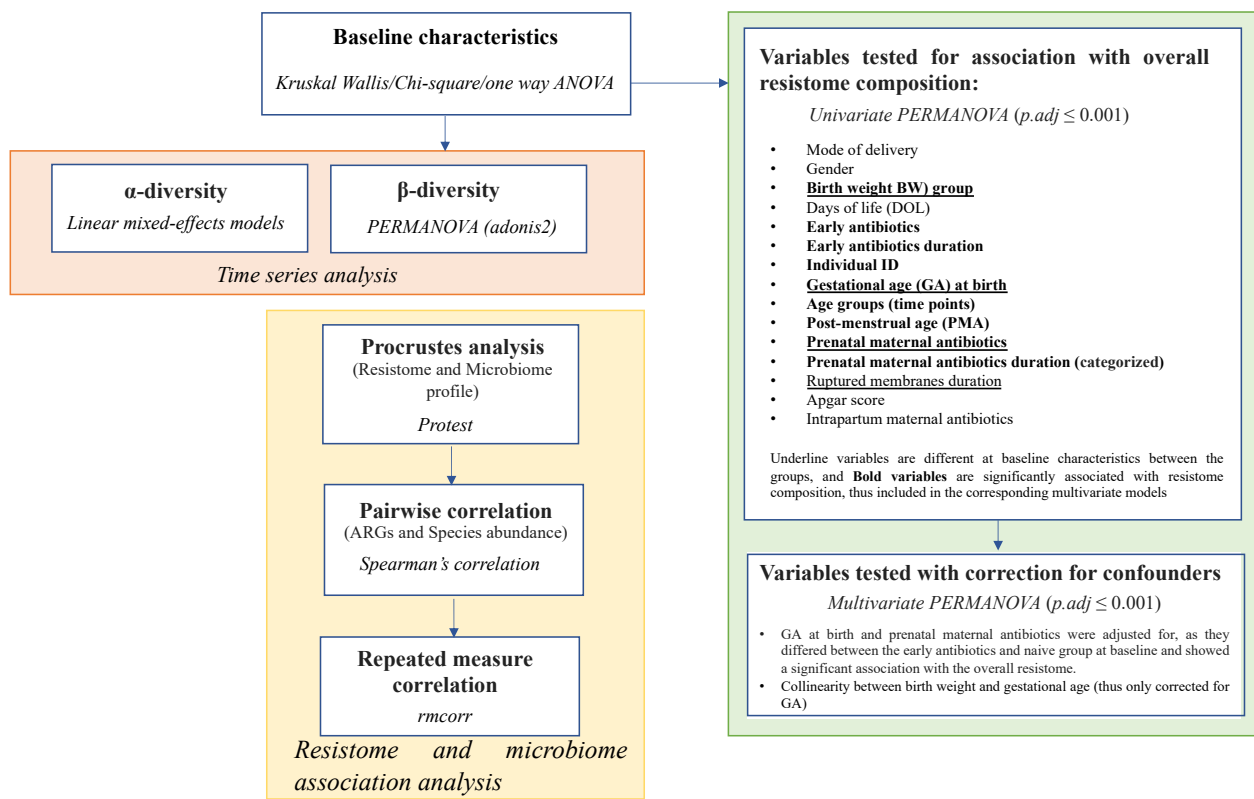




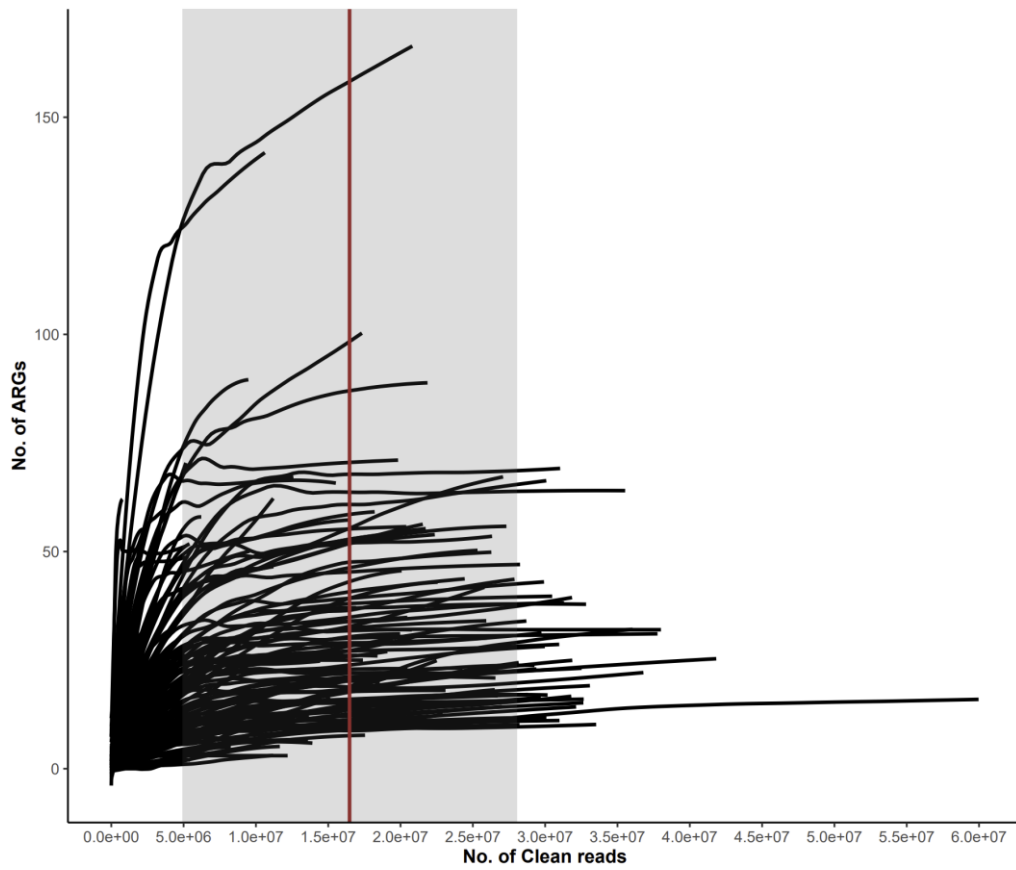
**Fig. 6: Association between resistotypes, community types, and ARG abundance profiles.** The PCA plot showing the association between overall ARG abundance profile with respect to their (A) resistotypes and (C) community types identified among all the nasopharyngeal samples using the DMM models.  $R^2$  was obtained using the PERMANOVA test. (B) Samples proportions for each resistotype shown as a function of community type. DMM = dirichlet multinomial mixture; Strep = *Streptococcus* ; Staph = *Staphylococcus*.



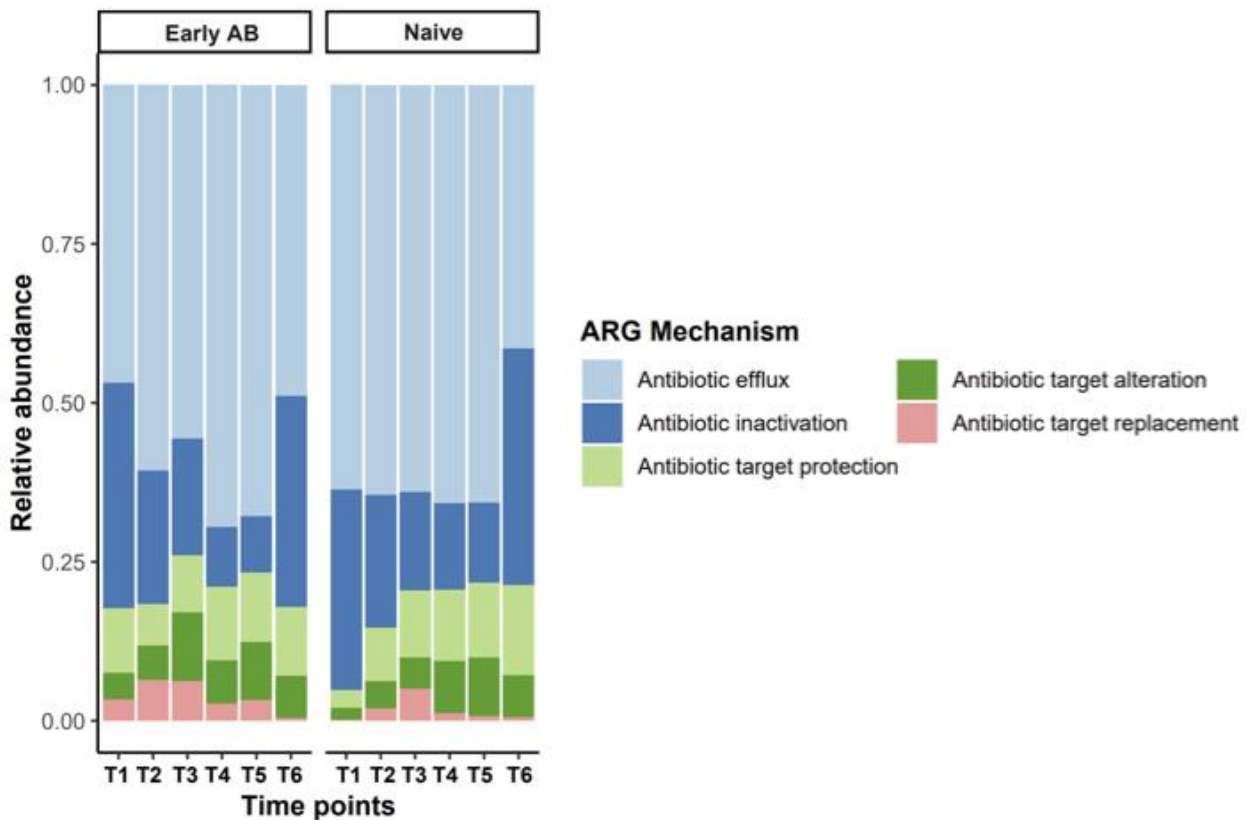
**Fig. 7: The resistome variation in nasopharynx is explained by microbiome variation.** The ARG and microbial abundance profiles were correlated with each other for antibiotic treated (A) and naive (B) groups using Procrustes analysis. The length of line connecting two points denotes the degree of dissimilarity or distance between the microbiome (filled square) and resistome (filled circle) composition of the same sample. The line and points are colored based on time points in both groups. The correlation coefficients and significance were calculated using the *protest* function in vegan package. (C) Heatmap depicts the Spearman's pairwise correlation coefficients ( $r$ ) between individual ARGs and microbial species from nasopharyngeal samples. Hierarchical clustering using Euclidean distance was performed on both rows and columns. Correlation coefficients ( $r$ ) was shown only where adjusted  $p$ -value  $< 0.05$ . ARGs = antibiotic resistance genes.



**Supplementary Fig. 1:** Statistical analysis design used in the study. ARGs = antibiotic resistance genes; PERMANOVA = permutational multivariate analysis of variance.



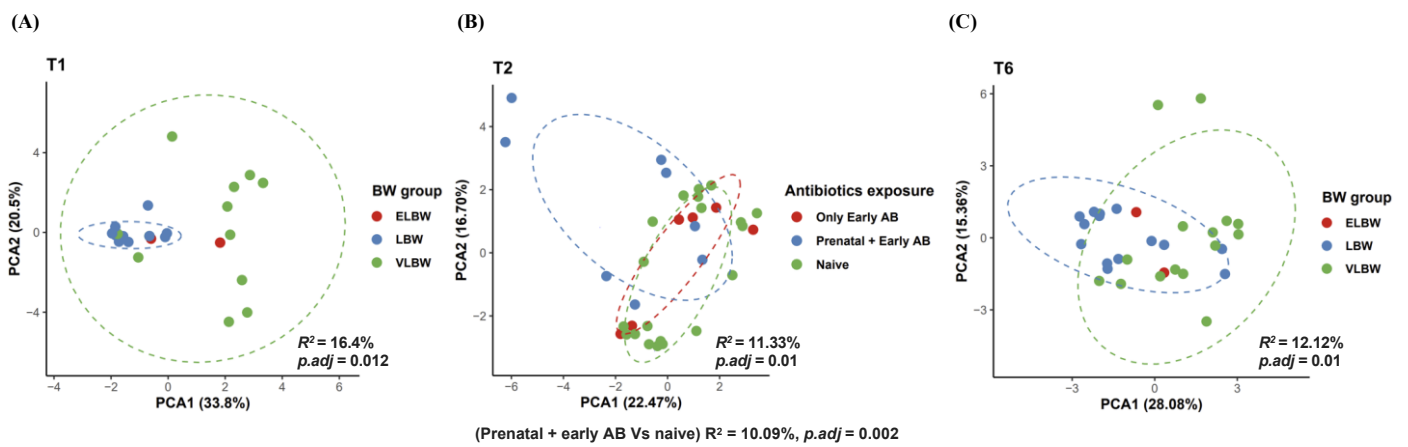
**Supplementary Fig. 2:** Rarefaction curve for the number of unique ARGs identified at subsampled sequencing depth. The red line indicates the median sequencing depth for all samples, while the gray shading represents one standard deviation.



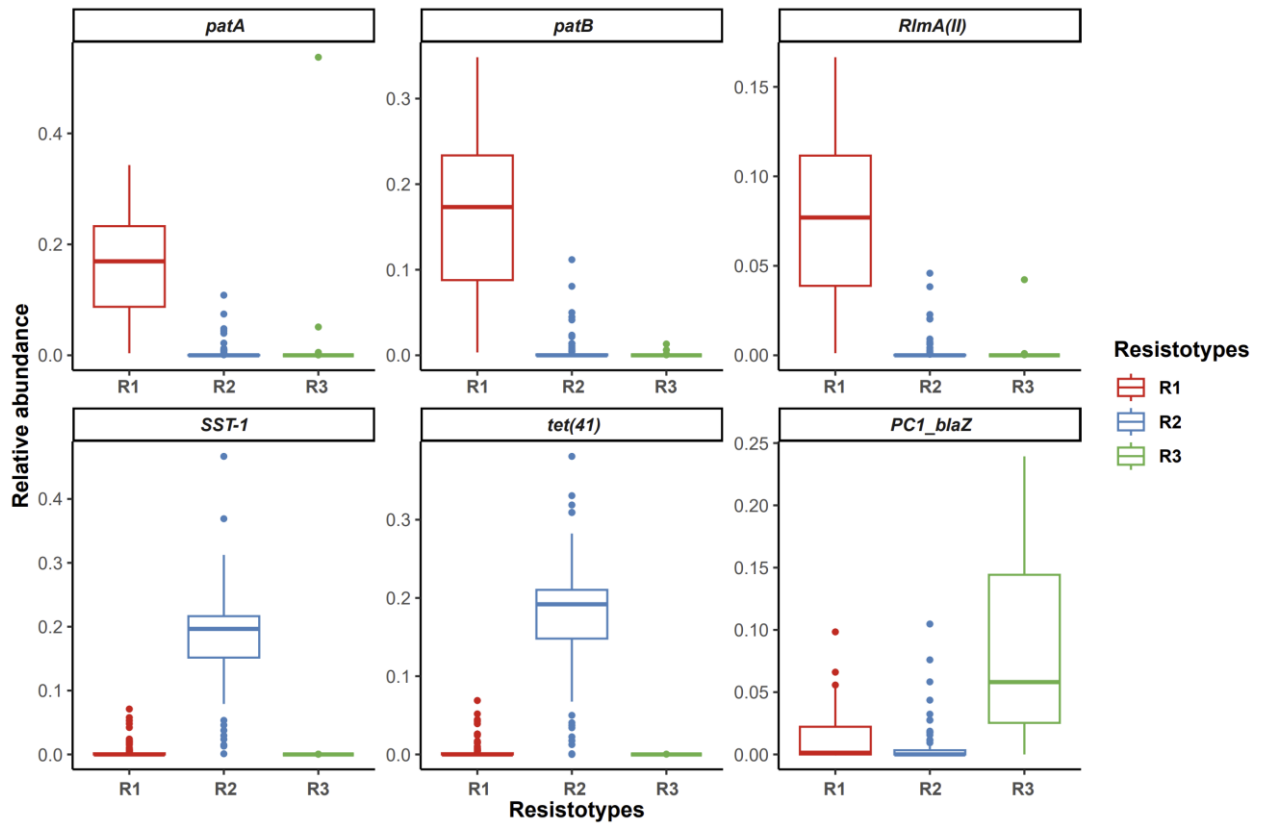
**Supplementary Fig. 3:** Mean normalized relative abundances of different ARG mechanisms found in the nasopharynx of early antibiotic treated and naive preterm infants at different time points (T1-T6).



