Downloaded from orbit.dtu.dk on: Apr 20, 2024

DTU Library

# Analysis of global emergence and spread of antimicrobial resistance in 214k host and environmental samples

**Martiny, Hannah-Marie**

*Publication date:*
2022

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

Link back to DTU Orbit

*Citation (APA):*
Martiny, H-M. (2022). *Analysis of global emergence and spread of antimicrobial resistance in 214k host and environmental samples*. Technical University of Denmark.

**General rights**
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

# Analysis of global emergence and spread of antimicrobial resistance in ~~150k~~ 214k host and environmental samples

Hannah-Marie Martiny

PhD thesis

DTU Food
National Food Institute

**Supervisor:** Associate Professor Thomas Nordahl Petersen

**Co-supervisor:** Assistant Professor Patrick Munk

**Co-supervisor:** Associate Professor Christian Brinch

**Co-supervisor:** Professor Frank M. Aarestrup



**DTU Food**

**Research Group of Genomic Epidemiology**

**Technical University of Denmark**

Kongens Lyngby, Denmark

# Summary

Antimicrobial resistance (AMR) has developed rapidly and now threatens to undermine our treatment of infectious diseases unless the spread of antimicrobial resistance genes (ARGs) is halted. AMR is a global challenge impacting human, animal, and planetary health, which calls for monitoring the abundance of ARGs across these three areas. Most of the current AMR surveillance systems only focus on clinical prevalences, failing to acknowledge that ARGs are also present in non-pathogenic microbes outside of hospital settings. In metagenomic sequencing, the goal is to recover genes for all organisms in a host or an environmental sample. That way, the abundance of ARGs can be quantified across all organisms in a sample, both unknown and known species.

Today, there is a vast amount of metagenomic sequencing datasets available in public repositories due to the good data-sharing practices during the academic publishing process. Most samples remain underutilized as few researchers have the computational and bioinformatic resources to analyze terabytes of data. However, the potential amount of information on microbial and AMR dynamics that can be extracted from these datasets using a standardized approach makes it worthwhile to explore. This has been the goal of this PhD project, where 214,095 metagenomic datasets were retrieved and analyzed to characterize the abundance of ARGs in host and environmental sources. With such a large pool of data, there are many questions on the distribution of ARGs that can be answered, and this project has focused on studying the differences in abundance for individual ARGs in local ecological settings and the overall co-occurrence of ARGs at a global scale. The three manuscripts enclosed in this PhD are presented below.

In **Manuscript I**, we carried out the download and processing of the 214,095 metagenomic datasets from the European Nucleotide Archive (ENA). Using the $442 \cdot 10^{12}$ basepairs of sequencing reads, we aligned the reads against reference sequences from two databases, Silva and ResFinder, to determine the abundance of ARGs and bacterial genes. In this publication, we presented a brief characterization of overall trends in this collection and observed differences in resistome and microbiome compositions between different sample types. We also made the count data and the curated metadata available to promote the reuse of publicly available samples and further encourage

sharing of raw sequencing data.

In **Manuscript II**, we studied the distribution of the family of *mcr* genes in the 214K collection of metagenomic samples. The *mcr* genes confer resistance to colistin, a last-resort antibiotic that is only used when all other treatment options fail. Our results confirmed that some of the *mcr* genes had spread around the world a while before being discovered. For example, we saw that the *mcr*-9 gene had been circulating for almost a decade before it was first reported. We also concluded that the differences in *mcr* abundances could largely be explained by the sampling source and location but that the genomic context of the *mcr* gene had not undergone significant changes. This manuscript confirmed the value of using publicly available metagenomic datasets for AMR surveillance and how the results can supplement existing surveillance programs.

In **Manuscript III**, following the characterization of only one group of genes in Manuscript II, we decided to investigate the abundance of all ARGs and how they co-occur. By inferring pairwise ARG correlations, we constructed correlation networks for different ecosystems that suggested that ARGs encoding resistance to different antimicrobials influences each other abundances. These observations suggest that using one antimicrobial in a specific environment induces the risk of resistance to multiple kinds of antimicrobials being indirectly selected, including resistance to the most critical antimicrobials for human medicine. We argued that the correlations could be used as risk profiles for guiding the safe use of antimicrobials in different settings.

Understanding how ARGs have spread through various ecological settings will be an effective tool in controlling and hopefully stopping the spread of AMR. The results presented in this thesis show how valuable it is to utilize sequencing datasets that are freely available in public repositories, coming one step closer to enabling global surveillance of AMR.

# Resumé

Antibiotikaresistens har udviklet sig med en alarmerende hastighed og truer nu med at gøre behandlinger imod infektionssygdome ineffektive, medmindre at spredningen af antibiotikaresistensgener stoppes. Antibiotikaresistens er en global sundhedsudfordring for både mennesker, dyr og naturen. Det er derfor nødvendigt at etablere et overvågningssystem af resistensgener på tværs af de tre fokusområder. Langt størstedelen af de eksisterende antibiotikaresistens overvågningsprogrammer fokuserer dog udelukkende på resistens i sygdomsfremkaldende bakterier i kliniske og andre udvalgte miljøer. De programmer negligerer derved det faktum, at resistensgener også kan fremkomme i ikke-sygdomsfremkaldende mikrober i alle mulige forskellige miljøer. I metagenomsekventering er målet at ekstrahere genetisk materiale for alle mikroorganismer i en prøve. Derved kan mængden af resistensgener i prøven kvantificeres for alle organismer i en prøve, både ukendte og kendte arter.

Takket være den gode praksis at dele ens rå sekventeringsdata i forbindelse med publiceringen af videnskabelige artikler, er der i dag en stor mængde af metagenomiske prøver tilgængelige i offentlige databaser. Desværre er der kun en lille del af datasættene, som bliver brugt af andre, hvilket skyldes at det kun er få forskere der har adgang til en tilstrækkelig computerkapacitet og de nødvendige bioinformatiske ressourcer. Anvendes der en standardiseret tilgang til at analysere prøverne, kan der udtrækkes en hel del ny viden om sammensætningen af mikrober og resistensgener. Dette har været det overordnede mål i denne Ph.d.-afhandling, hvor 214,095 metagenomiske datasæt er blevet brugt til at analysere fordelingen af resistensgener i lokale biologiske miljøer og tendensen til resistensgener til at fremkommer samtidigt på et globalt niveau. De tre videnskabelige artikler der er inkluderet i denne afhandling, er præsenteret nedenfor.

I **artikel I**, blev selve indsamlingen og behandlingen af de 214,095 metagenomiske datasæt fra European Nucleotide Archive (ENA) udført. De $442 \cdot 10^{12}$ basepar af rå sekventeringsreads blev alignet til referencesekvenser fra de to databaser ResFinder og Silva, således at hyppigheden af resistensgener og bakterier kunne bestemmes. I artiklen præsenterede vi selve kollektionen ved at give et overblik over de bakterielle- og resistensmønstre. Ud fra dette, kunne vi observere, at der var klare forskelle i både resistomet og mikrobiomet i de mange forskellige prøvetyper. Desuden offentliggjorde vi også alle resultater af vores read alignments og det standardiserede metadata med

det formål at få andre forskere til at drage nytte af vores kollektion og fremhæve værdien i, at rådata deles.

I **artikel II**, undersøgte vi fordelingen af gener for mobilt colistin resistens, *mcr*-generne, i de 214K metagenomer. *mcr*-gener danner resistens mod colistin, hvilket er et antibiotikum der kun bruges som den absolut sidste behandlingsmulighed. Vores resultater viste, at nogle af *mcr*-generne har været i kredsløb et stykke tid før de blev opdaget. Eksempelvis så vi, at *mcr*-9 havde cirkuleret i et årti, før det første gang blev rapporteret. Desuden observerede vi forskelle i hyppigheden af individuelle *mcr*-gener primært skyldtes prøvelokationen og prøvekilden, men at der dog ikke var nogle store ændringer i den genomiske placering af *mcr* genet. Denne artikel bekræftede den store værdi, der er i at analysere metagenomiske prøver fra offentlige databaser og hvordan resultaterne kan bruges til at udvide eksisterende antibiotikaresistens overvågningsprogrammer.

I **artikel III**, undersøgte vi forekomsten af alle resistensgener i stedet for kun én specifik familie. Vores analyse af korrelationen mellem par af to resistensgener viste, at resistensgener der virker mod forskellige antibiotika havde en tendens til oftest at følges ad. Dog var der en klar separation mellem forskellige økosystemer, hvilket vi argumenterede for kunne bruges til at skabe risikoprofiler for hvordan resistens udvikles. Risikoprofilerne indikerede, at hvis et antibiotikum bruges i et specifikt miljø, så bliver resistens til visse andre antibiotika indirekte selekteret for. Artiklen konkluderede, at risikoprofilerne bør blive brugt som en vejledning til hvordan antibiotika kan blive brugt med fornuft uden at der bliver indirekte selekteret for resistens.

Forståelsen for spredningen af resistensgener på et globalt plan er vigtigt for at kunne kontrollere, og på sigt stoppe, spredningen af antibiotikaresistens. Resultaterne fra dette Ph.d.-projekt viser værdien i hvor meget information der kan udledes fra metagenomiske analyser, og at det samlede projekt kan ses som et skridt på vejen til at etablere et verdensomspændende overvågningssystem af antibiotikaresistens.

# Preface

This PhD was carried out at the National Food Institute (Food) at the Technical University of Denmark (DTU) between 15.11.2019 and 14.11.2022. The project was supervised by Associate Professor Thomas Nordahl Petersen as the main supervisor and co-supervised by Associate Professor Christian Brinch, Assistant Professor Patrick Munk, and Professor Frank M. Aarestrup. A part of the research was completed at the University of Vic, Spain, under the guidance of Professor M. Luz Calle.

The PhD was funded as part of the Global Surveillance of Antimicrobial Resistance project under the Novo Nordisk Foundation grant NNF16OC0021856 and as part of the VEO project that got funding from the European Union Horizon 2020 program under grant agreement 874735.

Kongens Lyngby, 14$^{th}$ November 2022

Hannah-Marie Martiny

# Acknowledgements

When I began this PhD, the road to the finish line seemed quite straightforward, but despite different disruptions such as a couple of lockdowns, there are several people that have made this journey a very enjoyable one to which I owe my thanks.

First and foremost, thanks to my main supervisor Thomas Nordahl Petersen. Your support would not have made this PhD what it is, especially your calm and positive attitude has made this experience very enjoyable. Secondly, the co-supervision by Patrick Munk and Christian Brinch has also shaped this project greatly, whether it be from our bi-weekly meeting where everything from data analysis and idea generation to silly puns was discussed or always being ready to provide feedback and guidance. My thanks also go to Frank M. Aarestrup, you have been an integral part of this PhD by having tons of ideas and providing fact checks, spontaneous requests, and often well-deserved comments. The involvement of you four has been incredible in making this PhD what it is.

I was fortunate enough to go on an external research stay to work with Malu Luz Calle at the University of Vic in Spain, for which I want to give thanks for the many discussions about compositional data and I have enjoyed learning from you. To the people, I met at UVic that made my two visits very enjoyable.

Thanks to Markus Hans Kristofer Johansson, Derya Aytan-Aktug, Judit Szarvas, Nikiforos Pyrounakis, JD Martin, Amalia Bogri, Alexander Gmeiner, Baptiste Jacques Philippe Avot, Nermin Ghith, Timmie Lagermann, Laura Elmlund Kohl Birkedahl, Maja Lykke Brinch, Malte Bjørn Hallgren, Alfred Ferrer Florensa, Line Jensen Ostenfeld, Saria Otani, and the other current and former members of the Research Group for Genomic Epidemiology for all the chit-chats, lunches, and other great times during the last three years.

Thanks to my family and friends for supporting me through the various stages of the project, be it both for letting me vent my frustrations and enjoying the upsides over food and beers. To my friends that I met throughout my studies at DTU, Marianne Helenius, Michelle Lind Østrup, Marie Højmark Fischer, Cecilie Edelmann Bertelsen, and to those from VG, Rasmus Amund Henriksen, Marie Garnæs, Ida Sophie Brun, Nynne Kajs, and to all the others that I have met throughout the many years.

To my dad, Steen Martiny, who has always encouraged me to pursue science and always showed great interest in my education and work. To my mom Janne Olsen and sister Elisa Martiny for being just as encouraging and helpful with everything. For the scientific inquiries and motivation from Adam and Jennifer Martiny.

My warmest thanks to everyone.
- Hannah-Marie Martiny

# List of publications

## Publications included in the thesis

1. **Martiny, H. M.**, Munk, P., Brinch, C., Aarestrup, F. M., & Petersen, T. N. (2022). A curated data resource of 214K metagenomes for characterization of the global antimicrobial resistome. PLOS Biology 20(9): e3001792.

2. **Martiny, H. M.**, Munk, P., Brinch, C., Szarvas, J., Aarestrup, F. M., & Petersen, T. N. (2022). Global Distribution of *mcr* Gene Variants in 214K Metagenomic Samples. mSystems, 7(2), e00105-22.

3. **Martiny, H. M.**, Munk, P., Brinch, C., Aarestrup, F. M., Calle, M. L, & Petersen, T. N. Utilizing co-abundances of antimicrobial resistance genes to identify potential co-selection in the resistome. Manuscript in preparation.

## Publications not included in the thesis

1. **Martiny, H. M.**, Armenteros, J. J. A., Johansen, A. R., Salomon, J., & Nielsen, H. (2021). Deep protein representations enable recombinant protein expression prediction. Computational Biology and Chemistry, 95, 107596.

2. Thumuluri, V., **Martiny, H. M.**, Almagro Armenteros, J. J., Salomon, J., Nielsen, H., & Johansen, A. R. (2022). NetSolP: predicting protein solubility in Escherichia coli using language models. Bioinformatics, 38(4), 941-946.