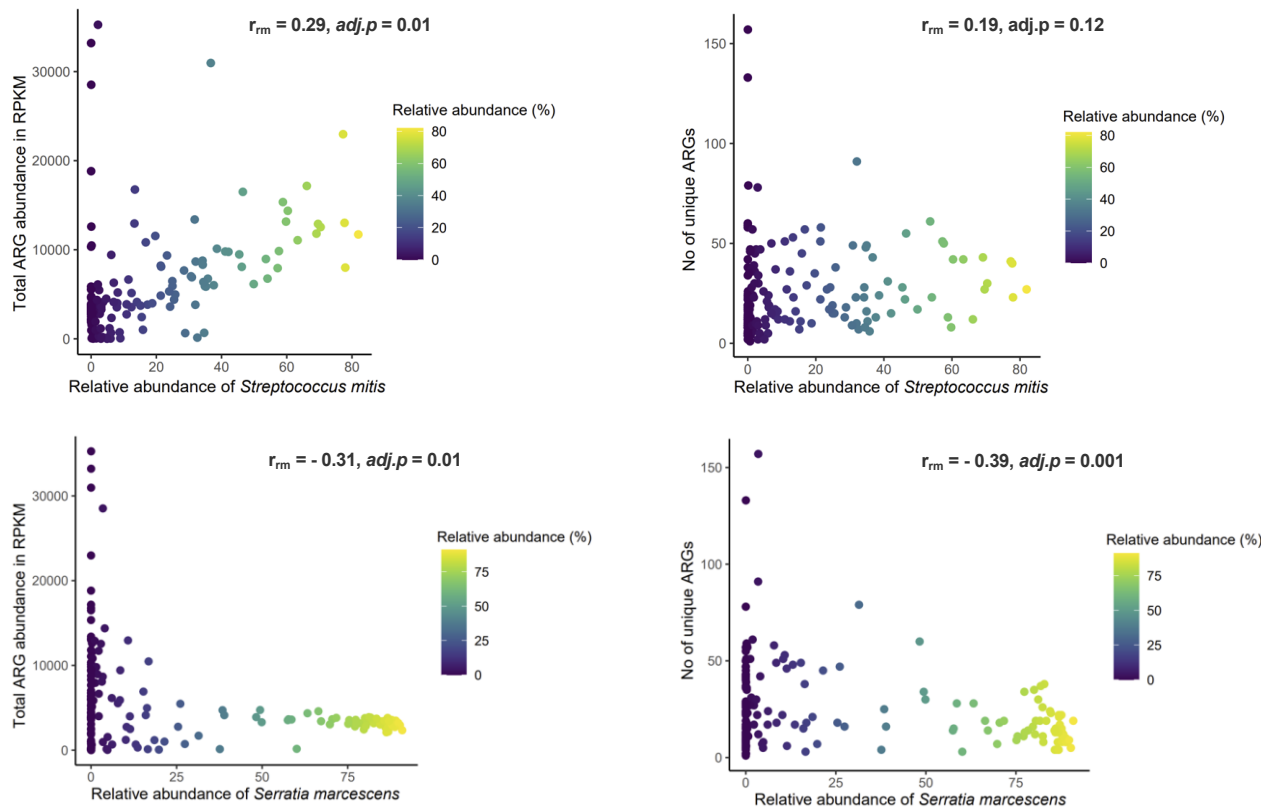
**Supplementary Fig. 4:** Heatmap showing the log<sub>10</sub>-transformed RPKM abundance of potential high-risk antibiotic resistance genes (ARGs) identified in the nasopharynx of all preterm infants across samples. The x-axis represents the names of samples, which are arranged by treatment group and time points. RPKM = reads per kilobase of reference gene per million bacterial reads.



**Supplementary Fig. 5:** PCA visualizations of beta diversity analysis using the Aitchison distances at time points: (A) T1, (B) T2 and (C) T6 (6 month corrected age). Each point corresponds to a sample colored based on birth weight group (A, C) and prenatal and postnatal antibiotic exposure (B). Ellipses indicate 95% confidence intervals (CI). Adjusted p-values and effect size ( $R^2$ ) are based on PERMANOVA test. CLR = centered-log ratio; BW= Birth Weight.



**Supplementary Fig. 6:** Box plots of the relative abundances of the main ARG contributors are shown for each Resistotype cluster identified across all the preterm infants' nasopharyngeal resistomes using DMM models. The X-axis displays the three identified clusters, and the Y-axis represents the relative abundance of the ARGs. DMM = dirichlet multinomial mixture.



**Supplementary Fig. 7:** Scatterplots showing the association between correlated species and overall resistome outcomes across all infant samples from birth to 6 months corrected age. The total ARG abundance in RPKM was positively correlated with the relative abundance of *Streptococcus mitis* (A) and non-linearly negatively associated with the relative abundance of *Serratia marcescens* (C). Plots (B) and (D) depict the relationship between the number of unique ARGs and the relative abundance of *Streptococcus mitis* and *Serratia marcescens*, respectively. Plot (A) and (C) are colored by the relative abundance of *Streptococcus mitis*, while plots (B) and (D) are colored by the relative abundance of *Serratia marcescens*. RPKM = reads per kilobase of reference gene per million bacterial reads.



**Table 1:** Baseline characteristics of preterm infants analyzed in this study

<b>NP cohort</b>	<b>Naive</b>	<b>Early-antibiotics</b>	<b>p</b>
<b>n</b>	21	15	
<b>Male, n (%)</b>	12 (57%)	11 (73%)	0.484
<b>GA, weeks (mean, SD)</b>	31, 6/7	29 5/7, 1	<b>&lt;0.001</b>
<b>BW, grams (mean, SD)</b>	1562, 188	1306, 277	<b>&lt;0.05 (0.0022)</b>
<b>BW, group, n (%)</b>			<b>&lt;0.05 (0.013)</b>
ELBW	0 (0%)	2 (13%)	
VLBW	8 (38%)	10 (67%)	
LBW	13 (62%)	3 (20%)	
<b>Mode of delivery, n (%)</b>			0.694
Vaginal	4 (19%)	4 (27%)	
C-section	17 (81%)	11 (73%)	
<b>Antibiotics during pregnancy, n (%)</b>			<b>&lt;0.001</b>
None	21 (100%)	6 (40%)	
< 10 days		0 7 (47%)	
>= 10 days		0 2 (13%)	
<b>Intrapartum antibiotics (IAP), n (%)</b>			0.50
None documented	2 (10%)	0 (0%)	
Given	19 (90%)	15 (100%)	
<b>ROM, hours (means, SD)</b>	3, 12	432, 733	<b>&lt;0.05 (0.0108)</b>
<b>Apgar10 (mean, SD)</b>	9, 1	8, 1	0.3039
<b>Nutrition (discharge to 6 months CA)</b>			<b>&lt;0.05 (0.039)</b>
fully breastfed	7 (33%)	4 (29%)	
at least 50% mother milk, but also some formula	3 (14%)	6 (42%)	
mostly formula and some mothers milk	6 (29%)	0 (0%)	
formula	2 (10%)	4 (29%)	
<b>Lost to follow up</b>	3 (14%)	1 (7%)	0.626

**Supplementary Table 1.** Sample flowchart. Flowchart showing number of nasopharyngeal aspirate samples from preterm infants ( $n=36$ ) available at several steps in the study including at sampling, DNA extraction, sequencing and bioinformatics analysis.

<b>Time points</b>	<b>T1</b>	<b>T2</b>	<b>T3</b>	<b>T4</b>	<b>T5</b>	<b>T6</b>	<b>Total</b>
Days of life (DOL)	0-3	4-9	12-19	20-31	32-76	229-288	
Extracted samples (before library prep)	35	37	36	35	23	32	<b>198</b>
Exclude samples (Low DNA yield)	2	1	1	0	0	0	<b>4</b>
Exclude samples (Low DNA yield) (%)	5.7%	2.7%	2.8%	0%	0%	0%	<b>2%</b>
Sequenced samples	33	36	35	35	22	32	<b>194</b>
Samples failed during Resistome profiling	5	1	2	0	1	1	<b>10</b>
Samples with double sampling time points	0	1	0	1	1	0	<b>3</b>
Included in the final analysis	28	34	33	34	21	31	<b>181</b>

**Supplementary Table 2: Sequence reads statistics**

<b>Sample</b>	<b>Raw reads</b>	<b>Reads after Trimming</b>	<b>Human reads</b>	<b>Remaining (Clean) reads</b>	<b>Reads aligning to CARD database</b>	<b>% of clean reads aligned to CARD</b>
S_103	26080501	26067155	16513464	9553691	17435	0.1825
S_104	22764460	22752830	20748076	2004754	10262	0.5119
S_105	20970229	20964496	3419609	17544887	506	0.0029
S_106	28332496	28321717	15003675	13318042	6043	0.0454
S_119	27475272	27457851	25839804	1618047	105	0.0065
S_120	22266609	22253052	20695679	1557373	82	0.0053
S_121	26201196	26188557	22868983	3319574	11179	0.3368
S_122	33790528	33772729	28966515	4806214	6608	0.1375
S_123	28686631	28676796	1679787	26997009	22095	0.0818
S_124	36678862	36580731	32369824	4210907	330	0.0078
S_125	37839541	37794781	27178195	10616586	302677	2.851
S_127	27688873	27682506	593223	27089283	181147	0.6687
S_128	31005030	30983206	27167769	3815437	15888	0.4164
S_129	28691309	28680175	17478848	11201327	17478	0.156
S_130	33163819	33076882	19478139	13598743	401	0.0029
S_137	31372046	31329501	14342527	16986974	13164	0.0775
S_138	21667917	21662586	846609	20815977	504320	2.4228
S_140	24503529	24497786	2632209	21865577	26203	0.1198
S_141	32823359	32806877	26585210	6221667	27575	0.4432
S_142	26789831	26776365	22169024	4607341	3279	0.0712
S_154	33191448	33139128	27972056	5167072	39874	0.7717
S_156	26884193	26874230	546850	26327380	586046	2.226
S_157	37364064	37340342	32241091	5099251	30575	0.5996
S_158	21680884	21674017	651114	21022903	25474	0.1212
S_159	32047093	32031086	26619128	5411958	4165	0.077

S_160	37292025	37257965	16861312	20396653	12196	0.0598
S_161	45235114	45148815	41399286	3749529	1241	0.0331
S_162	67632600	67539141	62221411	5317730	28064	0.5277
S_163	38901430	38869031	36914900	1954131	22451	1.1489
S_164	44585351	44483261	8951337	35531924	288242	0.8112
S_165	25628648	25540450	10013320	15527130	61843	0.3983
S_166	25531087	25512936	21798951	3713985	3399	0.0915
S_167	23680492	23663333	6734120	16929213	116337	0.6872
S_168	45431318	45401429	35864464	9536965	59284	0.6216
S_169	55896867	55855030	52825510	3029520	9542	0.315
S_170	30624654	30570075	19822065	10748010	26921	0.2505
S_171	27481415	27469324	11777603	15691721	2894	0.0184
S_172	37856150	37836322	25830529	12005793	16131	0.1344
S_173	32620409	32598666	2520144	30078522	144584	0.4807
S_174	29843735	29836353	8305978	21530375	41500	0.1928
S_175	28853583	28843865	8770011	20073854	42924	0.2138
S_176	40254779	40230614	38052826	2177788	4703	0.216
S_177	35925057	35889012	19426456	16462556	18834	0.1144
S_178	20274695	20263845	4157655	16106190	27812	0.1727
S_179	28497251	28482477	685432	27797045	64980	0.2338
S_180	23519660	23515651	1775004	21740647	87325	0.4017
S_181	23879524	23867846	23099772	768074	4218	0.5492
S_182	27132318	27127749	3087081	24040668	45811	0.1906
S_183	31972863	31964880	2266406	29698474	50764	0.1709
S_184	31108627	31101186	1126947	29974239	57650	0.1923
S_185	31622571	31616517	3735586	27880931	60664	0.2176
S_186	46757985	46730502	44193952	2536550	4820	0.19
S_187	30402571	30396836	3688615	26708221	47750	0.1788
S_188	33765529	33756376	1114356	32642020	62286	0.1908
S_189	33083199	33073575	2904430	30169145	57677	0.1912

S_190	30060393	30040023	26669757	3370266	18826	0.5586
S_191	33250280	33222469	16861645	16360824	101214	0.6186
S_193	47659319	47614810	14786449	32828361	56496	0.1721
S_214	52537220	52508963	49589032	2919931	5637	0.1931
S_215	29571723	29543447	8765633	20777814	47796	0.23
S_216	35053190	35033550	6769117	28264433	76971	0.2723
S_217	39517504	39487561	38163679	1323882	625	0.0472
S_218	30675257	30665911	1270181	29395730	56564	0.1924
S_219	27547953	27529357	972788	26556569	111748	0.4208
S_220	33157900	33149181	484150	32665031	66541	0.2037
S_221	21346517	21337082	831108	20505974	31847	0.1553
S_222	26286828	26273899	25340090	933809	919	0.0984
S_223	33944591	33939850	414140	33525710	66865	0.1994
S_224	40663366	40642802	38200609	2442193	2399	0.0982
S_225	31198454	31171888	11201529	19970359	96199	0.4817
S_226	28486798	28471399	21921787	6549612	11262	0.1719
S_227	23681140	23676886	758241	22918645	45401	0.1981
S_228	22840259	22827956	12959279	9868677	18766	0.1902
S_229	32833037	32813316	29804790	3008526	4748	0.1578
S_230	32797647	32777151	27045604	5731547	8726	0.1522
S_234	23368571	23345781	15243726	8102055	2282	0.0282
S_235	29631358	29616146	9188612	20427534	924	0.0045
S_236	24316377	24307110	10727801	13579309	5898	0.0434
S_237	31936082	31906503	3105384	28801119	81934	0.2845
S_238	27183251	27160033	9761467	17398566	61606	0.3541
S_241	21273396	21265742	2856634	18409108	35437	0.1925
S_242	37922507	37895061	34772369	3122692	17337	0.5552
S_243	49645886	49619252	46731003	2888249	14157	0.4902
S_244	66190561	66123529	63015475	3108054	9000	0.2896
S_245	24627154	24616003	12421581	12194422	31173	0.2556

S_247	42327108	42314690	484772	41829918	230967	0.5522
S_248	47197213	47176488	26244459	20932029	30315	0.1448
S_249	51501346	51474520	13469371	38005149	93014	0.2447
S_250	32355912	32328838	13242506	19086332	8192	0.0429
S_251	27317957	27289175	17759907	9529268	914	0.0096
S_252	61163417	61088642	47722474	13366168	36220	0.271
S_253	45402778	45377662	39162898	6214764	9007	0.1449
S_254	32164676	32143130	6335296	25807834	88170	0.3416
S_255	39009857	38993377	25088212	13905165	25987	0.1869
S_256	44877026	44838912	37951047	6887865	9660	0.1402
S_257	34093993	34087358	3603064	30484294	64242	0.2107
S_258	21659477	21650879	1178683	20472196	111693	0.5456
S_259	33243536	33235008	1377165	31857843	84420	0.265
S_26	30270838	30256072	26479542	3776530	14701	0.3893
S_260	35892679	35854245	17290178	18564067	36661	0.1975
S_266	33878209	33871694	778911	33092783	52270	0.1579
S_267	29882106	29867961	3938949	25929012	127336	0.4911
S_268	32997828	32992621	2002198	30990423	54318	0.1753
S_269	48795323	48743359	18996479	29746880	175072	0.5885
S_27	49542796	49483919	35050025	14433894	172882	1.1978
S_270	30056925	30048511	785470	29263041	22344	0.0764
S_271	51738432	51716699	48204756	3511943	5587	0.1591
S_278	29095206	29082097	17430590	11651507	147	0.0013
S_279	29498360	29486657	2167755	27318902	214738	0.786
S_28	37996875	37978461	36894669	1083792	1639	0.1512
S_280	39383489	39340791	14014208	25326583	176136	0.6955
S_281	20907270	20894105	2682760	18211345	106040	0.5823
S_282	31922829	31901011	7462386	24438625	93232	0.3815
S_283	37846534	37832711	1040092	36792619	139836	0.3801
S_284	29863131	29846583	24399378	5447205	67	0.0012

S_285	34228242	34214505	14192796	20021709	73436	0.3668
S_286	41765816	41726214	30517072	11209142	66965	0.5974
S_287	32011138	32000174	2080267	29919907	149226	0.4988
S_288	34869195	34833718	19684842	15148876	58922	0.389
S_289	39559968	39542758	1775686	37767072	105574	0.2795
S_29	38045611	38021180	34100541	3920639	5300	0.1352
S_290	37264718	37245276	26298029	10947247	18539	0.1693
S_291	32072825	32065914	1094390	30971524	56532	0.1825
S_292	63816388	63807235	3838178	59969057	109957	0.1834
S_293	29500665	29493003	5380549	24112454	41039	0.1702
S_294	33573848	33566767	1050194	32516573	68572	0.2109
S_295	29125110	29118244	889487	28228757	46684	0.1654
S_296	34647449	34641389	2501224	32140165	55314	0.1721
S_297	27768840	27758959	1480206	26278753	145877	0.5551
S_298	33622191	33613344	2308277	31305067	72539	0.2317
S_299	22661274	22654821	287602	22367219	66621	0.2979
S_30	28267450	28256137	5690538	22565599	6201	0.0275
S_300	20301912	20296120	296296	19999824	37094	0.1855
S_301	24829638	24825480	540037	24285443	50436	0.2077
S_302	31874316	31867866	833258	31034608	76222	0.2456
S_303	39962863	39916697	18377375	21539322	92682	0.4303
S_304	26868511	26861459	297886	26563573	62060	0.2336
S_305	32243814	32235699	429181	31806518	71735	0.2255
S_306	28410746	28404637	5017057	23387580	48409	0.207
S_307	27609393	27603414	1085057	26518357	56119	0.2116
S_308	32232733	32227923	351812	31876111	63333	0.1987
S_309	26565516	26560390	766174	25794216	57990	0.2248
S_310	36884546	36844010	16364759	20479251	43776	0.2138
S_311	24334581	24314320	1204496	23109824	50555	0.2188
S_312	31069699	31062500	976160	30086340	1325	0.0044

S_313	36354561	36333646	28072594	8261052	178	0.0022
S_314	39811266	39782498	37746037	2036461	205	0.0101
S_315	54249580	54222485	51664787	2557698	7333	0.2867
S_316	44061982	44039723	40399113	3640610	7186	0.1974
S_317	44116478	44095092	42622292	1472800	633	0.043
S_331	29644028	29624401	13828912	15795489	31192	0.1975
S_332	39672217	39644362	38196480	1447882	827	0.0571
S_333	26467120	26441980	13851310	12590670	97356	0.7732
S_334	28679182	28648634	11319308	17329326	105763	0.6103
S_335	29706172	29701470	1533135	28168335	46960	0.1667
S_342	4190563	4185620	997864	3187756	873	0.0274
S_343	26076305	26063824	20907404	5156420	4327	0.0839
S_344	28905903	28893140	19415003	9478137	8856	0.0934
S_345	36832919	36821990	776032	36045958	136753	0.3794
S_346	61044890	61004312	59371452	1632860	2000	0.1225
S_347	69251411	68623791	64379680	4244111	1229	0.029
S_354	27433235	27422512	16225329	11197183	53	0.0005
S_356	27229735	27221424	4719264	22502160	86767	0.3856
S_357	27546962	27535322	18225311	9310011	27877	0.2994
S_364	64520057	64429899	62001636	2428263	979	0.0403
S_365	24055590	24045067	10406711	13638356	24613	0.1805
S_366	55484020	55321634	52415053	2906581	4562	0.157
S_367	28470266	28444380	8614772	19829608	96588	0.4871
S_368	20964782	20957580	7515333	13442247	25918	0.1928
S_369	31720586	31709062	3007642	28701420	42042	0.1465
S_53	34434233	34414250	32019563	2394687	7072	0.2953
S_55	26470453	26463326	7666968	18796358	2006	0.0107
S_86	28423053	28408167	23773990	4634177	13285	0.2867
S_87	66938909	66905387	64680249	2225138	6394	0.2874
S_88	30976709	30958246	4982552	25975694	72811	0.2803



S_89	25774477	25760044	24970857	789187	46	0.0058
S_90	46594816	46555570	45060476	1495094	395	0.0264
S_91	36674226	36641370	35795529	845841	1172	0.1386
S_93	53178273	53147304	51682818	1464486	2822	0.1927
S_95	33533344	33517467	32593464	924003	1464	0.1584
S_96	26810898	26796813	22625616	4171197	492	0.0118

**Supplementary Table 3: PERMANOVA results**

<b>Univariate analysis</b>					
<b>Factor</b>	<b>Df</b>	<b>SumOfSqs</b>	<b>R2</b>	<b>F</b>	<b>p.adj</b>
Early antibiotics	1	3574	0.02283	4.1813	0.001
Individual ID	35	59041	0.37705	2.5075	0.001
Time points (age groups)	5	12832	0.08195	3.1241	0.001
Mode of Delivery	1	600	0.00383	0.6884	0.756
Gender	1	1856	0.01185	2.1468	0.02
Birth weight (BW) group	2	6393	0.04083	3.7885	0.001
Days of life (DOL)	1	2466	0.01575	2.8645	0.005
Post-menstrual age (PMA)	10	15136	0.09666	1.8191	0.001
Gestational age (GA) at birth	3	6313	0.04032	2.4787	0.001
Prenatal maternal antibiotics duration (categorized)	2	8162	0.05212	4.8942	0.001
Prenatal maternal antibiotics	1	7008	0.04475	8.386	0.001
Early antibiotics duration	1	4640	0.02963	5.4662	0.001
Ruptured membranes duration	1	1692	0.0108	1.9551	0.046
Apgar score	1	1412	0.00902	1.6285	0.082
Intrapartum maternal antibiotics	1	2311	0.01476	2.6813	0.007

<b>Multivariate analysis</b>					
<b>Factor</b>	<b>Df</b>	<b>SumOfSqs</b>	<b>R2</b>	<b>F</b>	<b>p.adj</b>
Individual ID	31	47028	0.30033	2.255	0.001
Early antibiotics	1	947	0.00605	1.1577	0.31
Early antibiotics duration	1	2009	0.01283	2.4664	0.009
Time points (age groups)	5	13213	0.08438	3.4531	0.001
Post-menstrual age (PMA)	10	15261	0.09746	1.959	0.001
Prenatal maternal antibiotics	1	5700	0.0364	6.9391	0.001
Prenatal maternal antibiotics duration (categorized)	2	6961	0.04445	4.25	0.001





# ResistoXplorer: a web-based tool for visual, statistical and exploratory data analysis of resistome data

Achal Dhariwal<sup>1</sup>, Roger Junges<sup>1</sup>, Tsute Chen<sup>2,3</sup> and Fernanda C. Petersen<sup>1,\*</sup>

<sup>1</sup>Institute of Oral Biology, Faculty of Dentistry, University of Oslo, 0372 Oslo, Norway, <sup>2</sup>Department of Microbiology, The Forsyth Institute, 02142 Cambridge, MA, USA and <sup>3</sup>Department of Oral Medicine, Infection, and Immunology, Harvard School of Dental Medicine, 02115 Boston, MA, USA

Received August 31, 2020; Revised December 25, 2020; Editorial Decision February 24, 2021; Accepted March 01, 2021

## ABSTRACT

The study of resistomes using whole metagenomic sequencing enables high-throughput identification of resistance genes in complex microbial communities, such as the human microbiome. Over recent years, sophisticated and diverse pipelines have been established to facilitate raw data processing and annotation. Despite the progress, there are no easy-to-use tools for comprehensive visual, statistical and functional analysis of resistome data. Thus, exploration of the resulting large complex datasets remains a key bottleneck requiring robust computational resources and technical expertise, which creates a significant hurdle for advancements in the field. Here, we introduce ResistoXplorer, a user-friendly tool that integrates recent advancements in statistics and visualization, coupled with extensive functional annotations and phenotype collection, to enable high-throughput analysis of common outputs generated from metagenomic resistome studies. ResistoXplorer contains three modules—the ‘Antimicrobial Resistance Gene Table’ module offers various options for composition profiling, functional profiling and comparative analysis of resistome data; the ‘Integration’ module supports integrative exploratory analysis of resistome and microbiome abundance profiles derived from metagenomic samples; finally, the ‘Antimicrobial Resistance Gene List’ module enables users to intuitively explore the associations between antimicrobial resistance genes and the microbial hosts using network visual analytics to gain biological insights. ResistoXplorer is publicly available at <http://www.resistoxplorer.no>.

## INTRODUCTION

Antimicrobial resistance (AMR) represents a major threat to global public health and the economy (1). Consequently,

examining the emergence and dissemination of AMR genetic determinants is one of the priorities in global research (2–4). Until recently, genetic determinants were mostly understood in the context of specific pathogens; however, to fully understand how antimicrobial resistance genes (ARGs) emerge and disseminate, a more holistic approach is required. In this respect, advancements in short-read based high-throughput DNA sequencing (HTS) technologies and computation methods have facilitated rapid identification and characterization of ARGs across microbial genomes present in a sample (metagenome) (5,6). They have been shown to provide unprecedented knowledge into the large reservoir of ARGs and contributed to elucidate the ARG composition and the spread of AMR between human, animal and environmental microbial communities (7–12). Currently, resistome profiles describing ARGs in complex and diverse microbial metagenomes are primarily generated using whole metagenome shotgun sequencing in which the total DNA extracted from a microbial community is sequenced. The resulting DNA fragments can be analyzed using read or assembly based approaches to characterize their resistome composition (5,6). These derived sequencing datasets are both large and complex, causing considerable ‘big data’ challenges in downstream data analysis.

The main computational effort in resistome analysis of metagenomic datasets so far has focused on processing, classification, assembly and annotation of sequenced reads. This has led to the development of a number of excellent bioinformatic tools and databases for detecting and quantifying ARGs in metagenomes (5,6,13,14). However, there is still no clear consensus with regards to standard analysis pipelines and workflows for high-throughput analysis of AMR metagenomic resistome data (14,15). Nonetheless, the outputs from most of these pipelines can be summarized as a data table consisting of ARGs abundance information across samples, i.e. resistome profiles, along with their functional annotations and sample metadata. For most researchers, the fundamental challenge in data analysis can often be centered on how to understand and interpret the information in the abundance tables especially within the context of different experimental factors and annotations.

\*To whom correspondence should be addressed. Tel: +47 2284 0312; Fax: +47 2284 0312; Email: f.c.petersen@odont.uio.no

Downstream analysis of resistome data can be separated into four main categories: (i) composition profiling—to visualize and characterize the resistome based on approaches developed in community ecology, such as alpha diversity, rarefaction curves or ordination analysis; (ii) functional profiling—to analyze resistome profiles at different functional categories (e.g. drug class, mechanism), thus gaining better insights on their collective functional capabilities; (iii) comparative analysis—to identify features having a significant differential abundance between studied conditions; and (iv) integrative analysis—to integrate the resistome and taxonomic data to understand the complex interplay and potential associations between microbial ecology and AMR. The computational methods and approaches to perform such analysis are fairly diverse and require deep understanding and programming skills, representing significant barriers for their broader and exploratory applications (16). The first category of analysis can be more straightforward to perform, but the last three are challenging.