# Abbreviations

| | |
|---|---|
| **ALR** | Additive log-ratio |
| **AMR** | Antimicrobial resistance |
| **ARG** | Antimicrobial resistance gene |
| **bp** | basepair |
| **CLR** | Centered log-ratio |
| **CoDa** | Compositional Data |
| **DA** | Differential Abundance |
| **ENA** | European Nucleotide Archive |
| **FAIR** | Findability, Accessibility, Interoperability, and Reusability |
| **HGT** | Horizontal gene transfer |
| **HTS** | High-throughput sequencing |
| **INSDC** | International Nucleotide Sequence Database Collaboration |
| **MAG** | Metagenome assembled genome |
| **MDR** | Multidrug-resistant |
| **MGE** | Mobile genetic element |
| **NGS** | Next-Generation Sequencing |
| **PCA** | Principal Component Analysis |
| **PCR** | Polymerase chain reaction |
| **rRNA** | ribosomal rRNA |
| **SDG** | Sustainable Development Goal |
| **WGS** | Whole-genome shotgun |
| **WHO** | World Health Organization |

# Contents

**Summary** i

**Resumé** iii

**Preface** v

**Acknowledgements** vii

**List of publications** ix

**Abbreviations** xi

**Contents** xiii

**I Introduction** 1

**1 The Threat of Antimicrobial Resistance** 3
1.1 The origins of AMR . . . 4
1.1.1 Natural resistance . . . 4
1.1.2 Acquired resistance . . . 5
1.2 Dissemination of AMR . . . 6
1.2.1 Emergence and distribution of AMR in clinical settings . . . 6
1.2.2 Environmental sources of AMR . . . 7
1.2.3 AMR, One Health, and the Sustainable Development Goals . . . 9
1.3 Problem statement . . . 9

**2 Epidemiological surveillance with (meta)genomics** 11
2.1 A brief history of DNA sequencing . . . 12
2.1.1 First generation sequencing . . . 13
2.1.2 Second generation sequencing . . . 13
2.1.3 Third generation sequencing . . . 14
2.2 The wealth of sequencing data . . . 16
2.2.1 *De-novo* assembly of NGS reads . . . 16

2.2.2 Mapping of NGS reads . . . . . 18
2.2.3 Epidemiological uses of assemblies and read mappings . . . . . 18
2.2.4 Reusability of sequencing data . . . . . 20

**3 Analyzing metagenomes using Compositional Methods 23**
3.1 Handling the sparsity of count data . . . . . 26
3.2 Analyzing metagenomic data with CoDa methods . . . . . 26

**II Manuscripts 31**

**4 Manuscript I 33**

**5 Manuscript II 49**

**6 Manuscript III 77**

**III Conclusion 109**

**7 Conclusion 111**

**Bibliography 115**

# Part I

# Introduction

CHAPTER 1

# The Threat of Antimicrobial Resistance

The wide distribution and diversity of microorganisms mean that some are good, some are neutral, and some are bad. The latter group can cause infections with sometimes fatal outcomes. Up until the 20th century, such microbial infections were tough to treat. It was not until 1904 that Paul Erlich began his search for a "magic bullet" that could target only disease-causing microbes. Erlich systematically screened hundreds of compounds to find the one drug that worked against syphilis, and today this approach is one of the most commonly used in the search for new drugs [1].

1928 marks the year of one of the most significant turning points in the history of modern medicine when Sir Alexander Fleming accidentally discovered penicillin [2]. While Fleming is credited with the discovery, it took another 12 years before the protocol for mass production and distribution of penicillin was published by Howard Florey and Ernest Chain in 1940 [1]. The administration of penicillin to treat infections during the Second World War was a huge success and inspired a worldwide search to find and produce more antibiotics [3, 4]. In fact, the period between the 1940s and 1960s is called the golden age of antibiotics because new antibiotics were discovered almost yearly [1], and most major classes of antibiotics were discovered during this period (Table 1.1) [5].

Medical agents that kill or inhibit microbes are called antimicrobials and are classified according to which microbial organism they work against. Antibiotics target bacteria, antifungals work on fungi, and antiparasitics for parasites. The widespread and successful use of antimicrobials has been estimated to have extended the average human life expectancy by 23 years [5]. During the previously mentioned golden age, many antimicrobials were discovered by screening microbes sampled from nature using the Waksman platform, which was built following the same idea as Ehrlich [6, 7]. The Waksman platform was used over the next 20 years and most of the major

antimicrobial classes were discovered during this time.

Unfortunately, the microbes have been fighting back by developing Antimicrobial resistance (AMR). Alongside the discovery of natural antimicrobials, AMR also started to emerge. To begin with, modifying existing antimicrobials, known as semisynthetic drugs, and, later, creating fully synthetic antimicrobial agents were sufficient to combat resistance [7, 8]. However, soon the rate of discovery started to decline, and the spread of AMR increased. It became apparent in the 1990s that AMR was rising faster than new antimicrobials were discovered [7]. As seen in Table 1.1, the time between the clinical introduction and the first case of resistance can be quite short, often only a few years.

Today, AMR threatens to unravel the last century of medical achievements. Experts estimate that if the resistance problem is not handled, by 2050, AMR will be the leading cause of death on a global scale [9]. A recent study by Murray et al. [10] estimated that in 2019, 4.95 million deaths were associated with bacterial AMR; of these, 1.27 million were directly attributed to bacterial AMR.

## 1.1 The origins of AMR

Considering that microorganisms develop antimicrobial resistance as a survival mechanism, it is not surprising that genes of AMR, the Antimicrobial resistance genes (ARGs), have been detected in 1-2,000-year-old human fecal samples [11] and samples from 30,000-year-old permafrost sediments [12]. Still, it is important to emphasize that the presence of ARGs has not always been to compete against humans but rather to create an equilibrium between antimicrobial-producing microbes and resistant microbes. Especially in nutrient-limiting environments, the benefit of producing antimicrobials or having AMR genes is to outcompete other microbes and gain access to more resources [13]. Many of those that produce antimicrobials also carry the corresponding resistance genes in their genome to avoid suicide [14]. With the introduction of antimicrobial agents in modern medicine, the pressure for microbes to obtain and keep ARGs has become much dire. The occurrence of ARGs in genomes is either due to natural or acquired mechanisms.

### 1.1.1 Natural resistance

Natural resistance mechanisms refer to ARGs found to be naturally occurring in a host's chromosome. ARGs can be grouped as intrinsic (always expressed) or induced upon exposure to an antimicrobial. Intrinsic resistance is typically defined as a trait

common to a species that does not depend on previous exposure and is not acquired via horizontal gene transfer [15, 16]. One example of an intrinsic resistance mechanism is the efflux pumps, where the antimicrobial is expelled directly from the cell [16]. ARGs can occur either as a natural phenomenon or acquired through several mechanisms.

### 1.1.2 Acquired resistance

The organism can also obtain resistance through mutations or acquisition. If mutations occur in genes encoding drug targets, antibiotic-modifying enzymes, drug transporters or regulators, the microbe might develop AMR. However, this mutation-aided AMR often comes at the cost of fitness decreasing [16].

Genetic material can also be shared between microorganisms of different species through Horizontal gene transfer (HGT) in one of the three main routes: transformation, where the bacteria can take up free DNA from the environment: transduction, where bacteriophages mediate the transfer of DNA: and conjugation, where the genetic material is transferred via a small tube from one bacteria to another [17].

Conjugation can be mediated by plasmids [15]. Plasmid-mediated resistance is one of the most common ways of acquiring resistance and can be considered platforms on which genes and genetic elements are arranged in a circular or linear form [18] and can replicate independently of the host chromosome [15, 19]. The emergence of pathogens being resistant to a multitude of different antimicrobials and their uncontrolled spread through clinical settings is believed to be due to their association with specific conjugative plasmids [20]. The basic component of all plasmids, the minimal replicon, consists of the origin, a region for initiation of plasmid replication, and an initiator gene. While plasmids encode their replication initiation, they use the replication machinery for DNA synthesis of their host. This exploitation is one of the factors that limit the host range of plasmids, but also factors for whether the plasmid is transmitted by conjugation or mobilization determine the host range [21].

Mobile genetic elements (MGEs) are pieces of DNA that promote the mobility of genetic material within and between bacteria. There are many different kinds of MGEs, such as insertion sequences, transposons, plasmids, and phage plasmids [22]. Transposable elements carry their own set of genes, including both ARGs and genes involved in the translocation. Integrons carry the machinery needed to perform site-specific recombination, passenger genes such as ARGs, and a promotor, allowing the integron to be expressed in a host [19, 23, 24]. Today, many acquired ARGs have been discovered (Table 1.1), and their widespread dissemination is largely due to the capability of MGEs to jump between different hosts and plasmids and taking the ARGs with them. There have been increased efforts to develop strategies that

can mitigate the actions of MGEs, such as considering the MGEs as pollutants and invasive species [22].

## 1.2 Dissemination of AMR

While the mechanism of developing or requiring resistance genes was described in the previous section, there is typically a fitness cost of keeping ARGs and mobile genetic elements around. However, the promotion of mobilization and maintenance of ARGs within a host typically happens under selection pressure [25]. Therefore, it is crucial to recognize that the AMR crisis goes beyond just studying the genetic functions and that characterizing other drivers of AMR is just as important. Drivers, such as trade and travel routes, environmental changes, and populations, have already been shown to influence the dissemination [26]. The increased accumulation of ARGs in various environments can, to some degree, be attributed to human activities, such as the overuse of antibiotics in clinical settings and the practice of adding antibiotics in livestock feed to promote growth and prevent disease. Spillover from these activities into the environment results in a buildup of resistance-carrying bacteria in water, soil, or air [25, 27].

Notably, the use of antimicrobials in agriculture has been associated with the prevalence of resistance in both pathogenic and commensal bacteria [28, 29]. Studies have even shown that, in some cases, resistance remains in an environment years after removing the selective pressure and that using different classes of antimicrobials increases the risk of co-occurrence of various resistance genes [30, 31, 32]. Fleming already warned about the dangers of inappropriate use of antibiotics leading to more severe forms of the disease in an interview in 1945 [33].

Based on these observations, the World Health Organization (WHO) developed a ranking system of antimicrobials to reflect their importance in human medicine (Table 1.1). These rankings can then be used to establish resistance risks associated with the use of different antimicrobials in animals [34, 35]. In the remaining part of this chapter, the distribution of resistance around the globe will be introduced, and, finally, present the One Health perspective on AMR.

### 1.2.1 Emergence and distribution of AMR in clinical settings

Most antibiotics were, in the beginning, used in hospitals, which is also where the first strains of drug-resistant pathogens emerged. One example is the emergence of the penicillin-resistant *Staphylococcus aureus* in London hospitals shortly after the introduction of penicillin in the 1940s [36]. Multidrug-resistant (MDR) bacteria

appeared in the late 1950s to early 1960s but continued to re-emerge and cause infections that are difficult to treat, sometimes with no successful outcome [37]. In the case of treating MDR pathogens, antimicrobials that were discontinued have been taken into rotation again as a last-resort treatment. One last-resort antibiotic is colistin or polymyxin E, which was first discovered in 1947 (Table 1.1) and used in human medicine. Due to neurotoxic and nephrotoxic side effects, colistin was abandoned in the 1980s but was taken into use again in the early 2000s [38, 39]. WHO has classified colistin as a critically important antibiotic (Table 1.1). However, due to a history of colistin used in veterinary medicine and as a growth promoter in pigs and poultry [40], plasmid-mediated resistance to colistin began to emerge, and in 2015 the first resistance gene was reported, namely the mobilized colistin resistance gene *mcr-1* [41]. Since then, multiple *mcr* genes have been reported and shown to have spread worldwide [42, 40, 43]. The newest member of the family of *mcr* genes is *mcr*-10 was reported in 2020 [44], suggesting that it is only a matter of time before the next gene in the family is discovered.

The high consumption level of antimicrobials in clinical settings and hospitals is also reflected in the waste that causes a spillover into the environment [45]. The hospital-generated waste accumulates antibiotics, exposing bacteria to the antibiotics and creating selective pressure for bacteria carrying resistance genes [46, 45]. There are examples of several antibiotics that could be detected in abundances strong enough to penetrate other environments, e.g., soil, from sewage treatment plants, despite a dilution of hospital sewage with communal sewage and wastewater treatment [46].

### 1.2.2 Environmental sources of AMR

As already highlighted in the previous sections, settings where historically the use of antimicrobials has been high have created hotspots of AMR that can cause ARGs to move into other environments. Examples include: ARGs conferring resistance to beta-lactams that have been detected in farm animals; tetracycline resistance in farm manure, wastewater treatment plants, and aquatic and soil environments; and resistance to sulfonamides (folic acid antagonists) in agricultural areas, wastewater treatment plants, rivers, and oceans [27].

The route of resistance-carrying bacteria might start with the emergence of novel resistance genes, then the genes become mobilized and move into various environments, and are ransferred into human pathogens [25]. The steps might not happen in the order written but highlights how resistance factors move. The One Health approach of recognizing the interplay between human health, the environment, and animals, is especially fitting for the issue of antimicrobial resistance [47].