Fundamental challenges within the different categories relate to the fact that metagenomic data is often characterized by differences in library sizes, sparsity, over-dispersion and compositionality (17,18). Hence, it is critical to normalize the data to achieve comparable and meaningful results (18–20). To deal with uneven library sizes, researchers often employ two common normalization approaches prior to analysis: subsampling the reads in each sample to the same number (rarefying) or rescaling the total number of reads in each sample to a uniform sum (using proportions). The former may entail loss of valuable information, while the latter could lead to issues related to data compositionality (21). To overcome such challenges, sophisticated scaling methods based on log-ratio transformations have been proposed (22,23). To identify differentially abundant genes, the development of statistical models that account for features of metagenomic data or the use of methods to transform data to have distributions that fit standard test assumptions is generally recommended (19). For instance, the metagenomeSeq algorithm incorporates cumulative sum scaling (CSS) normalization and a zero-inflated Gaussian (ZIG) mixture model to reduce false positives and improve statistical power for differential abundance analysis (24,25). It has also been demonstrated that algorithms developed for RNA-seq data such as edgeR and DESeq2, along with their respective normalization methods, outperform other approaches used for metagenomic abundance data (20,25,26). These standard strategies are widely employed, but do not explicitly account for the compositional nature of whole metagenomic sequencing data (27,28). To address this issue, promising Compositional Data Analysis (CoDA) approaches have been proposed (29,30).

Nonetheless, there is no one statistical method suitable for all types of metagenomic datasets (20,26). The best results are achieved from a trade-off between data characteristics (sample or group size, sequencing depth, effect sizes, genes abundances, etc.) and the normalization method, incorporated with the coupled exploratory and comparative analysis (31). Therefore, various statistical and normalization methods are required for different metagenomic datasets, analyses and research questions addressed (6,19,25,26,31). However, most of the approaches have been

implemented as R packages. Although flexible, learning R in order to use these methods can be challenging for most clinicians and researchers.

In the second category, functional profiling, the annotated resistome abundance profiles are analyzed by mapping ARGs either to the respective class of drugs to which they confer resistance (Class-level) or to their underlying molecular mechanism of resistance (Mechanism-level). Analyzing resistomes at such high level categories enables researchers to gain more biological, actionable and functional insights together with a better understanding of their data. However, these functional levels and categories, along with their classification scheme, vary considerably between AMR reference databases (15). Additionally, depending upon the database used for annotation, users need to manually collect and curate such information and then generate separate abundance tables for each functional level. Hence, collecting appropriate functional annotation information for hundreds of ARGs in resistomes for functional profiling and further downstream analysis can be confusing, arduous, time-consuming and error-prone. Some of these databases may also provide information regarding the microbial hosts that harbor or carry these reference ARGs. Information about such relationships can be complex as one microbe can carry multiple ARGs and single ARGs can in turn be present across multiple microbes. To explore such intricate ‘multiple-to-multiple’ relations, one option is to use a network-based visualization method. However, such visual exploration support is not present in current resistome analysis tools.

To address these gaps as well as to meet recent advances in resistome data analysis, we have developed ResistoXplorer, a user-friendly, web-based, visual analytics tool, to assist clinicians, bench researchers and interdisciplinary groups working in the AMR field to perform exploratory data analysis on abundance profiles and resistome signatures generated from AMR metagenomics studies. The key features of ResistoXplorer include:

1. Support of a wide array of common as well as advanced methods for composition profiling, visualization and exploratory data analysis;
2. Comprehensive support for various data normalization methods coupled with standard as well as more recent statistical and machine learning algorithms;
3. Support of a variety of methods for performing vertical data integrative analysis on paired datasets (i.e. taxonomic and resistome abundance profiles);
4. Comprehensive support for ARG functional annotations along with their microbe and phenotype associations based on data collected from >10 reference databases;
5. A powerful and fully featured network visualization for intuitive exploration of ARG-microbe associations, including functional annotation enrichment analysis support.

Collectively, these features consist of a comprehensive tool suite for exploratory downstream analysis of data generated from AMR metagenomics studies. ResistoXplorer is freely available at <http://www.resistoxplorer.no>.

## MATERIALS AND METHODS

### Design and implementation

ResistoXplorer is implemented based on Java, R and JavaScript programming languages. The framework is developed based on the Java Server Faces technology using the PrimeFaces (<https://www.primefaces.org/>) and BootsFaces (<https://www.bootsfaces.net>) component library. The network visualization uses the sigma.js (<http://sigmaj.js.org/>) JavaScript library. Additionally, D3.js (<https://d3js.org/>) and CanvasXpress (<https://canvasxpress.org/>) JavaScript libraries are utilized for other interactive visualization. All the R packages for performing back-end analysis and visualization are mentioned in the 'About' section of the tool. At the start of the analysis, a temporary account is created with an associated home folder to store the uploaded data and analysis results. All the analysis results will be returned in real-time. Upon completing their analysis session, users should download all their results. The system is deployed on a dedicated server with four physical CPU cores (Intel Core i5 3.4GHz), 8GB RAM and Ubuntu 18.04 LTS was used as the operation system. ResistoXplorer has been tested with major modern browsers such as Google Chrome, Mozilla Firefox, Safari and Microsoft Internet Explorer.

### Program description and methods

ResistoXplorer consists of three main analysis modules. The first is the 'ARG List' module that is designed to explore the functional and microbial hosts associations for a given list of ARGs of interest. The second is the 'ARG Table' module, which contains functions for analyzing resistome abundance profiles generated from AMR metagenomics studies. Lastly, the 'Integration' module enables users to perform integrative analysis on the paired taxonomic and resistome abundance profiles to further explore potential associations coupled with novel biological insights and hypotheses. Figure 1 represents the overall design and workflow of ResistoXplorer. We recommend users to try out our example datasets to get familiar with the basic steps and key features of the tool before proceeding with analysis of their own data. ResistoXplorer also contains manuals and a comprehensive list of frequently asked questions to assist users to easily navigate through different analysis tasks.

## RESULTS AND DISCUSSION

### Data upload and processing

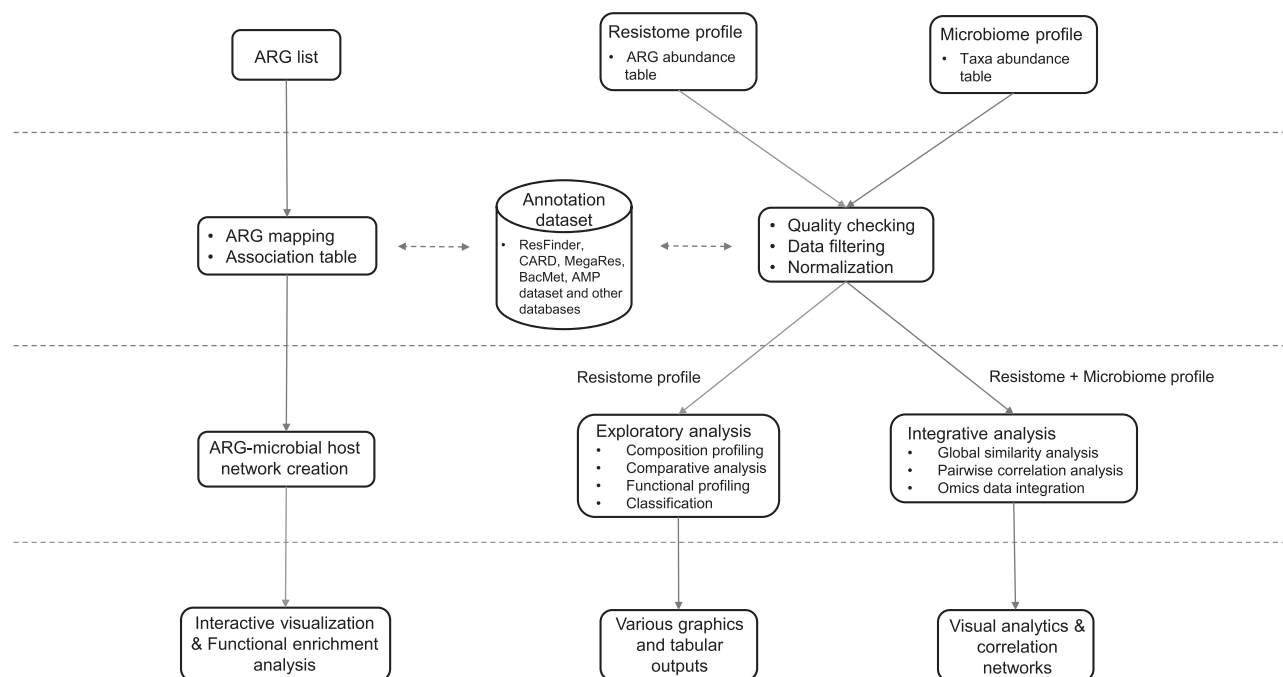
*Overview of data inputs.* The three analysis modules (ARG List, ARG Table and Integration) are represented as three interactive buttons in ResistoXplorer. Users must choose an analysis path based on their input. The input can be uploaded in two different ways—by entering a list of ARGs or by uploading an ARG abundance table along with a sample metadata file containing group information. In the latter case, the files can be uploaded as a tab-delimited text (.txt) or in comma-separated values (.csv). Further, users must also provide the annotation information of ARGs either by uploading a file (.txt or .csv) or by just selecting the same

database used for their annotation during upstream analysis. Additionally, the Integration module requires users to also upload a taxa abundance table in the same formats. The taxonomic and functional annotation files are optional in case of integrative analysis. Users can go to the corresponding 'Manuals' and 'Data Format' section, or download the example datasets for more details.

*ARG-functional annotations collection.* The annotation information and the classification scheme for reference ARGs (or sequences) are collected from nine widely used generalized AMR databases: CARD (32), ResFinder (33), MEGARes (15,34), AMRFinder (35), SARG (36), DeepARG-DB (37), ARGminer (38), ARDB (39) and ARG-ANNOT (40). Further, this annotation information is organized into tables containing the reference ARGs in rows and their functional annotation levels across columns for each of the databases. The headers (names) for reference ARGs are annotated as in the chosen database. Some of the annotation levels having multiple functional category assignments for ARGs were removed from the tables to avoid false counts inflation during downstream analysis. Additional manual curation for functional category naming and data redundancies were performed in some of the databases. Further, we also structured functional annotation information from the BacMet (41) database and antimicrobial peptide (AMP) resistance gene dataset (42) to enable users to perform functional profiling and downstream analysis of antibacterial biocides/metals and AMP resistance genes abundance profiles. It should be noted that ResistoXplorer does not perform any functional annotation of sequencing data. The tables are stored in RDS file format for less storage space and fast retrieval of data. Users can use these tables to analyze their resistome profile directly at different functional levels rather than manually collecting the annotation information from their corresponding database. The 'Data Format' and 'About' pages provide a detailed description on the format, structure of annotation table and database statistics, together with the links to allow users to download the annotation structure ('Downloads' section) available for the individual database.

*Data filtration and normalization.* By default, features with zero read count across all the samples or only present in one sample with extremely low count (e.g. 2) are removed from downstream analysis based upon statistical and biological approximations. Also, features present in only a small percentage of samples (e.g. 20%) with very few counts (e.g. 2) cannot be discriminated from sequencing errors or low-level contamination. It is also considered difficult to interpret their significance with respect to the whole community. By default, such low-quality features are filtered based on sample prevalence and their abundance levels to improve the comparative analysis. The default values are those used by the other tools and mostly found in the literature (43,44). Users can also choose to remove low abundant features by setting a minimum count cutoff based on their mean or median value. Conversely, some features remain constant in their abundances throughout the experimental conditions or across all the samples. These features are implausible to be informative in the comparative analy-





**Figure 1.** ResistoXplorer flow chart. ResistoXplorer accepts resistance gene list and ARG/taxa abundance tables as input data. Three successive steps are performed: data processing, data analysis and result exploration. The accompanying web interface offers a varied suite of options, and generates several tables and graphics to enable users to intuitively go over the data analysis tasks.

sis. Users can exclude such low variant features based on their inter-quantile ranges, standard deviations or coefficient of variations (43). Removing those uninformative features can increase the statistical power by reducing multiple testing issues during differential analysis (45). The filtered data is used for most of the downstream analysis except alpha diversity and rarefaction analysis. In case of integrative analysis, users can also choose to apply different data filtration criteria for both taxonomic and resistome abundance data.

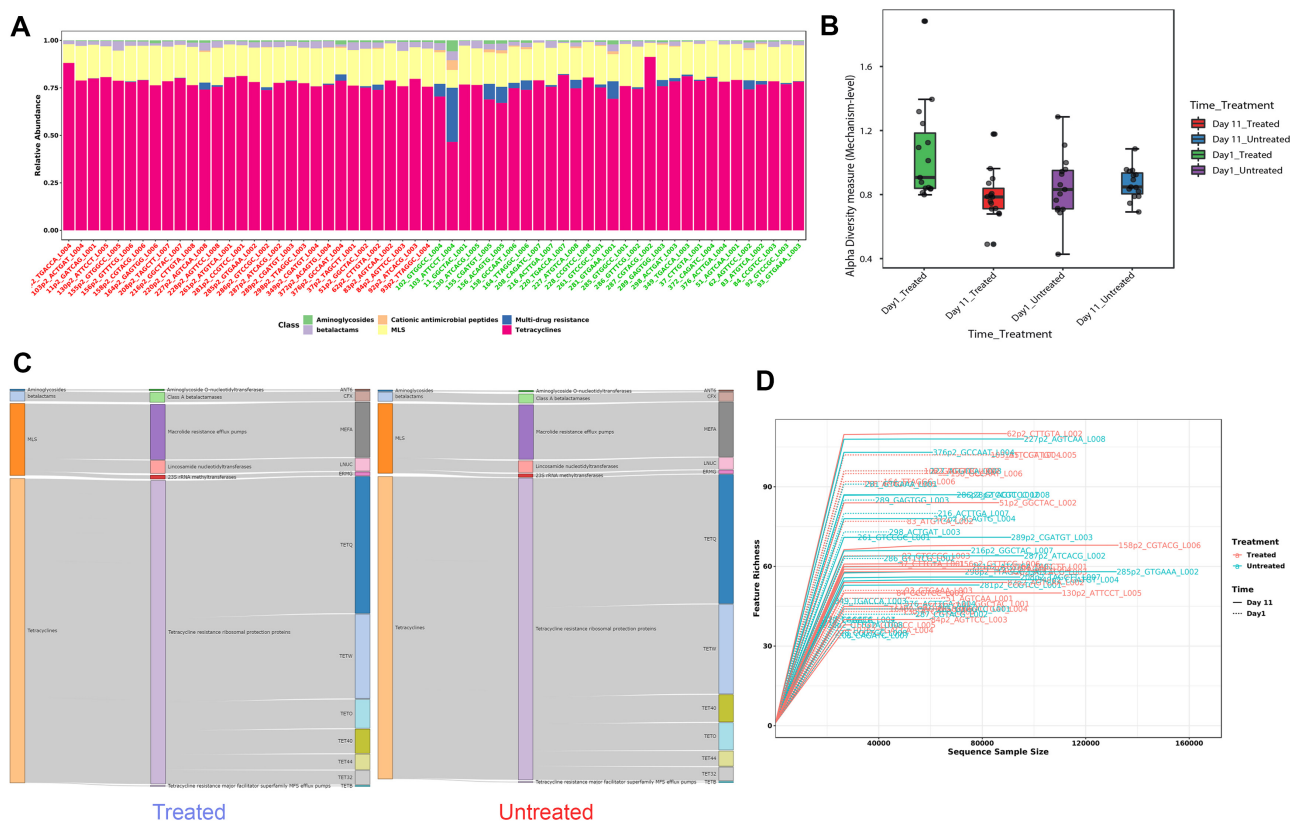
After data filtering, users must perform data normalization to remove the systematic variability between samples. Currently, ResistoXplorer offers three categories of data normalization—rarefying, scaling and transformation, based on various widely used methods for metagenomic abundance data (25). In addition, ResistoXplorer supports other normalization methods like centered log-ratio (clr) and additive log-ratio (alr) transformation to facilitate compositional data analysis. The choice of method is dependent upon the type of analyses to be performed (20,31). The normalized data is used for exploratory data analysis including ordination, clustering and integrative analysis. Users can explore different approaches and visually investigate the clustering patterns (*i.e.*, ordination plots, dendrogram and heatmap) to determine the effects of different normalization methods with regard to the experimental factor of interest. The total sum scaling (TSS) normalization is recommended for such type of analysis and has been set as the default option in ResistoXplorer (20,46,47). Also, comparative analysis using different approaches is performed on normalized data. However, each of these approaches will use its own specific normalization procedure

due to the lack of benchmark study evaluating which normalization methods should be combined with the different statistical approaches to achieve best performance for identifying differentially abundant genes (26). For example, the relative log expression normalization is used for DESeq2, and the centered log-ratio transformation is applied for ALDEx2. In the integrative analysis module, taxonomic and resistome datasets are normalized using the same approach.

### Composition profiling

**Visual exploration.** ResistoXplorer allows users to visually explore the resistome based on various intuitive visualization approaches used for metagenomic data. For instance, users can visualize resistome abundance data while simultaneously showing the functional hierarchical relationships and connectivity between features using an interactive sankey diagram, zoomable sunburst or treemap graphics (Figure 2C). In case of treemap, users can click a particular rectangular block of interest to further inspect its compositions at a lower functional level. The abundances can be represented as either absolute counts or relative proportions. However, such visualizations are more suitable for resistome profiles having an acyclic and hierarchical functional annotation structure. Abundance profiles can also be viewed at different functional levels using other common visualizations such as stacked area or stacked bar plot (Figure 2A). The plot is organized by experimental factors to help visualize the differences in resistome composition across different conditions.





**Figure 2.** Example outputs from composition profiling panel of ARG Table module in ResistoXplorer. **(A)** A stacked bar chart showing class level resistance abundance profiles across samples. **(B)** A box plot summary of the Shannon diversity index at mechanism level in different treatment groups across sampling time points. **(C)** A Sankey diagram showing the resistome abundance profile of treated (left) and untreated (right) cattle group at hierarchical functional levels including class, mechanism and group. **(D)** A rarefaction curve showing the number of unique ARGs identified in each sample as a function of sequence sample size.

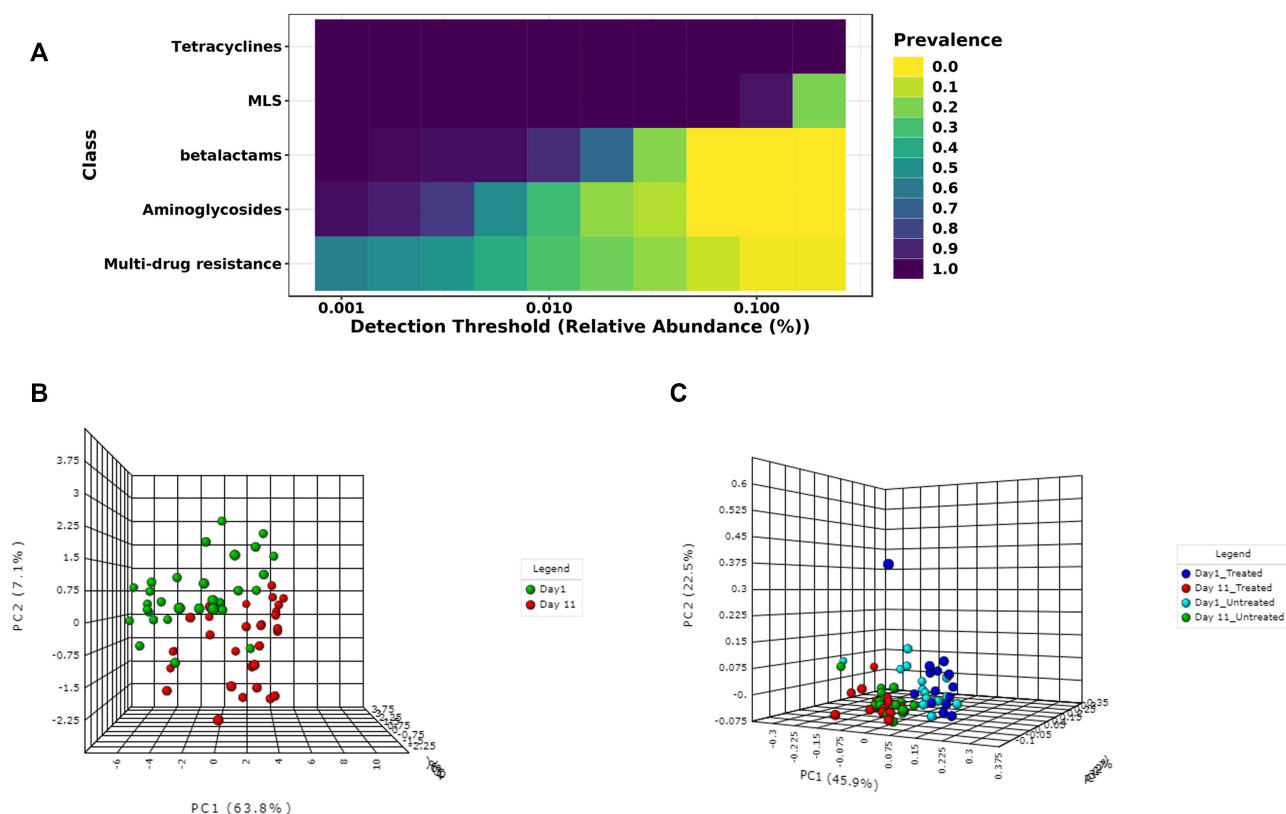
**Diversity profiling.** The resistome diversity profiling is implemented mainly based on the R *vegan* and *phyloseq* packages (48,49). Currently, users can perform alpha diversity (within-sample) analysis using eight common richness and/or evenness-based diversity measures. Since the Chao1 measure performs well and is recommended for estimating ARG diversity (50), it has been set as a default one in ResistoXplorer. The results are represented in the form of a dot plot for individual samples and box plots for each sample group (Figure 2B). The corresponding statistical significance is calculated automatically using either parametric or non-parametric tests. The analysis can be performed at different functional levels based on the available annotations. Additionally, the reliability of estimated diversity in samples can be assessed through rarefaction curves in which the number of unique features (ARGs) identified is plotted against the sequence sample size (Figure 2D).

**Ordination analysis.** The ordination analysis function allows users to explore and visualize the similarities or dissimilarities between samples based on their composition at different functional levels. The dissimilarity can be calculated using five non-phylogenetic-based quantitative or qualitative distance measures. Since the different types of distance measures have specific niches and can affect the

outcomes and the interpretation of the analysis, it is recommended by several authors to apply different measures to better understand the factors underlying composition differences (6,51,52). Currently, three widely accepted ordination methods are supported, including principal coordinate analysis (PCoA), non-metric multidimensional scaling (NMDS) and principal component analysis (PCA). In particular, users can follow a CoDA ordination approach by performing PCA on the centered log-ratio transformed data. The corresponding statistical significance is calculated using one of the three common multivariate statistical testing methods. Permutational multivariate analysis of variance (PERMANOVA) was set as the default option (53). By default, ordination analysis is performed using PCoA on a most widely used Bray–Curtis dissimilarity metric and assessed using PERMANOVA. The results are represented as both 2D and 3D sample plots. The samples visualized in these ordination plots are colored based on metadata or their alpha diversity measures to help users identify any underlying patterns in the datasets (Figure 3B).

**Comparative analysis**

**Differential abundance testing.** This section enables users to perform statistical testing to identify features that are



**Figure 3.** Illustration of core resistome and ordination analysis results in ResistoXplorer. (A) A heatmap showing the core resistome of cattle analyzed at class level. (B) A 3D PCA plot with sample colors based on time points. (C) A 3D PCoA plot with sample colors with regards to different treatment groups and time points.

significantly different in abundance across sample groups. ResistoXplorer supports standard tests, such as DESeq2 (22), edgeR (23), metagenomeSeq (24), LefSe (54), as well as more recent CoDA univariate analysis approaches such as ALDEx2 (55) and ANCOM (56). DESeq2 and edgeR are broadly used and generally considered as robust and powerful parametric statistical approaches for datasets with small group and equally distributed library sizes (18,20,25,26). They fit a generalized linear model and assume that read counts follow a negative binomial distribution to account for the features of count data. In contrast, the metagenomeSeq, with its recommended CSS normalization, has substantially higher performance with larger group sizes (26,48). LefSe uses the standard non-parametric tests for statistical significance coupled with linear discriminant analysis to assess the effect size of differentially abundant features. The CoDA methods perform statistical testing on the log ratios of features rather than their actual count abundances to deal with the compositional nature of sequencing data. ALDEx2, for instance, performs parametric or non-parametric statistical tests on log-ratio values from a modeled probability distribution of the data and returns the expected values of statistical tests along with effect size estimates. ANCOM tests the log-ratio abundance of all pairs of features for differences in means using non-parametric statistical tests. By default, the Benjamini–Hochberg correction is used for all approaches to control

the false discovery rate across significant genes. The differential analysis can also be performed at different functional levels.

The results from the differential analysis are displayed as a table. By default, the result table will show a maximum of 500 top features ordered according to their adjusted *P*-values. The significant features (if present) are automatically highlighted in orange color. Further, users can also see a boxplot summary for any feature of interest by just clicking the ‘Details’ icon. Since different statistical approaches may generate divergent *P*-values, it is often recommended to compare and visualize results using more than one statistical approach, as to increase confidence in the interpretation of results.

*Machine learning-based classification.* Prediction of microbiome signatures using machine learning algorithms has been gaining more recognition and shown to perform well in recent resistome data analyses and classifications (7,57,58). ResistoXplorer provides two such powerful supervised classification methods—Random Forest (59) and Support Vector Machine (SVM). Both can be applied to resistome data for identification of potential biomarkers. In particular, the Random Forest algorithm uses an ensemble of classification trees (forest), with final class prediction based on the majority vote of the ensemble. As the forest is constructed, it can provide an unbiased estimate of prediction errors by ag-

gregating cross-validation results using bootstrapped samples. Random forest also measures the importance of each feature based on the increase of the prediction errors when it is randomly shuffled. Alternatively, the SVM algorithm uses a training set of samples separated into classes to identify a hyperplane in higher dimensional feature space that generates the largest minimum distance (margin) between the samples that belongs to different classes (60). ResistoXplorer's SVM analysis is performed using recursive feature selection and sample classification via linear kernel (61). The features used by the best model are considered to be important and ranked based on their frequencies of being selected in the model. Figure 4D shows the classification performance of SVM with regards to decreasing number of features (variables).

**Other features.** There are several other valuable functions implemented in ResistoXplorer for exploratory analysis of resistome data. Users can perform core resistome analysis to detect core sets of features present in samples or sample groups based on their abundance and prevalence level (Figure 3A). ResistoXplorer also supports commonly used correlation analysis and hierarchical clustering. The results from hierarchical clustering can be visualized using heatmaps (Figure 4A) and dendrograms (Figure 4B). For publication purposes, all visualizations can be downloaded as either Scalable Vector Graphics (SVG) or Portable Document Format (PDF) files.

### Integrative analysis

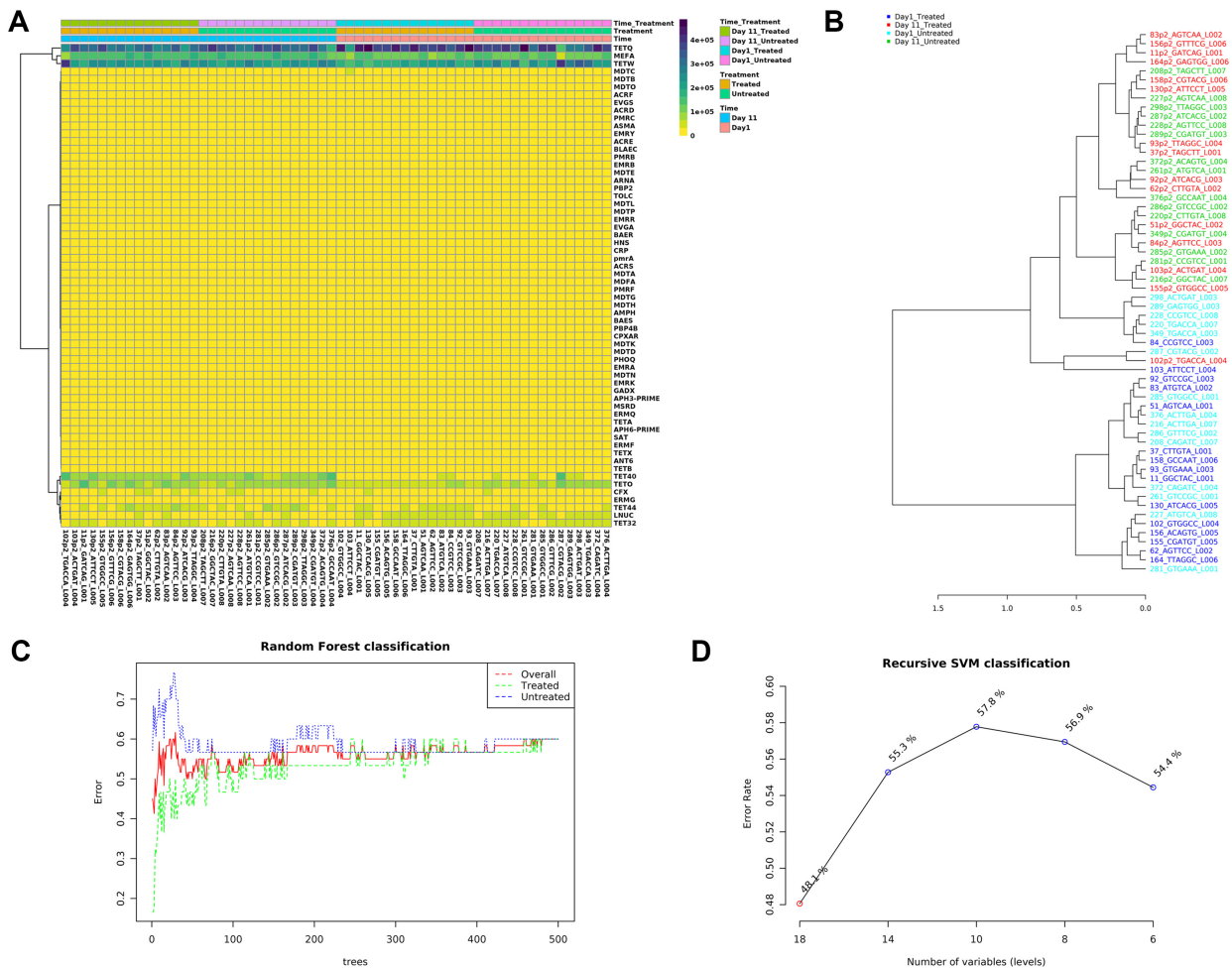
The main goal of this module is to explore and unveil potential hidden correlations between the microbiome and resistome using a variety of integrative data analysis approaches. Such analyses have been increasingly used to explore the associations between the bacteria and ARGs in different environments (11,62,63). Currently, ResistoXplorer offers several advanced and commonly used univariate and multivariate statistical methods for data integration and correlation analysis. All these analyses are performed on filtered and normalized datasets.