**Table 1.1.** Antimicrobial classes with year of first reported, medically introduction, and first resistance reported together with number of ARGs in ResFinder[48] (version 20200125), and WHO Classification of Antimicrobials Important for Human Medicine. Acronyms: CI: Critically Important. HI: Highly Important. I: Important.

| **Antimicrobial class** | **Year of discovery reported** | **Year of introduction** | **Year of first resistance reported** | **Number of ARGs** | **WHO classification [34]** |
|---|---|---|---|---|---|
| Aminoglycoside | 1944 [5] | 1946 [5] | 1946 [7] | 262 | CI |
| Beta-lactam | 1929 [7, 5] | 1938 [7] | 1945 [7] | 2,001 | CI |
| Fluoroquinolone | 1962 [5] | 1962 [5] | 1983[49] | 3 | CI |
| Folate pathway antagonist | 1908[1][50] | 1945[1][50, 5] | 1955 [51] | 160 | HI |
| Fosfomycin | 1969 [5] | 1971 [5] | 1976[52] | 40 | CI |
| Glycopeptide | 1954 [5] | 1958 [5] | 1960 [7] | 44 | CI |
| Lincosamide | 1962 [5] | 1963 [5] | 1955 [7] | 115 | HI |
| Macrolide | 1952 [5] | 1952 [5] | 1955 [7] | 127 | CI |
| Nitroimidazole | 1959[2][5] | 1960 [5] | 1942 [7] | 14 | I |
| Oxazolidinone | 1987 [5] | 2000 [5] | 2001 [7] | 28 | CI |
| Phenicol | 1947 [5] | 1949 [5] | 1950 [7] | 67 | HI |
| Pleuromutilin | 1951 [5] | 2007 [5] | 2006[53] | 21 | I |
| Polymyxin | 1950 [5] | 1959 [5] | 2016[41] | 53 | CI |
| Quinolone | 1961 [7] | 1968 [7] | 1968 [7] | 125 | CI |
| Rifampicin | 1957 [7] | 1958 [7] | 1962 [7] | 10 | CI |
| Steroid antibacterial | 1958[3][5] | 1962 [5] | 1971[3][54] | 3 | HI |
| Streptogramins | 1953[4][5] | 1965 [5] | 1964 [7] | 128 | HI |
| Tetracycline | 1948 [5] | 1952 [5] | 1950 [7] | 150 | HI |

[1] for sulfones.
[2] for sulfadrugs.
[3] for fusidic acid.
[4] for streptogramin B.

### 1.2.3 AMR, One Health, and the Sustainable Development Goals

Applying the concepts of the One Health approach to tackle AMR will require that all three domains (people, animals, and environments, see Figure 1.1) are well understood in their contribution to the development and spread of resistance [47]. Recording new ARGs as they emerge is a crucial step in establishing comprehensive monitoring across the three One Health pillars, as well as utilizing the many sequencing efforts happening across the globe. There are already existing surveillance programs in place, such as the Danish Integrated Resistance Monitoring Programme (DANMAP) started in 1995 that monitors AMR in clinical isolates, livestock, and along the food chain [55, 56], or the WHO Global Antimicrobial Resistance and Use Surveillance System (GLASS) that uses patient samples to survey the prevalence of pathogens common in human infections [57, 58].

AMR not only threatens health but is also a societal challenge considering the impact on the Sustainable Development Goals (SDGs). The UN developed the SDGs in 2015 to achieve a "better and more sustainable future for all" [59]. Besides the effect on SDG3 about good health and well-being, AMR is also linked to SDG1 on no poverty and SDG2 about zero hunger, as the increasing global population also increases the requirements of food production, increasing the use of antimicrobials as supplements in livestock feed [31]. Climate change (SDG13) and AMR are also linked; for example, a study has shown that AMR increases with a rise in temperature [60]. AMR is also linked to several other SDGs (see Gajdács et al. [61]), but hopefully, these examples illustrate why AMR has many ramifications beyond global health and how surveillance on the distribution of ARGs is needed on a global scale. Especially understanding where ARGs come from, how they move, and their abundances are important to make effective regulations on how to solve the AMR threat.

## 1.3 Problem statement

Viewing AMR as a global issue will require new multidisciplinary approaches, such as establishing how to characterize the distribution of AMR in humans, animals, and environments. The next two chapters will introduce how the distribution of ARGs can be surveyed by reusing publicly available sequencing datasets to answer the three main questions of this PhD project: what, where, and how much?

Creating a workflow for retrieving, processing, and analyzing sequencing reads from public data repositories to characterize the global distribution of AMR is the primary goal of this thesis. The studies conducted during this PhD focused on the following objectives:

1. Develop a pipeline for retrieving metagenomic samples from the European Nucleotide Archive (ENA), quality checking and trimming sequencing reads, and aligning the trimmed reads against reference sequences from two databases: ResFinder [48], consisting of acquired ARGs, and Silva [62], which contains ribosomal rRNA (rRNA) sequences.

2. Create a MySQL database for storing sampling metadata and output of the alignment procedure for more accessible analysis and data sharing.

3. Using techniques from compositional data analysis and genomics, profile the emergence and distribution of ARG abundances in different sampling locations, years, and sources to identify novel patterns.

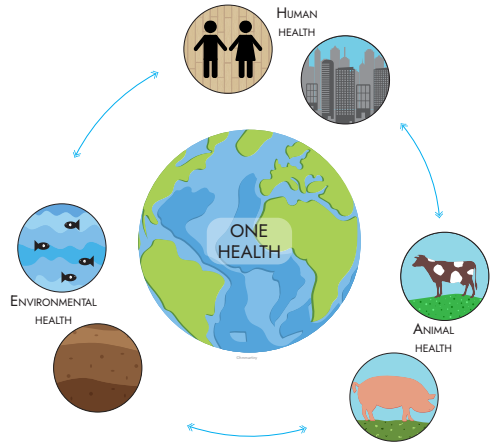



**Figure 1.1.** The One Health approach focuses on how the human health is connected to the health of both animals and the environment. In terms of AMR, there are clear links to each of the three pillars, and the study of all three domains is needed to understand how AMR emerges, evolves, and disseminates.

CHAPTER 2

# Epidemiological surveillance with (meta)genomics

Epidemiology is a field of study focusing on the distribution of disease in populations and which factors, or determinants, are involved in this distribution [63]. In terms of infectious diseases, determining the origin and monitoring the spread of pathogens can provide better and faster response and, in the end, control the disease. With the technical advances and decreasing cost of sequencing technologies, characterizing disease-causing agents at a genetic level has become more feasible and are routinely done as part of clinical diagnostics. Using genomic DNA sequences can give a faster and more reliable identification of the pathogen, identify phylogenetic relationships, and characterize of genomic traits relevant to epidemiological studies [64, 26]. Implementing a surveillance system of AMR embracing the One Health approach have the potential to elucidate how ARGs move across environments, hosts, and geographical borders.

Genomics, as a field, focuses on studying the genome of a single organism and was coined as a term in 1986 during a meeting about starting the project about sequencing the human genome [65]. A genomic workflow begins with the cultivation of a microbe. DNA is then extracted and amplified from the isolated microbe. Thirdly, the sequencing library is created, and finally, the generation of reads using a sequencing machine so that, in the end, the species' genome can be constructed (Figure 2.1). However, genomics is only done on organisms that can be cultivated, which does not capture the full diversity of microorganisms [66]. In metagenomic studies, all genetic material is recovered from an environmental sample and sequenced, which omits the cultivation step (Figure 2.1), i.e., a genetic snapshot of the environment is created.

To profile the composition of a microbial community, the sample can be sequenced using different methods, such as Whole-genome shotgun (WGS) sequencing or targeted

16S rRNA sequencing. WGS sequencing randomly breaks down long DNA molecules into smaller fragments that are sequenced [67]. Targeted 16S rRNA sequencing takes advantage of the universal presence of the 16S rRNA gene in prokaryotes. The 16S rRNA gene is about 1550 basepair (bp) long and consists of variable and conserved regions. Primers are designed for the conserved regions, and the variable regions are used to distinguish between the different bacteria [68]. Depending on the end goal, there are both advantages and disadvantages for doing shotgun or targeted 16S rRNA sequencing. Profiling of 16S rRNA does not need as high sequencing depth as WGS because 16S sequences are well-defined in reference databases; for example, GenBank has over 167,000 nucleotide 16S sequences deposited (retrieved 22-09-2022). WGS sequencing, instead, can capture all kinds of genetic sequences and not only 16S genes, which allows for the characterization of eukaryotes and viruses and functional analyses of, e.g., the distribution of ARGs. Because of the broader genetic information offered by WGS, it can be complex to analyze shotgun sequencing data [69].

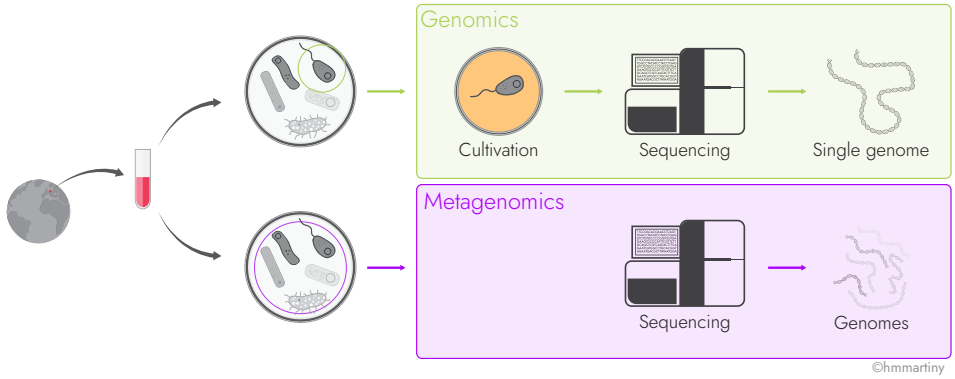



**Figure 2.1.** A simplified overview of genomic and metagenomic workflows to illustrate the difference between them, where the goal of genomics is to obtain the genome of interest and the metagenomic aim is to capture all genomes in the sample.

## 2.1 A brief history of DNA sequencing

While Alexander Fleming's discovery of penicillin kickstarted modern medicine, the starting point of DNA sequencing was when James Watson and Francis Crick solved the three-dimensional structure of DNA in 1953 [70] using crystallographic data provided by Rosalind Franklin and Maurice Wilkins [71]. It took almost another 25 years before the first generation of sequencing began with the Sanger method published in 1977, but since then, the field has seen many new sequencing platforms come and go, all in the quest to produce more accurate, longer, and cheaper DNA sequences.

### 2.1.1 First generation sequencing

The aforementioned Sanger sequencing [72] was a groundbreaking technique because of the chain-termination method. This technique involves adding chain-terminating dideoxynucleotides that have been, in the early days, radioactively labeled, to perform sequencing of a DNA fragment. After the addition of the labels, the fragments were then run through gel electrophoresis to separate based on fragment size and, lastly, analyzed to determine the sequence. In a later version of Sanger sequencing, the labeling was done using fluorescently instead of radioactively [73, 74]. The accuracy, robustness, and ease of use led Sanger sequencing to become the most common sequencing technology [73]. The first generation of automated DNA sequencing machines used the Sanger method [75], which generated sequencing reads of a little less than one kilobase (kb) in length [76].

### 2.1.2 Second generation sequencing

When the project to sequence the human genome began in the 1990s [77], the need for higher quantities and longer sequencing reads sped up the development of High-throughput sequencing (HTS) platforms. The pyrosequencing method [78] used in the 454 machines (later Roche) made it possible to do mass parallelization to obtain larger amounts of DNA in just one sequencing run. The 454 machines could produce reads around 400-500 bp long. Following the success of 454, new parallel sequencing techniques were developed that made it possible for individual research laboratories to build their sequencing capacities [79]. These new innovative sequencing technologies are called second generation or Next-Generation Sequencing (NGS) technologies.

Table 2.1 lists examples of NGS platforms, where the most notable of them is the Illumina machines, which today are the most common sequencing machine in use. Illumina machines produce short reads of up to about 300 bps in length and support paired-end sequencing, which means that the DNA fragment is sequenced from both ends [76]. Another NGS approach is the Ion Torrent sequencing by ThermoFisher, which instead of using fluorescence or luminescence detection, measures the difference in pH to determine the sequences.

All three sequencing platforms (454, Illumina, Ion Torrent) do sequencing-by-synthesis using Polymerase chain reaction (PCR), but an alternative approach was introduced with the SOLiD system from Applied Biosystems, later acquired by ThermoFisher (Table 2.1). SOLiD stands for sequencing by oligonucleotide ligation and detection using DNA ligase [73]. The length of the sequenced reads in a SOLiD platform is generally relatively short, at only about 35 bps, but the output per sequencing run was in the order of gigabases [80]. This large output made the SOLiD platform cost-effective

compared to the other NGS platforms [73].

The revolution of genome sequencing platforms has had a drastic impact on the ease and cost of performing sequencing experiments, as the number of bases that can be sequenced per unit cost has been growing at an exponential rate. Stein [81] showed in 2010 that the cost of genome sequencing has been decreasing faster than the cost of disk storage; in other words, it is cheaper to sequence a genome than to store the output.

### 2.1.3 Third generation sequencing

The second generation sequencing machines are known as short read technologies, as the read lengths are still in the bp range, up to a few 100 bps. The newest iteration of platforms, the third generation, uses long-read technologies that generate sequences of more than 10kb in length directly from the native DNA (Table 2.1). Pacific Biosciences (PacBio) and Oxford Nanopore Technology are the two main companies with third generation platforms on the market.