**Global Similarity analysis.** This section allows users to determine the overall similarity between the microbiome and resistome dataset using two multivariate correlation-based approaches: Procrustes analysis (PA) and Coinertia analysis (CIA). The datasets used for such analysis can be normalized using scaling and/or transformation approaches to account for uneven library sizes and compositional data. The analysis can be performed at various functional and taxonomic levels based on available annotations. These functions currently support five common distance measures. The ordinations from the distance matrices can be calculated using either PCoA, PCA or NMDS. By default, both these analyses are being assessed employing widely used PCoA on a Bray–Curtis distance metric. In case of PA, the microbiome ordination is scaled and rotated onto the resistome ordination to minimize the sum of squared differences between the two ordinations. For the CIA, the microbiome and resistome ordination are constrained so that the squared covariances between them are maximized to

measures the congruence between datasets. The results are represented using both 2D and 3D ordination plots, where samples are colored and shaped based on the datasets (Figure 5A). Users can also color the samples based on different experimental factors to identify some patterns or gain biological insights. The corresponding similarity coefficient and P-value from both these analyses are estimated automatically to assess the strength and significance of the association between the two datasets. The similarity coefficient ranges from 0 and 1, with 0 suggesting total similarity and 1 total dissimilarity between the two datasets. Users can perform both Procrustes and Coinertia analysis to gain more confidence by evaluating the congruency of the results.

**Omics data integration approaches.** ResistoXplorer offers other multivariate projection-based exploratory approaches, such as regularized canonical correlation analysis (rCCA) and sparse partial least square (sPLS) for the integration of microbiome and resistome data. These approaches aim at highlighting the correlations between high dimensional 'omics' datasets. They are implemented primarily based on the R mixOmics package (64). By default, both the datasets are normalized during such analyses using their recommended normalization technique (i.e. clr transformation). Users can choose or tune the number of components and regularization parameters for rCCA. In the case of sPLS, all variables are selected in both datasets by default. In addition to sample plots, other variable plots like clustered image maps (Figure 5C) and correlation circle plots (Figure 5D) are displayed to facilitate the interpretation of the complex correlation structure between datasets.

**Pairwise microbe-ARG correlation analysis.** This section enables users to determine if there are strong relationships (co-occurrence patterns) between individual microbial taxa (microbiome) and ARGs (resistome) using univariate correlation analysis. Users can perform such analysis using four different types of classical and more recent approaches, including Spearman, Pearson, CCLasso and Maximal Information Coefficient. Since the Spearman correlation analysis is most commonly used in resistome studies (65) and seems to perform overall better than other approaches in identifying pairwise associations between omics data, it has been present as a first choice in ResistoXplorer (66). In particular, features (ARGs or taxa) that are not present in half of the samples across all the groups are removed to eliminate an artificial association bias before performing Spearman correlation analysis (65,67). By default, these approaches (except CCLasso) use the microbiome and resistome relative abundances (proportions) for correlation inference, while the CCLasso is based on log-ratio normalization of microbiome and resistome compositional data. Due to the lack of consensus on each approach in different conditions, it is recommended to compare results from multiple methods (66,44). The analysis can be conducted on taxa at their different taxonomic annotations (i.e. phylum, genus and species) and on resistome at different functional levels (e.g. class, mechanisms, ARG, etc.) based on available annotations. Users can select strong and significant pairwise correlations using a combination of absolute correlation coefficients and adjusted P-value. The results are represented



**Figure 4.** Example outputs from clustering analysis and machine learning-based classifications in ARG Table module of ResistoXplorer. (A) A clustered heatmap showing the variation of resistome abundance at group level in samples organized based on time point. (B) A dendrogram showing the clustering of samples with colors based on treatment and time point. (C and D) A graphical summary of the classification performance on different treatment groups using the Random Forests and SVM algorithm, respectively.

as a co-occurrence network (Figure 6A) with each node indicating either a microbial taxon or a resistance determinant (ARG). The nodes can be sized based on their network topological measures (degree and betweenness). Users can double click on a node to highlight its corresponding correlated nodes in a network. The width and color of an edge indicate the strength and direction of the correlation between two nodes. The nodes are colored as well as shaped according to the dataset. The underlying correlation matrices and network centrality-based measures are also available to download.

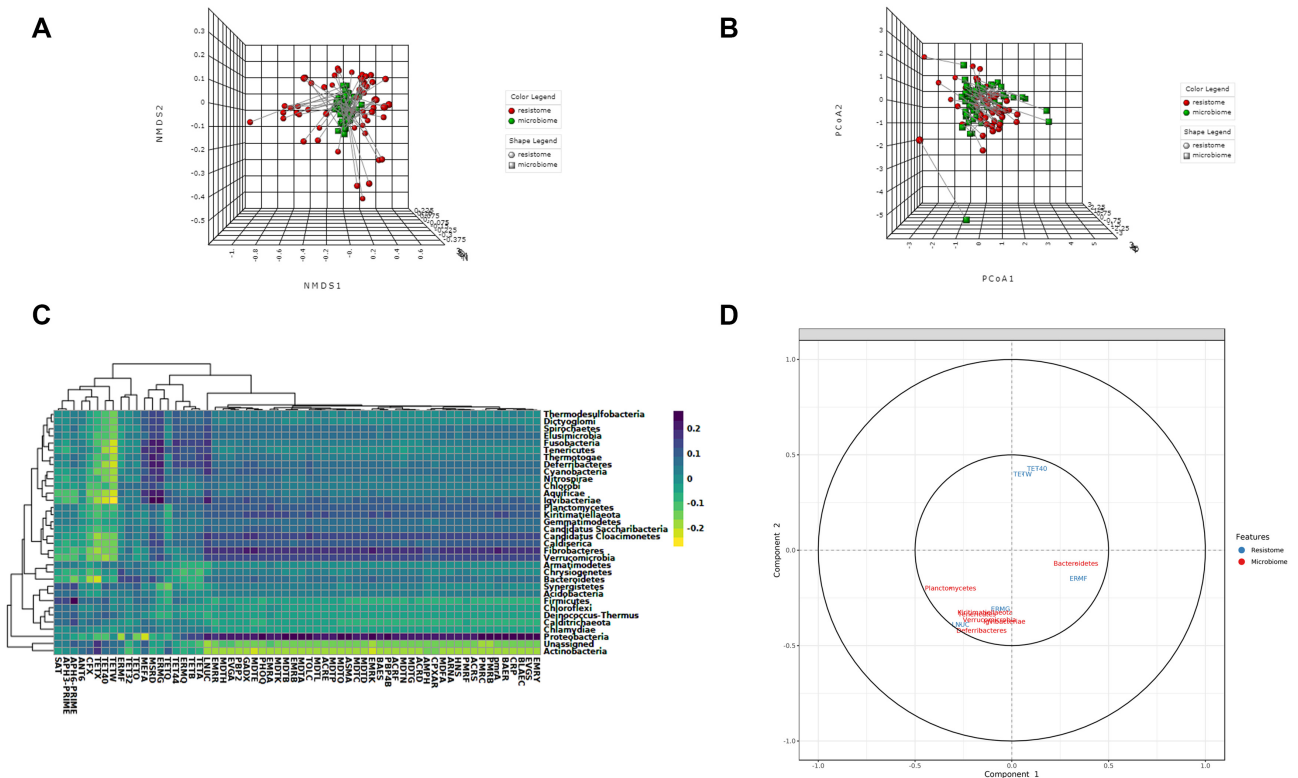
### ARGs-microbial host network exploration

The module offers users the possibility to understand the complex 'multiple-to-multiple' relations between ARGs and microbial hosts, using an advanced network-based visual analytics system. It is straightforward to identify key players from a network perspective, for instance, by looking for those ARGs that are found in multiple microbes or

by identifying those microbes that simultaneously contain multiple ARGs of interest.

**ARG-microbial host association data collection.** ResistoXplorer currently supports four reference databases (ResFinder (33), CARD (32), ARDB (39) and BacMet (41)) and a recently published AMP dataset (42) for network-based microbial host associations exploration of ARGs. These primary databases contain direct or indirect information regarding the microbial host for the reference ARGs. In the latter case, the information on microbial host associated with each of the reference ARGs has been collected from their corresponding GenBank accession number using a combination of text mining and manual curation, like in ResFinder and ARDB. Further, this information has been manually annotated to improve name readability and remove redundancy. Moreover, the available functional annotation information of ARGs was also collected and organized into sets to facilitate enrichment analysis.





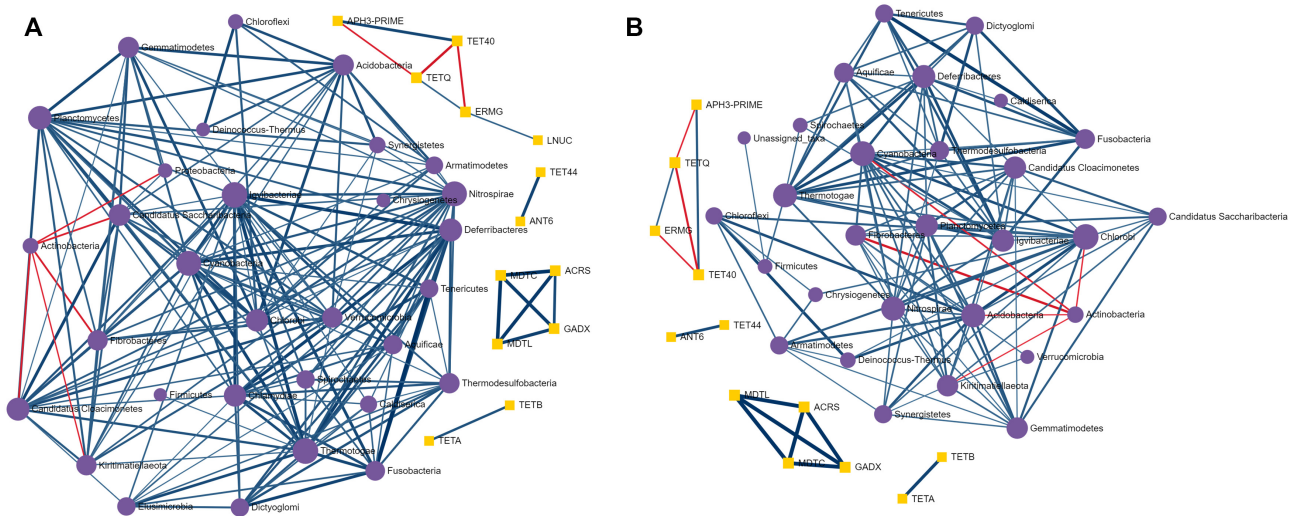
**Figure 5.** Example outputs from Integration module of ResistoXplorer. (A) A 3D NMDS plot from Procrustes analysis with samples shape and color with regards to datasets. (B) A 3D PCoA plot from Coinertia analysis, with the length of lines connecting two points indicating the similarity of samples between two datasets. (C) A clustered image heatmap showing the correlations between and among taxa (phylum level) and ARGs (group level). (D) A correlation circle plot showing the correlation structure of features (taxa/ARGs) present in two datasets.

*ARG–microbial host association table and network creation.* The uploaded list of ARGs are searched against the selected target database. This list can comprise significant ARGs detected in differential abundance testing or ARGs identified through high-throughput qPCR. The results will be represented as an association table with each row corresponding to a particular reference ARG (sequence) and its potential microbial host. When available, the table also provides other association information along with hyperlinks to the corresponding GenBank Accession number and PubMed literature. Users can directly remove each row by clicking the delete icon in the last column to keep only high-quality associations supported by literature or experimental evidence. The resulting ARG–host associations are used to build the default networks. Since not all the nodes will be connected, this approach may lead to the generation of multiple networks. The statistics of nodes and edges are provided for users to have an overview of the size and complexity of the generated networks. Further, users can also filter the nodes based on their topological measures (degree and betweenness) in case of large networks for better interpretation.