PacBio does single-molecule real-time sequencing by sequencing in thousands of reaction wells, in which the DNA template is added to the bottom. Four differently fluorescently labeled deoxyribonucleoside triphosphates (dNTPs) are added, and using a DNA polymerase, the sequence is called by detecting the small signal emitted by incorporating the dNTPs [82]. The resulting sequence fragments might contain random errors, but the errors are randomly distributed, and bias is reduced due to skipping the PCR amplification step [83]. The Oxford Nanopore Technology platforms (Table 2.1) pass a single-stranded DNA molecule through a specific protein pore, a nanopore, where the DNA fragment is pulled through the pore one base at a time. The sequence is then determined by detecting changes in an electrical current through the nanopore [84]. This kind of sequencing did suffer in the beginning from much higher error rates, around 15% [83], but is now down to less than 1% [85]. The benefit of nanopore sequencing is the very long reads, as detecting structural variants is much less cumbersome. There are only a few hundred reads to compare as opposed to a million reads from the earlier NGS platforms (Table 2.1) [84].

| Company | Platform | Maximum Output/Run | Maximum Reads/Run | Maximum Read Length |
|---|---|---|---|---|
| Roche (454)[1] | GS FLX | 500 Mb | $1 \cdot 10^6$ | 400 bp |
| Illumina | iSeq 100 | 1.2 Gb | $4 \cdot 10^6$ | 150 bp |
| | MiniSeq | 7.5 Gb | $25 \cdot 10^6$ | 150 bp |
| | MiSeq | 15 Gb | $25 \cdot 10^6$ | 300 bp |
| | NextSeq 550 | 120 Gb | $400 \cdot 10^6$ | 150 bp |
| | NextSeq 1000/2000 | 360 Gb | $1.2 \cdot 10^9$ | 150 bp |
| | NovaSeq 6000 | 6000 Gb | $20 \cdot 10^9$ | 250 bp |
| ThermoFisher | Ion 510 Chip | 1 Gb | $3 \cdot 10^6$ | 400 bp |
| | Ion 520 Chip | 2 Gb | $6 \cdot 10^6$ | 600 bp |
| | Ion 530 Chip | 8 Gb | $20 \cdot 10^6$ | 600 bp |
| | Ion 540 Chip | 30 Gb | $80 \cdot 10^6$ | 200 bp |
| | Ion 550 Chip | 50 Gb | $130 \cdot 10^6$ | 200 bp |
| ThermoFisher[2] | ABI SOLiD | 4 Gb | $1 \cdot 10^9$ | 35 bp |
| PacBio[3] | Sequel | 75 Gb | $5 \cdot 10^6$ | 30 kb |
| | Sequel II | 600 Gb | $4 \cdot 10^6$ | 30 kb |
| | Sequel Iie | 600 Gb | $4 \cdot 10^6$ | 30 kb |
| Oxford Nanopore | MinION | 50 Gb | >99% accuracy[4] | 4 Mb[5] |
| | MinION Mk1C | 50 Gb | >99% accuracy[4] | 4 Mb |
| | GridION | 250 Gb | >99% accuracy[4] | 4 Mb |
| | P2 Solo | 580 Gb | >99% accuracy[4] | 4 Mb |
| | P2 | 580 Gb | >99% accuracy[4] | 4 Mb |
| | PromethION 24 | 7 Tb | >99% accuracy[4] | 4 Mb |
| | PromethION 48 | 14 Tb | >99% accuracy[4] | 4 Mb |

[1] Roche has discontinued the 454 platforms, so data was retrieved from Voelkerding et al. [86].
[2] Formerly Applied Biosystems. Data for ABI SOLiD retrieved from Mardis [87].
[3] PacBio data retrieved from Hu et al. [74].
[4] Oxford Nanopore platforms reads the length of the fragment, no overlapping reads are generated.
[5] Highest reported length according to the Oxford Nanopore website per 22-09-2022.

**Table 2.1.** Examples of NGS platforms. All URLs were accessed 23-09-2022 to retrieve sequencing information, unless otherwise stated: `www.illumina.com`; `www.thermofisher.com`; `www.pacb.com`; `nanoporetech.com`. Acronyms: bp, basepair; kb, kilobases; Mb, megabases; Gb, gigabases; Tb, terabases.

## 2.2 The wealth of sequencing data

With the increased output and lowered cost of NGS technologies (Table 2.1), modern biology now utilizes sequence data routinely in research projects. As part of a scientific publication, it is highly encouraged to share the raw sequencing reads in public archives, resulting in an unprecedented amount of sequencing data being available online. Members of the International Nucleotide Sequence Database Collaboration (INSDC; `www.insdc.org`) have built the core infrastructure for sharing nucleotide sequencing data and the associated sample information, or metadata, in repositories that are publicly available. This collaboration is between three members that represent different parts of the world and facilitates a daily data exchange: the DNA Data Bank of Japan (DDBJ, `www.ddbj.nig.ac.jp`), the European Nucleotide Archive (ENA, `www.ebi.ac.uk/ena`), and GenBank (`www.ncbi.nlm.nih.gov/genbank`) [88]. Figure 2.2 shows the staggering amount of sequencing datasets, or runs, deposited on ENA since 2010. With more than 20 million sequencing runs encompassing more than 40 petabytes of data, it is safe to say that an enormous number of reads have been generated so far. These numbers are expected to continue to increase, as the current estimates say that the number of datasets doubles about every 22 months, and the disk storage doubles every 31 months.

### 2.2.1 *De-novo* assembly of NGS reads

Assembly of reads is the process of figuring out where the, typically, short reads overlap and reconstructing the original genomes based on these overlaps. This recovery can be done by creating assemblies either by using reference-based methods, *de novo* or a combination of both with reference-guided *de novo* assembly. However, if the microbial sample contains genomes not part of reference databases, reference-based methods cannot recover unknown genomes. *De novo* assembly has aptly been described as putting the pieces of a jigsaw puzzle together without knowing what the whole picture looks like [89]. It is a computationally expensive process, as it begins with assembling the NGS reads into long sequences, or contigs, by finding the overlapping regions. Then the contigs are scaffolded, ordering of contigs by using linkage information, such as mate pairs. Finally, the gaps between contigs in the scaffolds are filled out by independent reads. Scaffolding and gap-filling are sometimes done repeatedly in an iterative process to improve the overall assembly [90]. Many de novo assembly tools use *de Bruijn graphs* to build the contigs, such as metaSPADes [91]. The basic principle of building *de Bruijn* graphs is to construct k-mers from the reads and connect overlapping k-mers. However, efficiently storing the k-mers in memory is difficult as the usage increases with the length of the reads, coverage, and the complexity of the metagenomic sample, as illustrated by the approximation of gigabytes

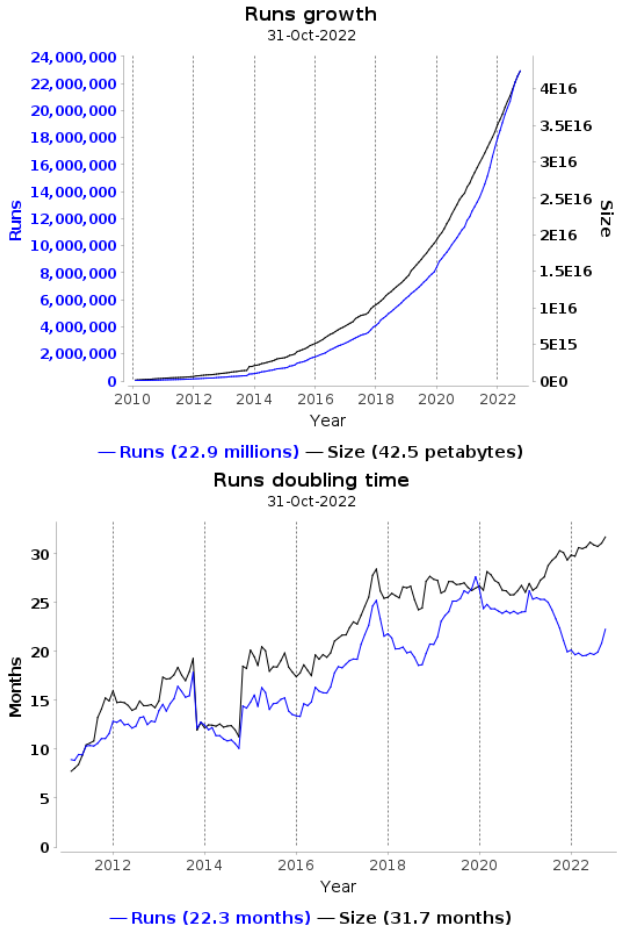



**Figure 2.2.** The growth of sequencing runs available on European Nucleotide Archive (ENA). Top: Count of sequencing datasets (runs) uploaded (left y-axis) and the disk storage size required to store all these datasets (right y-axis). Bottom: The doubling time of sequencing reads available and data size. The two figures were retrieved from `www.ebi.ac.uk/ena/browser/about/statistics`. Note that the graphs include all NGS datasets, not only metagenomic datasets.

needed of RAM to store a table of k-mers: $2(k+1)G$ gigabytes to store k-mers of size $k$ of and genome size $G$ [92]. Next, the question is what information can be extracted from the assembled contigs. Tools such as Kraken [93] or QIIME [94] can be used to assign taxonomic labels to genomes and analyze the microbial community, MGEs can be detected with MobileElementFinder [24], phylogenetic clustering with PhyloPhlAn [95]. These tools are just some examples of what can be used to study

assembled genomes.

### 2.2.2 Mapping of NGS reads

Instead of assembling reads, an alternative route is to characterize the unassembled sequence reads by mapping the reads against reference sequences. As the name suggests, using reference-based mapping happens by comparing the reads to reference genomes and is a relatively computationally inexpensive process, which can be pretty effective in, e.g., benchmarking studies in which the microbial composition is known [90]. Existing tools such as Bowtie 2 [96], BWA-MEM [97], and KMA [98] map the reads against entire databases of reference sequences. There are many different reference databases available, such as Silva [62] to profile the 16S rRNA genes or ResFinder [48] or CARD [99] to obtain ARG reference sequences. These methods of mapping reads to references scale well in their efficiency to complex datasets since each read is considered indecently to the rest but can still be slow if the reference database is large [69]. The result of mapping and aligning reads is typically a matrix with the count of reads matched to reference sequences if not doing assemblies.

An abundance matrix consists of multiple samples with their read counts that can be used to explore microbial diversity through different ecological indices. Alpha diversity measures the diversity of species or genes within a sample, whereas beta diversity quantifies the variations between samples [100]. There are also methods to test whether the abundances differ between categories [101], study the relationships by associating abundances of different genes or taxa [102, 103] or explore the beta-diversity with principal component analysis [104].

Although the process of assigning reads to references is less complex than doing assemblies, analyzing the abundance matrix to describe the community in the sample has a few key challenges that need to be addressed: not all species are often observed so the abundance table might be quite sparse; the counts depends on what is in the reference databases, so there might be a large unmapped part of the sample; and, the total number of reads depends mainly on the capacity of the platform, which might make the output of two sequencing runs on the same sample differ [69, 100]. The latter makes the data compositional and is the focus of the next chapter.

### 2.2.3 Epidemiological uses of assemblies and read mappings

In the epidemiological setting, recovering genomes is relevant for investigating outbreaks. One example is tracing patient-to-patient transmissions of methicillin-resistant *Staphylococcus aureus* (MRSA) in intensive care units by comparing the genetic sim-

ilarities of the isolated strains [67]. In retrospective studies, genetic variations are often used to find the origin and construct transmission networks. Another recent example, but not AMR related, is finding where the novel severe acute respiratory syndrome coronavirus (SARS-CoV-2) came from in 2019. In the study by Worobey et al. [105], they investigated where the coronavirus that caused a worldwide pandemic came from. By analyzing various environmental sources and modeling the likelihood of origin, they showed that the epicenter most likely was a seafood market in Wuhan, China, where SARS-CoV-2 jumped from animals into humans.

From the metagenomic perspective, there are many examples of studies assembling genomes from metagenomic samples, otherwise known as Metagenome assembled genomes (MAGs). With 38 metagenomic samples from activated sludge reactors treating antibiotic production wastewater, Zhao et al. [106] assembled 689 genomes from 2245 million paired-end reads to investigate the prevalence and mobility of ARGs in hosts. The authors found that ARGs are likely to be mobilized under high antibiotic selection pressures through a co-occurrence analysis of ARGs and MGEs in their MAGs. A characterization of MAGs recovered from fecal pig samples by Holman et al. [107] revealed that MAGs assigned to commensals in the gut were carrying not only ARGs but also specific enzymes involved in metabolism. They suggested that the functional identification likely could explain why macrolide and tetracycline resistance persisted in the gut in the absence of antimicrobial selective pressure. The Tara Oceans expedition [108] collected samples from the world oceans, where Cuadrat et al. [109] used the assembled metagenomes to explore the ARG distribution. Some of their results showed that specific ARGs are more abundant in coastal environments, which they hypothesize is due to the inflow of antibiotic-resistant strains by wastewater. From the intestines of deep-sea fish in the Atlantic Ocean, almost all the MAGs reconstructed by Collins et al. [110] lacked acquired antimicrobial resistance genes. This snapshot of the microbiome at deep-sea levels suggested that this environment remains largely unaffected by human activities.

Abundance analyses have revealed several associations of the resistome in various environments. For example, ARG transmission in soil environments was assessed by Knapp et al. [111], which resulted in that eight different ARGs being positively associated with copper levels in the soil. In another study by Wang et al. [112], they showed that the spread of ARGs across a soil-root continuum was a continuous stream by evaluating the attributions of environmental sources on the resistome. Urban wastewater environments have also gained a lot of attention for AMR surveillance, mainly due to the fact that globally an increased number of people are connected to sewage treatment plants. Some of the main benefits of sewage surveillance are that they cover large communities, a sampling procedure is straightforward to implement, sequencing and downstream analyses are easily standardized, and there are no ethical concerns in doing so [113]. Multiple studies have shown that higher abundances of ARGs exist in sewage [26, 114, 115, 116], but there are systematic differences in resistomes across world regions, likely due to sanitation and health factors [26].

### 2.2.4 Reusability of sequencing data

The growing number of sequencing reads offers many new possibilities to catalog and explore microbiomes in different environments, conditions, and geographical locations (Figure 2.2). Four principles were formulated to encourage the practice of sharing and building tools: Findability, Accessibility, Interoperability, and Reusability. Better known as the FAIR principles, they were designed to serve as guidelines for good data management [117]. The policy created by ENA and the other members of INSDC was actually used as the template for the FAIR principles [88], and arguably a lot of the reference sequence databases that exist also embrace these principles, such as ResFinder [48] and Silva [62].

However, working with terabytes, or even petabytes, of data, is not feasible for many researchers simply due to the lack of resources needed to handle such large amounts. Therefore, there is also a need for standardized pipelines that enable the sharing of assemblies, read counts, and other results of downstream analyses of DNA sequencing data that are easily accessible to the larger scientific community. Parallels can be drawn to experimental protocols, where each step is carefully documented, e.g., with concentrations and temperatures. The code, software, software versions, parameter values, and other details in a computational workflow have typically been omitted from published articles, but for the results to be reproducible, these details need to be shared [118]. Most journals nowadays require these for the manuscript to be published, but instead of just documenting it as part of a method section, several workflow managers are available that simplify the process of reimplementing. Two of them are Snakemake [119, 120] and Nextflow [121], which both uses a domain-specific language that improves readability and provides statements and declarations for controlling input files, variables, commands, and output [120]. Recently, a protocol for how to carry out metagenome analyses was published in Nature, in which the authors developed several easy-to-use scripts for the steps in the protocol [122]. Depositing data that is not the raw sequencing data, e.g., the output of an analysis, can be stored on other data repositories that provide a permanent digital object identifier, such as Zenodo (`zenodo.org`) [123]. From writing this, it could sound like the task of reusing the publicly available sequencing data might be straightforward; however, there are several challenges that one might face beyond just handling the sheer volume of data.

The information on the sample's origin, such as the sampling source, location, and date, is essential to put the results into context, but how this metadata is written is often filled with many errors. When entering metadata for uploading to a repository, e.g., ENA, there are fields with a restricted vocabulary of values to choose from, but others are free text giving ample risks for mistakes to be entered. While there today are several checklists in place to ensure minimum information about a sample is shared, these checks have not been in place from the beginning. Even if still encountering mistakes, updating the existing metadata records is currently only available for

the submitter of the data, not other users. Arguably, it should be a major point of a data submitter to ensure that their metadata is as clean as possible. More importantly, it should be possible for other users to flag incorrect data. Curating metadata can be a time-consuming and manual process but necessary to allow using the data to its full potential [124].

Creating a global overview of AMR also requires that the data be as much unbiased as possible. Historically it has been more expensive to perform sequencing experiments, as observed in public data repositories when looking at the sampling origins. For example, Abdill et al. [125] noted that human microbiome samples tended to come from Europe, USA, and Canada and fewer from central and southern Asian countries. The skewed distribution of sampling origins is not limited to only human samples but demonstrates well how the underrepresentation of specific world regions will restrict the interpretation of results from global analyses.

The rapid innovation in NGS platforms has also spawned a variety of experimental procedures for obtaining DNA sequences, and each iteration of the platforms has aimed to improve both read length and output size (Table 2.1). Illumina machines are the most dominant on the market, but that does not mean that there are no samples of older date and sequenced on the first machines. The variety of procedures and platforms also creates a non-removable bias in the available sequencing datasets, which should be considered when retrieving reads from different projects.

Nonetheless, there is still much to be gained from repurposing existing NGS datasets if the issues above are either handled in an appropriate manner or at least acknowledged when interpreting results. Especially the bias introduced by uneven sampling distributions will hopefully be less as the NGS platforms continue to become cheaper and better. Figure 2.3 shows how a pipeline can accomplish the reuse of a metagenomic dataset from ENA to characterize the abundance of ARGs. The pipeline takes the reads, does trimming, quality checking, and alignment to chosen reference databases, and stores the mapping results in a database. At the same time, a mix of automated and manual metadata curation is being done to conform the needed information into the same format (Figure 2.3a). The database then contains the output of both the processing of the reads and the curated metadata, where the data can then be extracted and analyzed to study the global resistome and microbiome (Figure 2.3b). The existing surveillance efforts should still continue, such as DANMAP and GLASS, as the proposed pipeline is not a replacement but should instead be considered as a supplement to broaden the understanding of AMR.

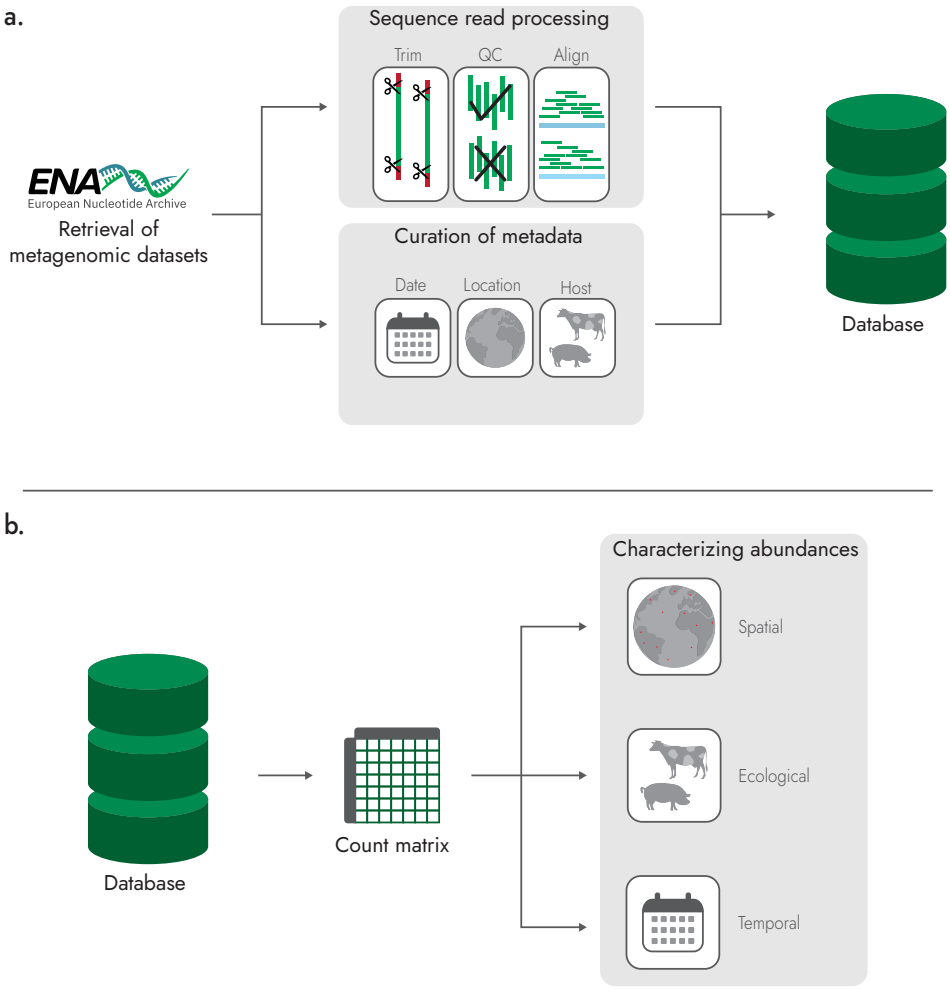



**Figure 2.3.** The proposed pipeline for retrieving and analyzing metagenomic datasets. **a**. The uniform processing of sequencing reads and metadata curation is stored in a database. **b**. The abundance matrix of counts and the associated metadata can then be analyzed to identify data structures across spatial, ecological, and temporal spaces.

CHAPTER 3

# Analyzing metagenomes using Compositional Methods

The sequencing reads produced by NGS technologies have many different uses, all with the goal of assessing the biological content of a sample. One of the use cases is to quantify the abundance of microbes or genes in the sample by aligning the reads to reference sequences. However, the different NGS techniques use different sample preparation and assay protocols, causing bias to be introduced into the number of reads generated and, subsequently, downstream analyses. Therefore, the total number of reads produced depends not only on the sample but also on the capacity of the sequencing platform. If one were to have a set of reads generated from two samples from the same environment, the two read distributions might be very different [126, 127].

Compositional Data (CoDa) refers to a set of vectors that consists of positive numbers that are part of a whole. That can easily be translated into the count matrix of reads aligned to different references. The read counts can only be interpreted by considering them as the relative proportion of the total number of reads. In 1897, Karl Pearson noted that the analysis of relative counts, or ratios, can produce spurious correlations and warned against attempts to interpret that [128]. Even so, it was not until the work of John Aitchison in 1982 that CoDa became its own field of research [129]. This chapter will not be a comprehensive review of CoDa but will focus on how CoDa methods are applicable for working with metagenomic samples. The mathematical notation in the accompanying sections follows the notation of Pawlowsky-Glahn et al. [130].

A composition is a vector $x = [x_1, x_2, , x_D]$ of $D$ relative numbers, where $D$ is the number of components, i.e., the number of genes, and each $x_i \in \mathbb{R}^D$ is the count of reads assigned to gene $i$. The counts can only be considered as carrying relative information as a count is only interpretable in relation to all the other counts. Due to this relativeness, the compositional data is constrained to be on the simplex $\mathcal{S}$, which is the sample space with a dimension size of one less than the number of components (Figure 3.1). The compositional data in the simplex can be normalized by closing the composition to express the counts proportionally to a specific constant, i.e., the counts of reads are normalized to be like percentages by using definition 3.1:

**Closure**

**Definition 3.1.** For a compositional vector $x$ of $D$ parts, closure is defined as

$$\mathcal{C}(x) = \frac{\kappa}{\sum_{i=1}^{D} x_i} \cdot x$$

where $\kappa > 0$ is the closure constant.

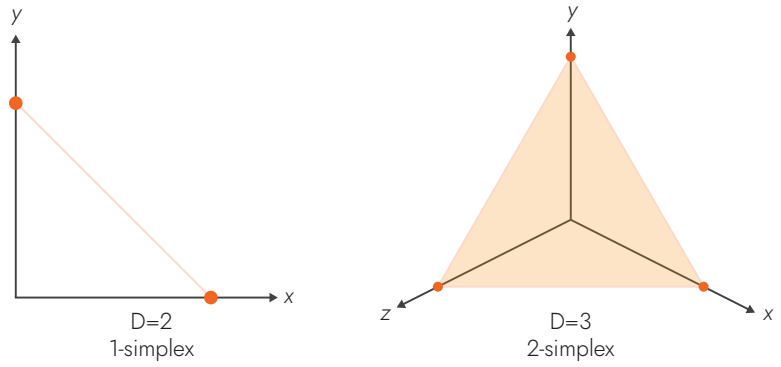



**Figure 3.1.** Example of 1-simplex and 2-simplex as graph representations.

The traditional statistical methods do not apply in the compositional sample space because they operate on real numbers. Aitchison did define geometrical functions [129], but using these functions makes the calculations much more cumbersome. Instead, the starting point of any statistical analysis of compositional counts should be to do a log-transformation of count ratios. Transforming the relative counts into a log-space makes the data symmetric and linearly related and converts them into a log-ratio space of real numbers [126]. Another way to put it is that the ratios are now relative to other features in the data directly related to the sampling origin and no longer on the NGS platform used to produce the reads. The Additive log-ratio (ALR) and the Centered log-ratio (CLR) are often used and can be found in definitions 3.2

and 3.3.

**Additive log-ratio transformation**

**Definition 3.2.** The ALR transformation uses one part as the reference and transform the composition $x \in \mathcal{S}^D$ into $\mathbb{R}^{D-1}$ by

$$\text{ALR}(x) = \left[ \ln \frac{x_1}{x_D}, \ln \frac{x_2}{x_D}, ..., \ln \frac{x_{D-1}}{x_D} \right]$$

**Centered log-ratio transformation**

**Definition 3.3.** Instead of using one of the components as a reference, the CLR transformation uses the geometric mean $g_m(x)$ of the sample as the reference:

$$\text{CLR}(x) = \left[ \ln \frac{x_1}{g_m(x)}, \ln \frac{x_2}{g_m(x)}, ..., \ln \frac{x_D}{g_m(x)} \right]$$

where $g_m(x)$ is

$$g_m(x) = \left( \prod_{i=1}^{D} x_i \right)^{1/D} = \exp \left( \frac{1}{D} \sum_{i=1}^{D} \ln x_i \right)$$

ALR might not be so unfamiliar to the reader, as there have been variations of ALR used for a while. One variation is log(FPKM) or the logarithm of fragments per kilobase per million reads. log(FPKM) has been designed to reflect the number of reads aligned to a reference, scaled by the length of the reference sequence in kilobases, and then divided with the total number of fragments in millions in a sample. However, the choice of $x_D$ does not need to be the output size but can be what the analyst chooses. An interpretation of ALR is that the counts are compared to a specific reference [127]. There is still the issue of using the total number of reads available as $x_D$ does not give information about the environment [126], which is why the sum of bacterial reads is often chosen in AMR research instead. By using the sum of bacterial reads as the reference, the ALR abundance will reflect the fraction of ARGs to the amount of bacteria in the environment.

The CLR transformation does not require a specific reference being chosen but instead uses the geometric average of the compositional vector. The most significant advantage of CLR-transformed data is that the values can be used for multivariate hypothesis testing and building models [126]. That is because CLR values are invariant to scaling (multiplying by a number), perturbation (unit conversion), and permutation (change of order) and show sub-compositional dominance (a subset of

the composition is less informative than the full) [130, 127].

## 3.1 Handling the sparsity of count data

During the process of aligning read fragments to reference sequences, there might be millions of references to pick between. An abundance matrix of at least two samples can therefore be quite large and filled with zeroes, as there might be references that did not have any read hits in one sample but did in the other one. This often results in a sparse abundance matrix. A quick glance at the definitions of ALR and CLR will, hopefully, raise the question of what to do, as taking the logarithm of zero is impossible.

Considering what a zero represents is necessary before tackling the issue of its presence. It might be easy to jump to that zero means that the gene is absent because no reads were matched, but that cannot be true if we consider the pool of reads a fixed-size random sample of the true distribution. A zero might mean the gene is not there or that, by chance, the reads matching the gene were simply just not observed. Since there is no way to know which of the two statements is correct, there is a tendency to believe the latter: the deeper a sample is sequenced, the more references will get at least one matching read [131].

There is an ongoing discussion on how to best handle zeroes [127]. Zeroes can be replaced with a small, fixed number, but this imputation is more a way of masking the missing data than actually handling it [132]. Another way to replace zeroes is to use a Dirichlet sampling procedure, which converts the count of reads into proportions of reads and thus eliminates the zeroes. Using this sampling from a probability distribution is much more computationally expensive than just adding a small value. Still, it does have the benefit of incorporating the mapping evidence [133].

## 3.2 Analyzing metagenomic data with CoDa methods

Many methods used to describe microbiomes stem from ecology, where the goal typically is to investigate the relationship between species and their environment. These parallels are also valid for analyses of metagenomic samples, just that the goal is not to count animals or plants but the genetic material. As already touched upon earlier, calculating the diversity of the samples is an excellent first step in exploring the microbiome data.

Specifically, measuring the diversity between samples, the beta-diversity, can reveal similarities or differences in microbiomes. Between two samples, it can be quantified with the Aitchison distance, a compositional measurement that functions the same as using Euclidean distances on CLR values. Distances can be used for clustering and ordination of samples. The variance in samples can be visualized with Principal Component Analysis (PCA) biplots, displaying the relationship between variances in counts in samples can be studied (Figure 3.2). The beta diversity is best explored through a PCA biplot [126].