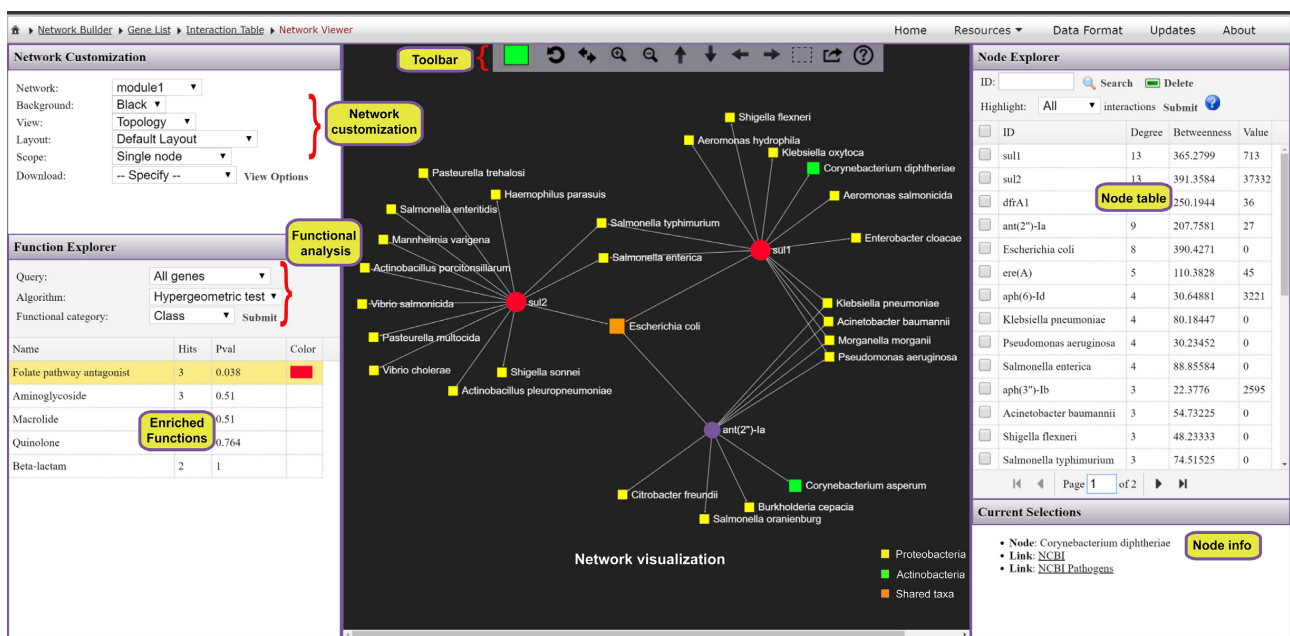
*Network visualization and functional analysis.* The resulting networks are visualized using HTML5 canvas and JavaScript-based powerful and fully featured visualization system. This system is implemented based on a previously

published visual analytics tool (68). It is comprised of three main components: the central network visualization area, the network customization and functional analysis panel on the left, and the right panel containing a node table (Figure 7). Users can intuitively visualize and manipulate the network in the central area using a mouse with a scroll wheel. For example, users can scroll the wheel to zoom in and out the network, hover the mouse over any node to view its name, click a node to display its details on the bottom-right corner or double click a node to select it. The horizontal toolbar to the top exhibits basic functions to manipulate the network. The first is the color picker to enable users to choose a highlight color for the next selection. Users can also select and drag multiple nodes by using the dashed square icon in the toolbar.

The network customization panel provides various options to configure the general visualization features of the default network or to specify the range of mouse operation. The ‘Layout’ option enables users to perform automatic network layout using different algorithms; the ‘Background’ option enables users to select between a white and black background. The range of mouse operations during highlighting and dragging-and-dropping can be varied using the ‘Scope’ option. For instance, in ‘single node’ mode, only the node which has been clicked or dragged will be highlighted or affected, whereas all the nodes that are being selected by the users will be affected in the ‘all highlights’



**Figure 6.** Illustration of pairwise correlation analysis results. (A and B) A co-occurrence network showing the strong and significant pairwise correlations between taxa (phylum level) and ARGs (class level) identified using Spearman and Pearson correlation analysis, respectively.



**Figure 7.** A screenshot of ResistoXplorer network visual analytics system. The view is divided into three main compartments with the network visualization (toolbar on top) at the center, the node table on the right and the network customization panel together with functional annotation table on the left. Users can easily highlight and manually organize different groups of nodes based on either their annotations or connectivity patterns. It is straightforward to identify those ARGs that are found in multiple microbes, or those microbes that simultaneously contain multiple ARGs of interest.

mode. Additionally, the ‘Download’ option allows users to either save the current network in different formats or to download the network file in GraphML format for visualization in other software. The node table on the right panel displays the ARGs and their microbial hosts along with corresponding network topological measures. The corresponding abundance values, if available, will also be presented in the last column. Users can directly click on any row of interest to select, and the network view will automatically zoom to the related node. The bottom right panel provides more

detailed info related to the node(s) being highlighted or currently selected on the network.

ResistoXplorer also supports functional enrichment analysis of the ARGs present within the current network using hypergeometric tests. This approach coupled with the network visualization system has the potential to provide better interpretation of AMR resistance mechanisms and inform on possible dissemination routes of ARGs. The categories or levels at which enrichment analysis can be performed will be based on the initially selected database. The

enrichment analysis results are shown as a table on the left panel. By clicking on a row of the result table, users can highlight all nodes related to an enriched function within the network. There are also several other options and functionalities present, which allow users to intuitively explore, manipulate and customize the ARG-microbial host association networks.

### Use case

To illustrate the functionality of the tool, we have selected two recent resistome studies with publicly available metagenomic datasets (69,70). These datasets have been mounted as an example sets in the 'ARG Table' module of our tool. In Doster *et al.* (70), the authors have examined the effects of tulathromycin (antimicrobial drug) on gut microbiome and resistome using commercial feedlot cattle. Two groups of cattle were used, with one untreated group and the other treated with metaphylaxis. Fecal samples were collected from 15 cattle within each group at two time points—baseline (Day 1) and after 11 days (Day 11). Shotgun sequencing was performed on the extracted metagenomic DNA, and reads were aligned by the authors to MEGARes and custom Kraken database for resistome and microbiome characterization. The resulting resistome abundance profile is uploaded to the ARG Table module of ResistoXplorer for further downstream analysis and exploration. Since the reads were annotated using the MEGARes database, we have directly selected the precompiled functional annotation information of the corresponding database to annotate and classify ARGs (gene accessions) at higher functional levels. All the ARGs are mapped and classified at three functional hierarchical levels—class, mechanism and group as per the MEGARes classification scheme. We first compared the resistome alpha diversity at the mechanism level; the Shannon diversity indices of the treatment group decrease over time, but the diversity changes are not prominent in the untreated group (Figure 2B). The composition profiling is carried out to explore and represent the gut resistome of cattle. As shown in Figure 2A, the resistome composition at the class level is dominated mainly by the ARGs that confer resistance to tetracycline and the macrolide-lincosamide-streptogramin (MLS) class of antibiotics in all the samples. The hierarchical composition profiling using the Sankey diagram showed that almost all of the ARGs that belongs to tetracycline-class confer resistance through ribosomal protection proteins. In contrast, most of the ARGs that belong to the MLS-class confer resistance through macrolide efflux pumps. Moreover, the resistome composition at different functional levels was quite similar between treated and untreated cattle (Figure 2C). The ordination analysis at AMR mechanism level using PCA (CoDA-based) and PCoA indicated that the resistome composition of treated and untreated groups were significantly different between time points (ANOSIM:  $R = 0.49$ ,  $P$ -value  $< 0.05$ ; Figure 3B and C). We also performed differential abundance testing on fecal resistome profile using metagenomeSeq and ALDEx2 (CoDA-based) at class, mechanism, and group level. No significant features were found to be differentially abundant between treatment groups using both these methods at all levels. All these analyses confirm and replicate the previous findings and results

of the original publication. We also performed additional analyses of the data to highlight the utility and exploratory capabilities of ResistoXplorer. The rarefaction curve analysis indicated that enough sequencing depth was achieved to describe the ARG richness in all the metagenomic samples (Figure 2D). The heatmap showed that most of the features (except TETQ, MEFA, TETW and TETO) at group level have a very low abundance and sparse representation across all the 60 samples. The distinct abundance pattern of TET40 group is observed when comparing the Day1 with Day11 for both the treatment groups (Figure 4A). The hierarchical clustering analysis showed that samples belonging to both treated and untreated groups are clustered effectively based on time points (Figure 4B). The ARGs belonging to tetracyclines and MLS classes comprise the core resistome in cattle, based on their abundance and prevalence level (Figure 3A). Additionally, both Random Forest and SVM algorithms suggested that the treatment groups could not be predicted with high accuracies based on the resistome profiles of fecal samples, confirming the findings that tulathromycin does not seem to influence the gut resistome in cattle (Figure 4C and D).

Furthermore, the bacterial and ARG abundance profiles are uploaded to the Integration module of ResistoXplorer to explore the relationship between the fecal microbiome and resistome in cattle. The application of Procrustes analysis suggested that there were no significant overall similarities between the resistome and bacterial abundance profile ( $M2 = 0.23$ ,  $P$ -value  $> 0.05$ ; Figure 5A). However, the results from Coinertia analysis indicated that resistome and bacterial composition are moderately correlated with statistical significance (RV coefficient = 0.47,  $P$ -value  $< 0.05$ ; Figure 5B). To deal with the result discrepancies between approaches, we also investigated the pairwise associations between individual taxa (microbiome) and ARGs (resistome) using Spearman and Pearson correlation analysis. The results showed that no significant and strong pairwise correlations (criteria: absolute correlation coefficient  $> 0.7$ ; adjusted  $P$ -value  $< 0.05$ ) were identified between any taxa (phylum level) and ARG (group level) (Figure 6A and B), suggesting that the gut resistome was not correlated and structured by bacterial composition.

### Comparison with other tools

A variety of tools or pipelines have been developed in recent years to support resistome analysis of metagenomic data (15,33,36,37,71,72). Most of these tools have been designed primarily for raw reads processing and annotation, with limited or no support for interactive visual exploration and downstream analysis. ResistoXplorer complements these tools and resources by providing real-time visual analytics experience along with comprehensive support for statistical, visual and exploratory analysis on the metagenomic resistome data. AMR++ Shiny (15) is a web-based R/shiny application dedicated for basic exploratory and statistical analysis of metagenomic resistome and microbiome data. More recently, resistomeAnalysis (67) R package also supports visualization, comparative and integrative analysis of resistome abundance profile. Based on the detailed comparisons among those tools (Table 1), it is clear that ResistoXplorer offers a unique set of features and functions with



**Table 1.** Comparison of ResistoXplorer with other web-based tools (except resistomeAnalysis R package) supporting downstream analysis of metagenomic resistome data

Tools	ResistoXplorer	AMR++ Shiny	resistomeAnalysis	WHAM!
Platform	Web-based	Web-based + locally installable	R package	Web-based
Registration	No	No		No
<b>Data processing</b>				
Data input	Abundance tables	Abundance tables	Abundance tables	Abundance tables (Biobakery + EBI)
Functional annotation	User-defined + collected from >10 AMR databases	MEGARes	CARD	User-defined
Filtering	Abundance, variance	Abundance (quantile)		Variance
Normalization	Scaling, transformation, rarefying	CSS, rarefying	TSS, proportion	Proportion, clr
<b>Composition profiling</b>				
Visual profiling	Stacked bar plot, stacked area, sankey diagram, zoomable sunburst, treemap	Stacked bar plot	Stacked bar plot	Interactive stacked bar plot
Alpha diversity	Multiple		Richness	
Ordination analysis	PCoA, NMDS & PCA (2D & 3D)	PCA & NMDS (Bray–Curtis) (2D)	PCoA (2D)	
<b>Comparative analysis</b>				
Differential analysis	DESeq2, metagenomeSeq, EdgeR, LEfSe, ALDEx2, ANCOM	metagenomeSeq	DESeq2	ALDEx2
Classification	Random Forests, SVM			
Other functions	Heatmaps, dendrogram, correlation, core resistome, alpha rarefaction curves	Heatmaps, Alpha rarefaction bar plots	Heatmaps, dendrogram, correlation, core resistome	Interactive Heatmaps, correlation
<b>Integrative analysis</b>	Procrustes, Coinertia, rCCA, sPLS, Spearman, Pearson, CCLasso, MIC		Spearman	Visual comparisons
<b>Gene (ARG) List exploration</b>	ARG-microbial host network visual analytics & functional analysis			

ResistoXplorer: <http://www.resistoxplorer.no>

AMR ++ Shiny: <https://github.com/lakinsm/amrplusplus-shiny>

resistomeAnalysis (R package): <https://github.com/blue-moon22/resistomeAnalysis>

WHAM!: [https://ruggleslab.shinyapps.io/wham\\_v1/](https://ruggleslab.shinyapps.io/wham_v1/)

**Note:** Tools exclusively dedicated for sequence annotations are not included.

regards to comprehensive statistical and exploratory data analysis, visualization, integrative data analysis and ARG-microbial network visual analysis.

### Limitations and future directions

The ARG table module can be used for visualization and analysis of resistome profiles characterizing different genetic determinants present within an AMR reference database. However, ResistoXplorer does not allow users to choose multiple precompiled functional annotation databases to analyze the entire ecologies (antimicrobial drugs, biocide, metal and other resistance drivers) of resistance determinants. More importantly, the functional annotations collected in ResistoXplorer mainly depend on the information and classification scheme present in the original databases. Hence, there might still be some acyclic and hierarchical functional annotation structure in databases, which users need to curate for accurate count-based analyses. The biocuration of ARGs and their functional annotation structure for the supported databases is beyond the scope of the study. Although some of the supported databases are no longer updated, such as ARDB and ARG-ANNOT (5,6), excluding them would have limited the possibility of exploring and analyzing previous datasets, as well as a variety of present studies that still use them. In re-

gard to the ARG-microbial host associations module, these are limited by the type and quality of information available in the databases. Currently, ResistoXplorer offers limited functionalities and features for vertical data integration and pairwise correlation analysis. As most of the advanced approaches for performing this type of analysis on multidimensional datasets are based on computationally intensive re-sampling (cross-validation) and permutation-based approaches to calculate statistical significance, which adds layers of complexity and computational power demands, that is often challenging for a real-time interactive web application. In future versions, we plan to continuously update and expand our database and analysis support for exploration of mobilome, virulome and other resistance determinants of relevance in AMR metagenomic studies.

### CONCLUSION

Whole metagenomic sequencing studies are providing unparalleled knowledge on the diversity of resistomes in the environment, animals and humans, and on the impact of interventions, such as antibiotic use (7–12,67,69). Currently, such studies and data analyses are mainly exploratory in nature. In spite of the continuous development of many new statistical approaches, there is no exclusive method that unflinchingly performs well, as demonstrated by several



benchmarking studies (20,25,26,46). Indeed, it has been recently suggested that metagenomic analysis should be explored comparatively using different available approaches (31). However, this is a time consuming task and requires knowledge and bioinformatics training on the implementation of each statistical method employed. Therefore, it is critical to assist researchers and clinical scientists in the field to easily explore their own datasets using a variety of approaches, in real-time and through interactive visualization, to facilitate data understanding and hypothesis generation. ResistoXplorer meets these requirements by offering comprehensive support for composition profiling, statistical analysis, integrative analysis and visual exploration of resistome data. Conversely, such analysis is entirely dependent on the comprehensiveness and quality of the AMR reference databases (5,6). Hence, the use of continuously updated and curated databases with a simple, acyclic and hierarchical functional annotation scheme is desired for accurate downstream analysis. Lastly, ResistoXplorer will continuously be updated to follow the advancements in approaches for resistome analysis. We believe ResistoXplorer will have the potential to find large applicability as a useful resource for researchers in the field of AMR.

## ACKNOWLEDGEMENTS

This work has been supported by and utilized the computational resources at the University of Oslo.

## FUNDING

University of Oslo; Research Council of Norway [273833, 274867]; Olav Thon Foundation [421258].  
*Conflict of interest statement.* None declared.