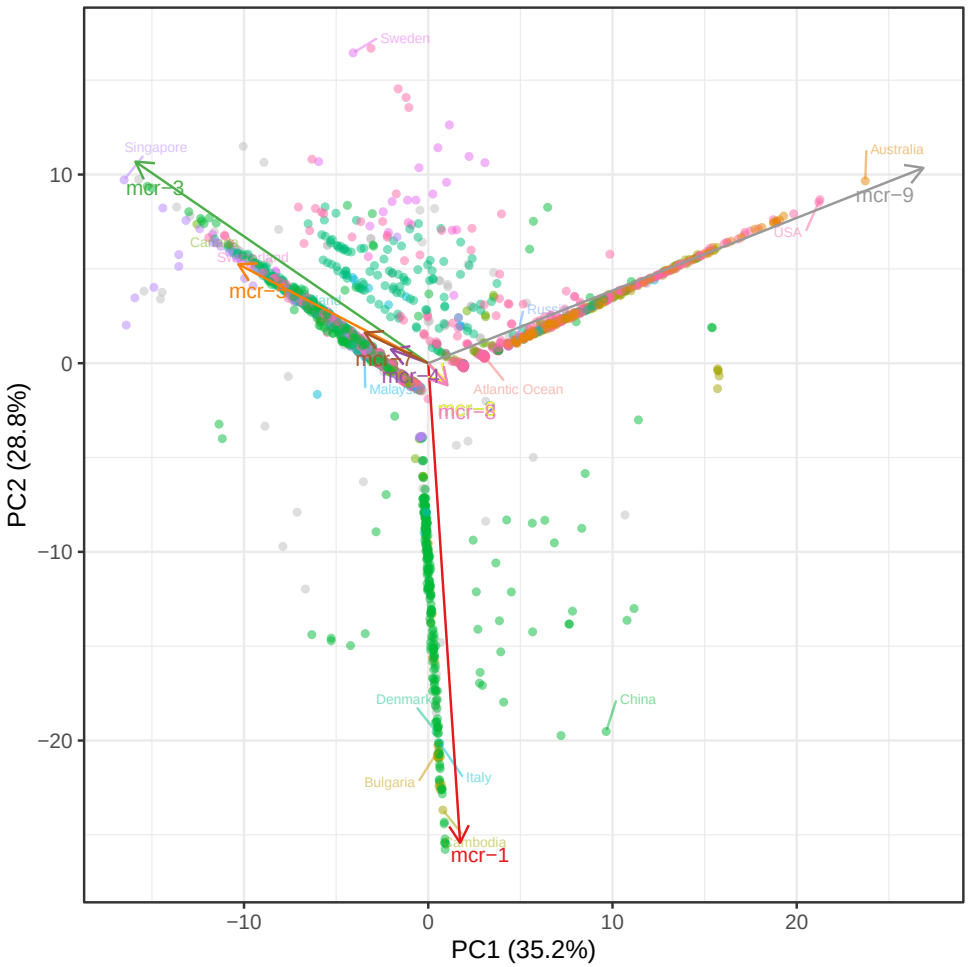



**Figure 3.2.** An example of a PCA biplot that is part of manuscript II. It shows the variance in mcr gene abundances, where, for instance, *mcr*-1 and *mcr*-9 is separated by the sampling location.

While a PCA biplot might display variances in the data, it does not say whether the abundance of a reference sequence is significantly different between sample groups. Differential Abundance (DA) testing has been done in different ways, but not all DA-models are designed for compositional data [134]. Benchmarking studies have shown that some of these tools, such as edgeR [135] and metagenomeSeq [136], suffer from high false-positive discovery rates [134]. ALDEx2 adheres to compositional principles and performs an ANOVA-Like Differential Expression (ALDEx) analysis on CLR values [101]. Another tool for DA that uses log-ratio transformations is ANCOM which relies on ALR values and tackles the issue of choosing which part a reference is by using each of them in a regression [137]. Nearing et al. [138] found that out of the many different DA models available, ALDEx2 and ANCOM-II [139], an updated version of ANCOM, produced the most consistent results across different studies. They did recommend that multiple DA tools should be used to ensure robust results.

Since microbes exist together in a microbiome and interact with each other, measuring these relationships can enhance understanding of the functional microbiome [140]. Correlations are one way to study these relationships, where compositional tools such as SparCC [102] and SpiecEasi [103] are available. Yet, obtaining all pairwise correlations for multiple samples can be a computationally heavy task if the count matrix is both large and sparse. Imagine if a matrix has counts for 500 parts, there is a possibility of getting $(n * (n-1))/2 = (500 * (500-1))/2 = 124,750$ correlations [141] and only increases with more features to test. SparCC does a log-ratio analysis of the count data and aims to estimate the basis, or truth, through an inference procedure with Aitchison's formula for variance [102]. In contrast, SpiecEasi seeks to infer correlations with a directional dependence using either a penalized regression or maximum likelihood selection on CLR transformed values [103]. Weiss et al. [141] showed that the tools infer different numbers of correlation coefficients and that some types of relationships, such as mutualism and commensalism, are better detected overall. They highlighted SparCC as being able to identify competitive relationships. Another way to study relationships in abundance is to find positively and negatively associated parts [142]. Selbal [143] finds a balance of geometric means from two groups of parts associated with a sampling group or another response variable, a so-called signature of the data for a specific outcome is identified.

Interpreting the results uncovered by compositional analyses in the context of microbial research is a difficult task. For example, it can be quite challenging to decipher a microbiome's functional role since there is no information on which genome the gene with aligned reads sits in. There might also be several confounders in the data collection, such as unreported reference sequences, sample conditions, or erroneous metadata annotations. Some results could most likely be verified through rigorous literature research or by comparing the read abundances with fully assembled genomes. The actual distribution of microbes in a sample is also somewhat unknown, so it is hard to determine how many NGS reads are needed. While there is a tendency to believe that all genes are present, for example, when working with the Dirichlet

distributions, it might not be entirely correct.

All the metagenomic NGS data available online are ready to be repurposed for AMR surveillance by analyzing the read abundances across different metagenomic samples. With so many samples, biases introduced by various sequencing platforms, environments, and sparsity of counts should be significantly less prevalent, and analyzing across so many samples, the many compositional data techniques, some of them presented in this thesis, can be used to find new patterns, and confirm existing ones, on the global dissemination and distribution of ARGs so that in the end effective regulations can be implemented to halt the spread of AMR.

# Part II

# Manuscripts

CHAPTER 4

# Manuscript I

## A curated data resource of 214K metagenomes for characterization of the global antimicrobial resistome

Hannah-Marie Martiny[1], Patrick Munk[1], Christian Brinch[1], Frank M. Aarestrup[1], Thomas Nordahl Petersen[1]

1. National Food Institute, Technical University of Denmark, Kongens Lyngby, Denmark.

Published in PLOS Biology.

DOI: https://doi.org/10.1371/journal.pbio.3001792

PLOS BIOLOGY

METHODS AND RESOURCES

# A curated data resource of 214K metagenomes for characterization of the global antimicrobial resistome

**Hannah-Marie Martiny*, Patrick Munk, Christian Brinch, Frank M. Aarestrup, Thomas N. Petersen**

Research Group for Genomic Epidemiology, Technical University of Denmark, Kongens Lyngby, Denmark

* hanmar@food.dtu.dk

Check for updates

OPEN ACCESS

**Citation:** Martiny H-M, Munk P, Brinch C, Aarestrup FM, Petersen TN (2022) A curated data resource of 214K metagenomes for characterization of the global antimicrobial resistome. PLoS Biol 20(9): e3001792. https://doi.org/10.1371/journal.pbio.3001792

**Academic Editor:** Tobias Bollenbach, Universitat zu Koln, GERMANY

**Received:** May 12, 2022

**Accepted:** August 9, 2022

**Published:** September 6, 2022

**Peer Review History:** PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pbio.3001792

**Copyright:** © 2022 Martiny et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** The code to produce the figures is available at https://github.com/hmmartiny/mARG. The data has been deposited at https://doi.org/10.5281/zenodo.6919377, and

## Abstract

The growing threat of antimicrobial resistance (AMR) calls for new epidemiological surveillance methods, as well as a deeper understanding of how antimicrobial resistance genes (ARGs) have been transmitted around the world. The large pool of sequencing data available in public repositories provides an excellent resource for monitoring the temporal and spatial dissemination of AMR in different ecological settings. However, only a limited number of research groups globally have the computational resources to analyze such data. We retrieved 442 Tbp of sequencing reads from 214,095 metagenomic samples from the European Nucleotide Archive (ENA) and aligned them using a uniform approach against ARGs and 16S/18S rRNA genes. Here, we present the results of this extensive computational analysis and share the counts of reads aligned. Over $6.76 \cdot 10^8$ read fragments were assigned to ARGs and $3.21 \cdot 10^9$ to rRNA genes, where we observed distinct differences in both the abundance of ARGs and the link between microbiome and resistome compositions across various sampling types. This collection is another step towards establishing global surveillance of AMR and can serve as a resource for further research into the environmental spread and dynamic changes of ARGs.

## Introduction

The vast amount of genomic data available in public data repositories is a unique and potentially important resource for doing research and genomic surveillance of antimicrobial resistance (AMR). Using these datasets collected from locations all over the world across different years and from various sampling sources might further aid our understanding of the emergence and distribution of antimicrobial resistance genes (ARGs).

The sharing of genomic sequence data to one of the available repositories is today a major and often mandatory step in peer-reviewed journals, for which several repositories were created by the members of the International Nucleotide Sequence Database Collaboration (INSDC) [1], including the European Nucleotide Archive (ENA) [2]. The number of sequencing data available at ENA continues to increase with an estimated doubling time of 18 months (https://www.ebi.ac.uk/ena/browser/about/statistics; accessed 2022-03-08).

documentation of the various tables can be accessed at https://hmmartiny.github.io/mARG.

**Funding:** This work was supported by the European Union's Horizon H2020 grant VEO (874735) and the Novo Nordisk Foundation (grant NNF16OC0021856: Global Surveillance of Antimicrobial Resistance). HMM, PM, CB, TNP, and FMA were all supported by both grants. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

**Abbreviations:** AMR, antimicrobial resistance; ARG, antimicrobial resistance gene; ENA, European Nucleotide Archive; INSDC, International Nucleotide Sequence Database Collaboration; mcr, mobilized colistin resistance.

Several approaches for analyzing genomic data depending on the sample types are already well established.

However, the exploration of these resources is often restricted to a few research groups only since both sufficient skills in bioinformatics and access to high-performing computer resources are needed to handle the large amount of available data.

Existing collections of analyzed datasets tend to focus on either specific sample sources, such as humans [3,4], marine [5], or urban sewage [6,7], or focus on specific genera [8]. Especially the COVID-19 pandemic has highlighted the value of data sharing to trace the spread and evolution of the virus [9]. Despite the attempts to standardize the analysis workflows of these databases, they are limited in their ability to generalize across environments and locations. A recent study [10] has shared a searchable collection of 661K bacterial genomes for exploring the global bacterial diversity across different origins, providing an easy-to-access resource for genomic research. While this is an impressive data-sharing effort, the authors did not include metagenomic samples in their pipeline. Metagenomic techniques aim to sequence all DNA in a sample and can be used to characterize the microbiome in different environments [11,12], discover novel organisms [13], monitor disease [14,15], and specific genes, such as ARGs [5,6,16].

Here, we present a large-scale metagenomic analysis of 214,095 metagenomic samples retrieved from ENA. We have carried out an assembly-free approach by aligning sequencing reads against ARGs and 16S/18S ribosomal RNA genes. We have previously published an in-depth analysis of the distribution of mobilized colistin resistance [17] based on those data. Now we both share the entire collection of mapping results and showcase how to characterize the global resistome and microbiome with this dataset. The curated metadata and mapping results are available at https://doi.org/10.5281/zenodo.6919377 and documentation at https://hmmartiny.github.io/mARG/Tables.html.

## Materials and methods

### Retrieval of metagenomes

We retrieved metagenomic datasets from ENA [2] uploaded between 2010-01-01 and 2020-01-01 that had library source as "METAGENOMIC" and library strategy of "WGS." We collected 214,095 sequencing runs from 146,732 samples from 6,307 projects corresponding to 442 Tbp of raw reads taking up 300 TB of storage. The associated metadata for each sample was also retrieved.

### Preprocessing and mapping of sequencing reads

The retrieved raw FASTQ reads were trimmed and aligned against reference sequences, as outlined in Martiny (2022) [17]. In brief, we used FASTQC v.0.11.15 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) for read quality checking and BBduk2 v.36.49 [18] for trimming the raw sequencing reads. With the k-mer-based alignment tool KMA 1.2.21 [19], the trimmed reads were mapped against reference sequences from 2 different databases: The AMR gene database ResFinder [20] (downloaded 2020-01-25), which contained 3,085 sequences of acquired ARGs, and the ribosomal rRNA Silva [21] gene database (version 138, downloaded 2020-01-16), which had 2,225,272 reference sequences with more than 88% of them being 16/18S rRNA genes. For KMA, we used the following alignment parameters: 1, -2, -3, -1 for a match, mismatch, gap opening, and gap extension. For read pairing, we used a value of 7 and a minimum relative alignment score of 0.75. Data retrieval, quality checking, trimming, and read alignments were done using the Danish National Supercomputer for Life Sciences (https://www.computerome.dk/).

## Standardization of metadata

The following attributes for each metagenome were standardized: sampling location, sampling host or environment (referred to as a host below), and sampling date.

To standardize the label for sampling locations, we looked at the values entered in the two fields "country" and "location." First, the latitude and longitude coordinates were mapped to a country using the Python library Shapely 1.7.1 [22] to find the matching area defined in one of the 3 public domain map datasets (countries, marine, and lakes) available in the Natural Earth Data collection. If the lookup failed or the coordinates were not given, the second step was to match the text attribute in the country label to ISO 3166 country codes with a fuzzy search with the Python library PyCountry 20.7.3 (https://github.com/flyingcircusio/pycountry). Finally, if the 2 lookup searches did not yield a match, we did a manual lookup of the country labels to standardize the text.

For the standardization of host labels, we mapped the taxonomic id given by the attribute "host_tax_id" to the NCBI Taxonomy database [23], or if the feature was missing, the "tax_id" was used instead.

Since the only way to curate entered collection dates is to look up suspicious dates in published studies manually, and that was deemed too time-intensive, we decided to replace dates entered as later than 2020-01-01 in the sample attribute field "collection_date" with the missing value NULL.

## Measuring the abundance of ARGs

Since we report the fragment count aligned to each reference gene, the mapping results are compositional and should be treated as such [24]. In the simplest form, the ARG abundance for a sample or sample group can be calculated as the log-ratio of the count of reads, $n_i$, aligned to each ARG $i$ over the total sum of rRNA read fragments $n_B$:

$$x = [n_1, n_2, \ldots, n_D, n_B], i = 1..D$$

$$\text{Abundance}(x) = \left[\log\frac{n_1}{n_B}, \log\frac{n_2}{n_B}, \ldots, \log\frac{n_D}{n_B}\right]$$

where $D$ is the number of ARGs and $n_B = \frac{\sum_j^{D_B} n_j}{1 \cdot 10^6}$ with $D_B$ being the number of read fragments aligned to rRNA genes. Each ARG count $n_i$ has been adjusted with the length of the gene in kilobases.

The relative abundance resistance classes were calculated as the proportion of ARG resistance assigned to different classes and scaled with $\kappa = 100$:

$$\text{Relative abundance}(x) = \frac{\kappa}{\sum n_i} n_i$$

## Diversity measurements

Besides the read abundance values, we report the species richness, Shannon diversity index [25], and the Gini–Simpson [26] diversity index of read counts of ARGs, genera, and phyla per sample. Species richness is the number of different genes or taxonomic groups present in the sample with at least 1 read fragment aligned.

The Shannon index ($H'$) was calculated using the proportions of reads $p_i = \frac{n_i}{\sum n}$:

$$H' = -\sum_{i=1}^{R} p_i \ln p_i$$

whereas the Gini–Simpson index ($GS$) was calculated using the read counts $n = [n_1, \ldots, n_D]$ and $N = \sum n$ is the total count of reads for the group:

$$GS = 1 - \frac{\sum n_i \cdot (n_i - 1)}{N \cdot (N - 1)}$$

Together with these 2 indices, we also report the sample-wise unique number of reference sequences or taxonomic groups matched.

## Results

Here, we present a large-scale mapping of 442 Tbp of raw reads of 214,095 metagenomic samples suitable for analyzing the distribution of acquired antimicrobial resistance genes and 16S/18S rRNA genes. Furthermore, we have spent considerable effort standardizing 3 main sample attributes: sampling date, location, and source. To facilitate easy access and usage, we have shared the mapping results and corrected metadata in 3 different data formats (TSV, HDF, and MySQL dumps). We also provide tutorials with code examples in R and Python on using the data in different scenarios. Data files are all available at https://doi.org/10.5281/zenodo.6919377.

By collecting the sequencing reads from ENA, we could also verify the inherited bias of specific sample types or sources being overrepresented simply due to the availability in the public repository. While the 214,095 metagenomic datasets were collected from 797 different hosts, most were either of human or marine origin (Fig 1A). A similar skewed geographical distribution towards European and North American countries was observed in the sampling locations

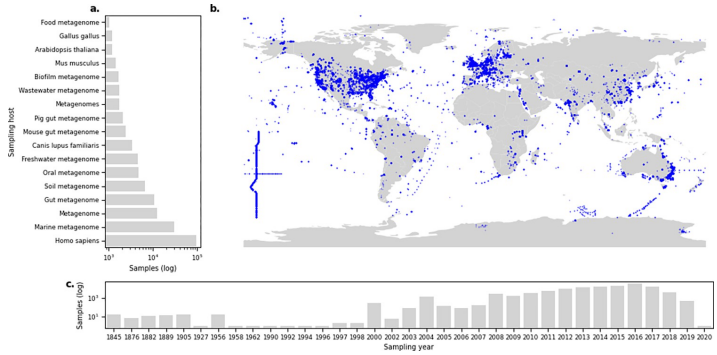



**Fig 1. Distribution of metagenomes reveals the overrepresentation of samples from specific sources.** (**a**) Number of samples grouped per sampling host, where only hosts with more than 1,000 samples are plotted. (**b**) Sample locations for metagenomes with available GPS coordinates; each marker is a sample. A total of 83,903 samples did not have coordinates available. (**c**) Year of which a sample was collected. A total of 84,238 of the samples did not have a valid sampling date recorded. The data underlying this figure can be found at https://doi.org/10.5281/zenodo.6919377, and the base layer map was created with data from https://www.naturalearthdata.com/.

https://doi.org/10.1371/journal.pbio.3001792.g001

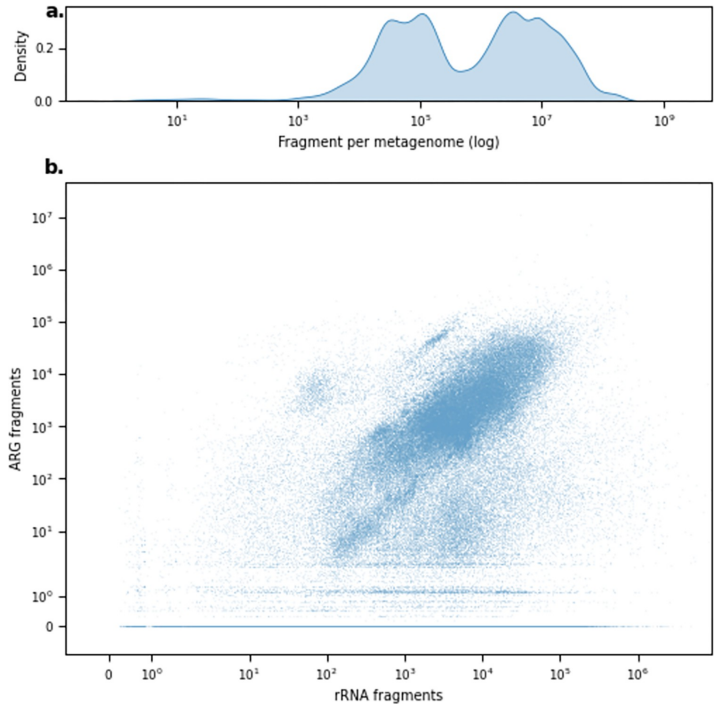



**Fig 2. Distribution of available and aligned fragments. (a)** Density distribution of available fragments per sample. **(b)** The distribution compares the number of fragments mapped to rRNA genes and ARGs. The data underlying this figure can be found at https://doi.org/10.5281/zenodo.6919377.

https://doi.org/10.1371/journal.pbio.3001792.g002

(Fig 1B). The distribution of samples according to the sampling year reveals that a considerable number were collected between 2010 and 2020 (Fig 1C).

Of the more than $1.8{\cdot}10^{12}$ raw sequencing reads, corresponding to 442.1 Tbp, 93% of the reads were generated using Illumina sequencing technologies (S1 Fig). We mapped over $1.69{\cdot}10^{12}$ trimmed read fragments, with a median of 784,748 fragments per sample (range 1 to 916,901,400) (Fig 2A). Approximately 0.04% of all read fragments could be aligned to ARGs, and 0.19% to rRNA genes. Overall, the amount of sequencing reads and bases available did increase the count of aligned read fragments (S3 Fig). The number of ARG fragments aligned increased with the number of aligned rRNA fragments, although for 34% of the samples, we did not find any ARGs despite having read fragments aligning to 16S rRNA genes (Fig 2B). The microbial differences in the different sampling origins were highlighted in the number of aligned fragments (S4 Fig).

## The global abundance of antimicrobial resistance

To measure the global distribution of ARGs and the composition of the resistome, we calculated the abundance of ARGs as the log-ratio of ARG fragments over summed rRNA sequence

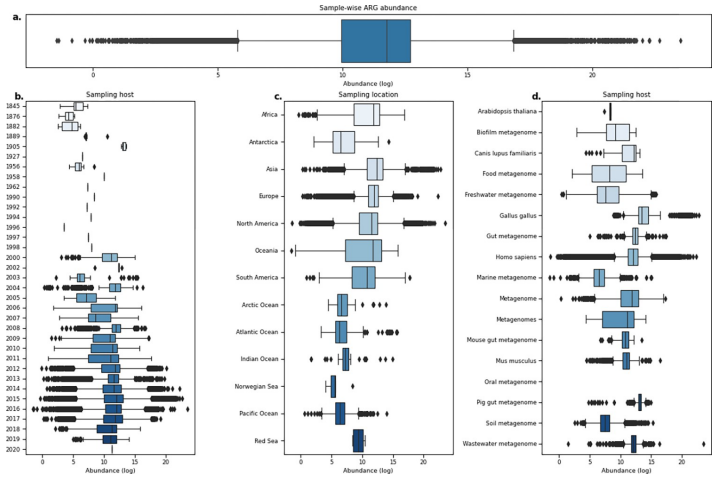



**Fig 3. Boxplots of ARG abundances in metagenomic samples show that levels vary across different origins. (a)** Distribution of ARG abundance per sample. **(b)** Distribution of sample-wise ARG abundance grouped by sampling year. **(c)** Sample-wise ARG abundance per sampling location. **(d)** Sample-wise ARG abundance grouped by hosts. Only hosts with more than 1,000 metagenomes analyzed are shown. The data underlying this figure can be found at https://doi.org/10.5281/zenodo.6919377.

https://doi.org/10.1371/journal.pbio.3001792.g003

fragments. Almost all of the reference sequences from the ResFinder database had at least 1 fragment aligned, and only 94 ARGs had no hits (S2 Fig). The median observed resistance load per metagenomic sample was 11.74 (log range: −1.45 to 23.52) (Fig 3A), which appeared to be mainly dependent on the geographic origin and environment (Fig 3B–3D) and not on which year the sample was taken. For example, samples originating from locations within Europe showed similar abundance levels for most of the samples but with several outliers, whereas multiple samples from locations in the Oceania region had a much broader load distribution with few outliers (Fig 3C).

While the distribution of sample-wise resistance loads illustrates the high variability in this data collection (Fig 3), we saw that once we stratified the relative ARG read proportions per resistance class and sample type, there were clear separations between different groups (Fig 4). For the sampling years with a considerable number of samples available (2004 to 2019), the relative proportion of classes was relatively consistent, with Tetracycline reads being the most common, except for a spike of Beta-lactam reads in 2017 (Fig 4A). Across the continents and large water bodies, we observed that ARGs conferring resistance to Aminoglycosides or Beta-lactam antimicrobials were more common in water environments, whereas mainland regions had a more diverse distribution (Fig 4B). Once we stratified by sampling host or source, the distribution of resistance classes was very dependent on the group, as seen by the high proportion of read fragments aligned to, for example, Phenicol for marine and soil samples and Tetracycline reads being highly prevalent in mice (*Mus musculus*) samples (Fig 4C).

## Linking the microbiome diversity with resistance diversity

The relationship between the diversity of the microbiome and the resistance genes was quantified by calculating the species richness and 2 alpha diversity measurements (Shannon and

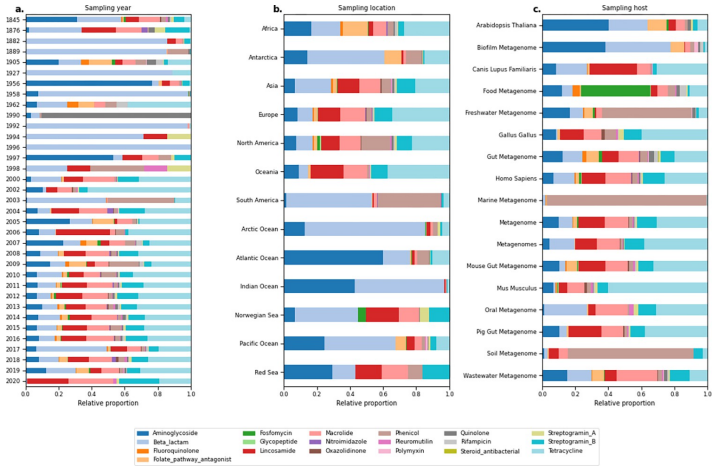



**Fig 4. Composition of reads assigned to ARGs from different resistance classes grouped by sampling origin.** **(a)** Grouped by sampling year. **(b)** Grouped per sampling location. **(c)** Grouped per sampling host. Only hosts with more than 1,000 metagenomes analyzed are shown. The data underlying this figure can be found at https://doi.org/10.5281/zenodo.6919377.

https://doi.org/10.1371/journal.pbio.3001792.g004

Gini–Simpson) on ARG levels and phyla and genera taxonomic levels. Without looking at the sample origin, we observed that a majority of the samples had both high microbial diversity and ARG diversity (Figs 5 and S5). However, the relationship between genera and ARG diversity indexes differed between sampling sources, with several groups containing samples that

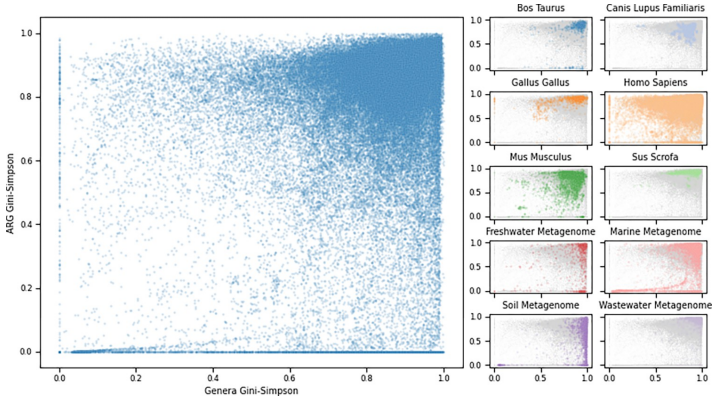



**Fig 5. The genus–ARG diversity relationship for all metagenomic samples.** The Gini–Simpson diversity indexes were calculated on genus categories (x-axis) compared to ARG levels (y-axis). Left: scatterplot of all samples. Right: samples colored by selected host or environmental origins. The data underlying this figure can be found at https://doi.org/10.5281/zenodo.6919377.

https://doi.org/10.1371/journal.pbio.3001792.g005

did not follow the assumption of the 2 diversity measurements following each other, suggesting that increased diversity of microbes in, for example, soil samples does not necessarily lead to a higher diversity of resistance genes. Contrarily, the chicken (*Gallus gallus*) samples showed that they still had elevated ARG diversity despite having lower microbial diversity (Fig 5).