## REFERENCES

- Interagency Coordination Group on Antimicrobial Resistance (2019) No time to wait—securing the future from drug-resistant infections. *Rep. Secret. Gen. Nations.* <https://www.who.int/antimicrobial-resistance/interagency-coordination-group/final-report/en/>.
- Simonsen, G.S., Tapsall, J.W., Allegranzi, B., Talbot, E.A. and Lazzari, S. (2004) The antimicrobial resistance containment and surveillance approach—a public health tool. *Bull. World Health Organ.*, **82**, 928–934.
- Cecchini, M., Langer, J. and Slawomirski, L. (2015) Antimicrobial Resistance in G7 Countries and Beyond: Economic Issues, Policies and Options for Action. *Paris: Organization for Economic Co-operation and Development*, 1–75.
- Metcalfe, S., Baker, M.G., Freeman, J., Wilson, N. and Murray, P. (2016) Combating antimicrobial resistance demands nation-wide action and global governance. *NZ Med. J.*, **129**, 8–14.
- Boolchandani, M., D'Souza, A.W. and Dantas, G. (2019) Sequencing-based methods and resources to study antimicrobial resistance. *Nat. Rev. Genet.*, **20**, 356–370.
- Gupta, C.L., Tiwari, R.K. and Cytryn, E. (2020) Platforms for elucidating antibiotic resistance in single genomes and complex metagenomes. *Environ. Int.*, **138**, 105667.
- Xia, Y., Zhu, Y., Li, Q. and Lu, J. (2019) Human gut resistome can be country-specific. *PeerJ*, **7**, e6389.
- Forslund, K., Sunagawa, S., Kultima, J.R., Mende, D.R., Arumugam, M., Typas, A. and Bork, P. (2013) Country-specific antibiotic use practices impact the human gut resistome. *Genome Res.*, **23**, 1163–1169.
- Munk, P., Andersen, V.D., de Knecht, L., Jensen, M.S., Knudsen, B.E., Lukjancenko, O., Mordhorst, H., Clasen, J., Agersø, Y. and Folkesson, A. (2016) A sampling and metagenomic sequencing-based methodology for monitoring antimicrobial resistance in swine herds. *J. Antimicrob. Chemother.*, **72**, 385–392.
- Pal, C., Bengtsson-Palme, J., Kristiansson, E. and Larsson, D.J. (2016) The structure and diversity of human, animal and environmental resistomes. *Microbiome*, **4**, 54.
- Forsberg, K.J., Patel, S., Gibson, M.K., Lauber, C.L., Knight, R., Fierer, N. and Dantas, G. (2014) Bacterial phylogeny structures soil resistomes across habitats. *Nature*, **509**, 612–616.
- Hendriksen, R.S., Munk, P., Njage, P., Van Bunnik, B., McNally, L., Lukjancenko, O., Röder, T., Nieuwenhuijse, D., Pedersen, S.K. and Kjeldgaard, J. (2019) Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nat. Commun.*, **10**, 1124.
- McArthur, A.G. and Wright, G.D. (2015) Bioinformatics of antimicrobial resistance in the age of molecular epidemiology. *Curr. Opin. Microbiol.*, **27**, 45–50.
- Hendriksen, R.S., Bortolaia, V., Tate, H., Tyson, G., Aarestrup, F.M. and McDermott, P. (2019) Using genomics to track global antimicrobial resistance. *Front. Pub. Health*, **7**, 242.
- Lakin, S.M., Dean, C., Noyes, N.R., Dettenwanger, A., Ross, A.S., Doster, E., Rovira, P., Abdo, Z., Jones, K.L. and Ruiz, J. (2016) MEGARes: an antimicrobial resistance database for high throughput sequencing. *Nucleic Acids Res.*, **45**, D574–D580.
- Chong, J. and Xia, J. (2017) Computational approaches for integrative analysis of the metabolome and microbiome. *Metabolites*, **7**, 62.
- Li, H. (2015) Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annu. Rev. Stat. Appl.*, **2**, 73–94.
- Calle, M.L. (2019) Statistical analysis of metagenomics data. *Genomics Inform.*, **17**, e6.
- Bengtsson-Palme, J., Larsson, D.J. and Kristiansson, E. (2017) Using metagenomics to investigate human and environmental resistomes. *J. Antimicrob. Chemother.*, **72**, 2690–2703.
- Weiss, S., Xu, Z.Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J.R., Vázquez-Baeza, Y. and Birmingham, A. (2017) Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, **5**, 27.
- Filzmoser, P., Hron, K. and Reimann, C. (2009) Univariate statistical analysis of environmental (compositional) data: problems and possibilities. *Sci. Total Environ.*, **407**, 6100–6108.
- Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
- Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
- Paulson, J.N., Pop, M. and Bravo, H.C. (2013) metagenomeSeq: statistical analysis for sparse high-throughput sequencing. *Bioconductor Package*, **1**, 91.
- Pereira, M.B., Wallroth, M., Jonsson, V. and Kristiansson, E. (2018) Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics*, **19**, 274.
- Jonsson, V., Österlund, T., Nerman, O. and Kristiansson, E. (2016) Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. *BMC Genomics*, **17**, 78.
- Quinn, T.P., Erb, I., Richardson, M.F. and Crowley, T.M. (2018) Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, **34**, 2870–2878.
- Gloor, G.B. and Reid, G. (2016) Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Can. J. Microbiol.*, **62**, 692–703.
- Gloor, G.B., Macklaim, J.M., Pawlowsky-Glahn, V. and Egozcue, J.J. (2017) Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.*, **8**, 2224.
- Quinn, T.P., Erb, I., Gloor, G., Notredame, C., Richardson, M.F. and Crowley, T.M. (2019) A field guide for the compositional analysis of any-omics data. *GigaScience*, **8**, giz107.
- Pérez-Cobas, A.E., Gomez-Valero, L. and Buchrieser, C. (2020) Metagenomic approaches in microbial ecology: an update on whole-genome and marker gene sequencing analyses. *Microbial Genomics*, **6**, mgen000409.
- Alcock, B.P., Raphenya, A.R., Lau, T.T., Tsang, K.K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A.-L.V., Cheng, A.A. and Liu, S. (2020) CARD 2020: antibiotic resistance surveillance with the

- comprehensive antibiotic resistance database. *Nucleic Acids Res*, **48**, D517–D525.
33. Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O., Aarestrup, F.M. and Larsen, M.V. (2012) Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.*, **67**, 2640–2644.
  34. Doster, E., Lakin, S.M., Dean, C.J., Wolfe, C., Young, J.G., Boucher, C., Belk, K.E., Noyes, N.R. and Morley, P.S. (2019) MEGARes 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. *Nucleic Acids Res.*, **48**, D561–D569.
  35. Feldgarden, M., Brover, V., Haft, D.H., Prasad, A.B., Slotta, D.J., Tolstoy, I., Tyson, G.H., Zhao, S., Hsu, C.-H. and McDermott, P.F. (2019) Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob. Agents Chemother.*, **63**, e00483-19.
  36. Yin, X., Jiang, X.-T., Chai, B., Li, L., Yang, Y., Cole, J.R., Tiedje, J.M. and Zhang, T. (2018) ARGs-OAP v2. 0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. *Bioinformatics*, **34**, 2263–2270.
  37. Arango-Argoty, G., Garner, E., Pruden, A., Heath, L.S., Vikesland, P. and Zhang, L. (2018) DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, **6**, 23.
  38. Argoty, G.A.A., Guron, G.K., Garner, E., Riquelme, M.V., Heath, L., Pruden, A., Vikesland, P. and Zhang, L. (2020) ARG-miner: a web platform for crowdsourcing-based curation of antibiotic resistance genes. *Bioinformatics*, **36**, 2966–2973.
  39. Liu, B. and Pop, M. (2008) ARDB—antibiotic resistance genes database. *Nucleic Acids Res.*, **37**, D443–D447.
  40. Gupta, S.K., Padmanabhan, B.R., Diene, S.M., Lopez-Rojas, R., Kempf, M., Landraud, L. and Rolain, J.-M. (2014) ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob. Agents Chemother.*, **58**, 212–220.
  41. Pal, C., Bengtsson-Palme, J., Rensing, C., Kristiansson, E. and Larsson, D.J. (2013) BacMet: antibacterial biocide and metal resistance genes database. *Nucleic Acids Res.*, **42**, D737–D743.
  42. Kintsés, B., Méhi, O., Ari, E., Számel, M., Györkei, Á., Jangir, P.K., Nagy, I., Pál, F., Fekete, G. and Tengölics, R. (2019) Phylogenetic barriers to horizontal transfer of antimicrobial peptide resistance genes in the human gut microbiota. *Nat. Microbiol.*, **4**, 447–458.
  43. Dhariwal, A., Chong, J., Habib, S., King, I.L., Agellon, L.B. and Xia, J. (2017) MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res.*, **45**, W180–W188.
  44. Ni, Y., Yu, G., Chen, H., Deng, Y., Wells, P.M., Steves, C.J., Ju, F. and Fu, J. (2020) M2IA: a web server for microbiome and metabolome integrative analysis. *Bioinformatics*, **36**, 3493–3498.
  45. Bourgon, R., Gentleman, R. and Huber, W. (2010) Independent filtering increases detection power for high-throughput experiments. *Proc. Natl Acad. Sci. U.S.A.*, **107**, 9546–9551.
  46. McMurdie, P.J. and Holmes, S. (2014) Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.*, **10**, e1003531.
  47. McKnight, D.T., Huerlimann, R., Bower, D.S., Schwarzkopf, L., Alford, R.A. and Zenger, K.R. (2019) Methods for normalizing microbiome data: an ecological perspective. *Methods Ecol. Evol.*, **10**, 389–400.
  48. McMurdie, P.J. and Holmes, S. (2013) phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, **8**, e61217.
  49. Dixon, P. (2003) VEGAN, a package of R functions for community ecology. *J. Veg. Sci.*, **14**, 927–930.
  50. Bengtsson-Palme, J. (2018) The diversity of uncharacterized antibiotic resistance genes can be predicted from known gene variants—but not always. *Microbiome*, **6**, 125.
  51. Goodrich, J.K., Di Rienzi, S.C., Poole, A.C., Koren, O., Walters, W.A., Caporaso, J.G., Knight, R. and Ley, R.E. (2014) Conducting a microbiome study. *Cell*, **158**, 250–262.
  52. Kuczynski, J., Liu, Z., Lozupone, C., McDonald, D., Fierer, N. and Knight, R. (2010) Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat. Methods*, **7**, 813–819.
  53. Anderson, M.J. and Walsh, D.C. (2013) PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing? *Ecol. Monogr.*, **83**, 557–574.
  54. Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S. and Huttenhower, C. (2011) Metagenomic biomarker discovery and explanation. *Genome Biol.*, **12**, R60.
  55. Fernandes, A.D., Reid, J.N., Macklaim, J.M., McMurrough, T.A., Edgell, D.R. and Gloor, G.B. (2014) Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, **2**, 15.
  56. Mandal, S., Van Treuren, W., White, R.A., Eggesbø, M., Knight, R., Peddada, S.D. and disease (2015) Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbiol. Ecol. Health*, **26**, 27663.
  57. Gasparrini, A.J., Wang, B., Sun, X., Kennedy, E.A., Hernandez-Leyva, A., Ndao, I.M., Tarr, P.I., Warner, B.B. and Dantas, G. (2019) Persistent metagenomic signatures of early-life hospitalization and antibiotic treatment in the infant gut microbiota and resistome. *Nat. Microbiol.*, **4**, 2285–2297.
  58. Rahman, S.F., Olm, M.R., Morowitz, M.J. and Banfield, J.F. (2018) Machine learning leveraging genomes from metagenomes identifies influential antibiotic resistance genes in the infant gut microbiome. *MSystems*, **3**, e00123-17.
  59. Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
  60. Burges, C.J. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining Knowledge Discov.*, **2**, 121–167.
  61. Zhang, X., Lu, X., Shi, Q., Xu, X.-q., EL, H.-., Harris, L.N., Iglehart, J.D., Miron, A., Liu, J.S. and Wong, W.H. (2006) Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics*, **7**, 197.
  62. Ruppé, E., Ghoulane, A., Tap, J., Pons, N., Alvarez, A.-S., Maziers, N., Cuesta, T., Hernando-Amado, S., Clares, I. and Martínez, J.L. (2019) Prediction of the intestinal resistome by a three-dimensional structure-based method. *Nat. Microbiol.*, **4**, 112–123.
  63. Liu, J., Taft, D.H., Maldonado-Gomez, M.X., Johnson, D., Treiber, M.L., Lemay, D.G., DePeters, E.J. and Mills, D.A. (2019) The fecal resistome of dairy cattle is associated with diet during nursing. *Nat. Commun.*, **10**, 4406.
  64. Rohart, F., Gautier, B., Singh, A. and Lê Cao, K.-A. (2017) mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.*, **13**, e1005752.
  65. Li, B., Yang, Y., Ma, L., Ju, F., Guo, F., Tiedje, J.M. and Zhang, T. (2015) Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. *ISME J.*, **9**, 2490–2502.
  66. You, Y., Liang, D., Wei, R., Li, M., Li, Y., Wang, J., Wang, X., Zheng, X., Jia, W. and Chen, T. (2019) Evaluation of metabolite-microbe correlation detection methods. *Anal. Biochem.*, **567**, 106–111.
  67. Carr, V.R., Witherden, E.A., Lee, S., Shoaie, S., Mullany, P., Proctor, G.B., Gomez-Cabrero, D. and Moyes, D.L. (2020) Abundance and diversity of resistomes differ between healthy human oral cavities and gut. *Nat. Commun.*, **11**, 693.
  68. Fan, Y., Siklenka, K., Arora, S.K., Ribeiro, P., Kimmins, S. and Xia, J. (2016) miRNet-dissecting miRNA-target interactions and functional associations through network-based visual analysis. *Nucleic Acids Res.*, **44**, W135–W141.
  69. Munk, P., Knudsen, B.E., Lukjancenko, O., Duarte, A.S.R., Van Gompel, L., Luiken, R.E., Smit, L.A., Schmitt, H., Garcia, A.D. and Hansen, R.B. (2018) Abundance and diversity of the faecal resistome in slaughter pigs and broilers in nine European countries. *Nat. Microbiol.*, **3**, 898–908.
  70. Doster, E., Rovira, P., Noyes, N.R., Burgess, B.A., Yang, X., Weinroth, M.D., Lakin, S.M., Dean, C.J., Linke, L. and Magnuson, R. (2018) Investigating effects of tulathromycin metaphylaxis on the fecal resistome and microbiome of commercial feedlot cattle early in the feeding period. *Front. Microbiol.*, **9**, 1715.
  71. Kaminski, J., Gibson, M.K., Franzosa, E.A., Segata, N., Dantas, G. and Huttenhower, C. (2015) High-specificity targeted functional profiling in microbial communities with ShortBRED. *PLoS Comput. Biol.*, **11**, e1004557.
  72. Inouye, M., Dashnow, H., Raven, L.-A., Schultz, M.B., Pope, B.J., Tomita, T., Zobel, J. and Holt, K.E. (2014) SRST2: rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med.*, **6**, 90.