## Discussion

Global surveillance of AMR based on genomics continues to become more accessible due to the advancement in NGS technologies and the practice of sharing raw sequencing data in public repositories. Standardized pipelines and databases are needed to utilize these large data volumes for tracking the dissemination of AMR. We have uniformly processed the sequencing reads of 214,095 metagenomes for the abundance analysis of ARGs.

Our data sharing efforts enable users to perform abundance analyses of individual ARGs, the resistome, and the microbiome across different environments, geographic locations, and sampling years.

We have given a brief characterization of the distribution of ARGs according to the collection of metagenomes. However, in-depth analyses remain to be performed to investigate the influence of temporal, geographical, and environmental origins on the dissemination and evolution of antimicrobial resistance. For example, analyzing the spread of specific ARGs across locations and different environments could reveal new transmission routes of resistance and guide the design of intervention strategies to stop the spread. We have previously published a study focusing on the distribution of mobilized colistin resistance (*mcr*) genes using this data resource, showing how widely disseminated the genes were [17]. Another use of the data collection could be to explore how the changes in microbial abundances affect and are affected by the resistome. Furthermore, our coverage statistics of reads aligned to ARGs could be used to investigate the rate of new variants occurring in different reservoirs. Even though we have focused on the threat of antimicrobial resistance, potential applications of this resource can be to look at the effects of, for example, climate changes on microbial compositions. Linking our observed read fragment counts with other types of genomic data, such as evaluating the risk of ARG mobility, accessibility, and pathogenicity in assembled genomes [27,28], and verifying observations from clinical data [29].

We recommend that potential users consider all the confounders present in this data collection in their statistical tests and modeling workflows, emphasizing that the experimental methods and sequencing platforms dictate the obtained sequencing reads and that metadata for a sample might be mislabeled, despite our efforts to minimize those kinds of errors. Furthermore, it is essential to consider the compositional nature of microbiomes [30]. The reads do not depend on the distribution of genetic material in the sample but on the capacity of the sequencing platform [24,31]. Various statistical methods already exist that consider the compositionality [24,32,33]. Finally, it is important to highlight that the results we have presented here include fragment counts of 1 for the sake of transparency, but we also recommend potential users consider appropriate filters in their analysis.

The sequencing data in public repositories has continued to grow, giving us plenty of opportunities to continue to expand our data collection even more. To establish a truly global surveillance program of AMR, sequencing data should be analyzed as soon as published in these archives. Although this would require access to even more computational resources, we hope to achieve this soon and compare our approach with other methods, such as AMRFinderPlus [34] and CARD [35]. As new sequencing technologies are becoming more used, our settings for our alignment procedure should also be tuned to better take advantage and be aware of the flaws of different sequencing platforms.

With this data resource, we have taken a step towards enabling the scientific community to utilize the wealth of information in these metagenomic samples to broaden our understanding of the dissemination of antimicrobial resistance and changes in microbiomes at both local and global scales through time and environments.

## Supporting information

**S1 Fig. Distribution of samples per sequencing instrument platform. (a)** Sample count per platform. **(b)** Distribution of raw sequencing read counts per platform. The data underlying this figure can be found at https://doi.org/10.5281/zenodo.6919377.
(TIFF)

**S2 Fig. More than 96% of ARG templates had at least 1 aligned fragment.** The bars illustrate the percentage of ARGs per resistance class without and with at least 1 aligned fragment. The parenthesis after each class label contains the number of genes found out of the total available templates. The data underlying this figure can be found at https://doi.org/10.5281/zenodo.6919377.
(TIFF)

**S3 Fig. The sample-wise distribution of aligned (a) ARG or (b) rRNA fragments compared to raw sequencing base counts.** The data underlying this figure can be found at https://doi.org/10.5281/zenodo.6919377.
(TIFF)

**S4 Fig. The sample-wise distribution of aligned rRNA fragments and ARG fragments, colored by selected host and environmental sources.** The data underlying this figure can be found at https://doi.org/10.5281/zenodo.6919377.
(TIFF)

**S5 Fig. Additional distributions showing the relationship between ARGs and genera for all metagenomic samples. (a)** The richness of genus groups (x-axis) vs. ARG richness (y-axis). **(b)** The relationship between Shannon diversity index calculated on genus level (x-axis) and ARGs (y-axis). Right: samples colored by selected host or environmental origins. The data underlying this figure can be found at https://doi.org/10.5281/zenodo.6919377.
(TIFF)

## Author Contributions

**Conceptualization:** Frank M. Aarestrup, Thomas N. Petersen.

**Data curation:** Hannah-Marie Martiny.

**Formal analysis:** Hannah-Marie Martiny.

**Funding acquisition:** Frank M. Aarestrup, Thomas N. Petersen.

**Investigation:** Hannah-Marie Martiny, Patrick Munk.

**Methodology:** Hannah-Marie Martiny, Patrick Munk, Thomas N. Petersen.

**Project administration:** Patrick Munk, Frank M. Aarestrup, Thomas N. Petersen.

**Resources:** Hannah-Marie Martiny, Frank M. Aarestrup.

**Software:** Hannah-Marie Martiny.

**Supervision:** Patrick Munk, Christian Brinch, Frank M. Aarestrup, Thomas N. Petersen.

**Validation:** Hannah-Marie Martiny, Christian Brinch.

**Visualization:** Hannah-Marie Martiny.

**Writing – original draft:** Hannah-Marie Martiny.

**Writing – review & editing:** Hannah-Marie Martiny, Patrick Munk, Christian Brinch, Frank M. Aarestrup, Thomas N. Petersen.

## References


1. Arita M., Karsch-Mizrachi I., Cochrane G. The international nucleotide sequence database collaboration. Nucleic Acids Res. (2021) 49, D121. https://doi.org/10.1093/nar/gkaa967 PMID: 33166387
2. Leinonen R. et al. The European nucleotide archive. Nucleic Acids Res. (2011) 39, 44–47.
3. Shao L., Liao J., Qian J., Chen W., Fan X. MetaGeneBank: a standardized database to study deep sequenced metagenomic data from human fecal specimen. BMC Microbiol. (2021) 21, 1–12.
4. Almeida A. et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol. (2021) 39, 105–114. https://doi.org/10.1038/s41587-020-0603-3 PMID: 32690973
5. Cuadrat R. R. C., Sorokina M., Andrade B. G., Goris T., Dávila A. M. R. Global ocean resistome revealed: Exploring antibiotic resistance gene abundance and distribution in TARA Oceans samples. Gigascience. (2020) 9, 1–12. https://doi.org/10.1093/gigascience/giaa046 PMID: 32391909
6. Hendriksen R. S. et al. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nat Commun, (2019) 10. https://doi.org/10.1038/s41467-019-08853-3 PMID: 30850636
7. Fresia P. et al. Urban metagenomics uncover antibiotic resistance reservoirs in coastal beach and sewage waters. Microbiome. (2019) 7, 1–9.
8. Zhou Z., Alikhan N. F., Mohamed K., Fan Y., Achtman M. The EnteroBase user's guide, with case studies on Salmonella transmissions, Yersinia pestis phylogeny, and Escherichia core genomic diversity. Genome Res. (2020) 30, 138–152. https://doi.org/10.1101/gr.251678.119 PMID: 31809257
9. Khare S. et al. GISAID's Role in Pandemic Response. China CDC Wkly. (2021) 3, 1049–1051. https://doi.org/10.46234/ccdcw2021.255 PMID: 34934514
10. Blackwell G. A. et al. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. PLoS Biol. (2021) 19, e3001421. https://doi.org/10.1371/journal.pbio.3001421 PMID: 34752446
11. Fierer N. et al. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. Proc Natl Acad Sci U S A. (2012) 109, 21390–21395. https://doi.org/10.1073/pnas.1215210110 PMID: 23236140
12. Gill S. R. et al. Metagenomic analysis of the human distal gut microbiome. Science (80-). (2006) 312, 1355–1359. https://doi.org/10.1126/science.1124234 PMID: 16741115
13. Al-Shayeb B. et al. Clades of huge phages from across Earth's ecosystems. Nature. (2020) 578, 425–431. https://doi.org/10.1038/s41586-020-2007-4 PMID: 32051592
14. Nieuwenhuijse D. F. et al. Setting a baseline for global urban virome surveillance in sewage. Sci Rep. (2020) 10, 1–13.
15. Liu P., Chen W., Chen J. P. Viral Metagenomics Revealed Sendai Virus and Coronavirus Infection of Malayan Pangolins (Manis javanica). Viruses 2019, Vol 11, Page 979 (2019) **11**, 979. https://doi.org/10.3390/v11110979 PMID: 31652964
16. Forsberg K. J. et al. Bacterial phylogeny structures soil resistomes across habitats. Nature. (2014) 509, 612–616. https://doi.org/10.1038/nature13377 PMID: 24847883
17. Martiny H.-M. et al. Global distribution of mcr gene variants in 214,095 metagenomic samples. mSystems. (2022). https://doi.org/10.1128/msystems.00105-22 PMID: 35343801
18. Bushnell B. BBMap. (2014).
19. Clausen P. T. L. C., Aarestrup F. M., Lund O. Rapid and precise alignment of raw reads against redundant databases with KMA. BMC Bioinformatics. (2018) 19, 1–8.
20. Zankari E. et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. (2012) 67, 2640–2644. https://doi.org/10.1093/jac/dks261 PMID: 22782487
21. Quast C. et al. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. (2013) 41, 590–596. https://doi.org/10.1093/nar/gks1219 PMID: 23193283

22. Gillies S., Others A. Shapely: manipulation and analysis of geometric objects. (2007).
23. Federhen S. The NCBI Taxonomy database. Nucleic Acids Res. (2012) 40, D136–D143. https://doi.org/10.1093/nar/gkr1178 PMID: 22139910
24. Gloor G. B., Macklaim J. M., Pawlowsky-Glahn V., Egozcue J. J. Microbiome datasets are compositional: And this is not optional. Front Microbiol. (2017) 8, 1–6.
25. Shannon C. E. A mathematical theory of communication. Bell Syst Tech J. (1948) 27, 379–423.
26. Jost L. Entropy and diversity. Oikos. (2006) 113, 363–375.
27. Zhang A. N. et al. An omics-based framework for assessing the health risk of antimicrobial resistance genes. Nat Commun. (2021) 12, 1–11.
28. Zhang Z. et al. Assessment of global health risk of antibiotic resistance genes. Nat Commun, (2022) 13. https://doi.org/10.1038/s41467-022-29283-8 PMID: 35322038
29. Karkman A., Berglund F., Flach C. F., Kristiansson E., Larsson D. G. J. Predicting clinical resistance prevalence using sewage metagenomic data. Commun Biol. (2020) 3, 1–10.
30. Aitchison J. The Statistical Analysis of Compositional Data. J R Stat Soc Ser B. (1982) 44, 139–160.
31. Quinn T. P. et al. A field guide for the compositional analysis of any-omics data. Gigascience. (2019) 8, 1–14. https://doi.org/10.1093/gigascience/giz107 PMID: 31544212
32. Fernandes A. D., Macklaim J. M., Linn T. G., Reid G., Gloor G. B. ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. PLoS ONE. (2013) 8. https://doi.org/10.1371/journal.pone.0067019 PMID: 23843979
33. Friedman J., Alm E. J. Inferring Correlation Networks from Genomic Survey Data. PLoS Comput Biol. (2012) 8, 1–11.
34. Feldgarden M. et al. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. (2021) 11.
35. Alcock B. P. et al. CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res. (2020) 48, D517–D525. https://doi.org/10.1093/nar/gkz935 PMID: 31665441

## Supplementary Material

Martiny, H. M., Munk, P., Brinch, C., Aarestrup, F. M., & Petersen, T. N. (2022). A curated data resource of 214K metagenomes for characterization of the global antimicrobial resistome. PLOS Biology 20(9): e3001792.

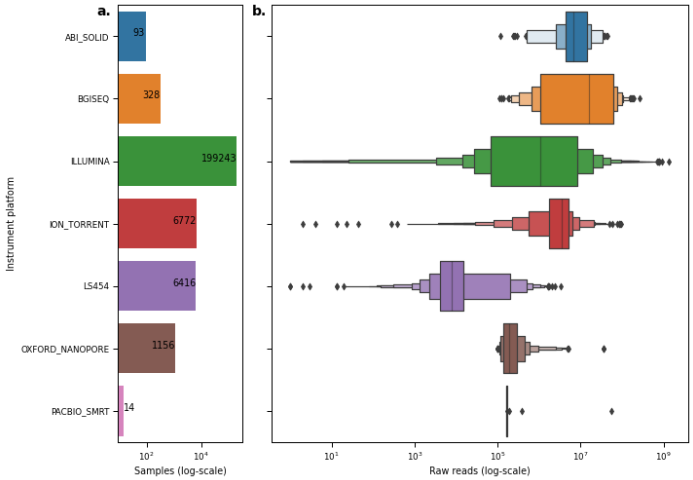



*Figure S1:* ***Distribution of samples per sequencing instrument platform. a.*** *Sample count per platform.* ***b.*** *Distribution of raw sequencing read counts per platform. The data underlying this figure can be found at https://doi.org/10.5281/zenodo.6919377.*

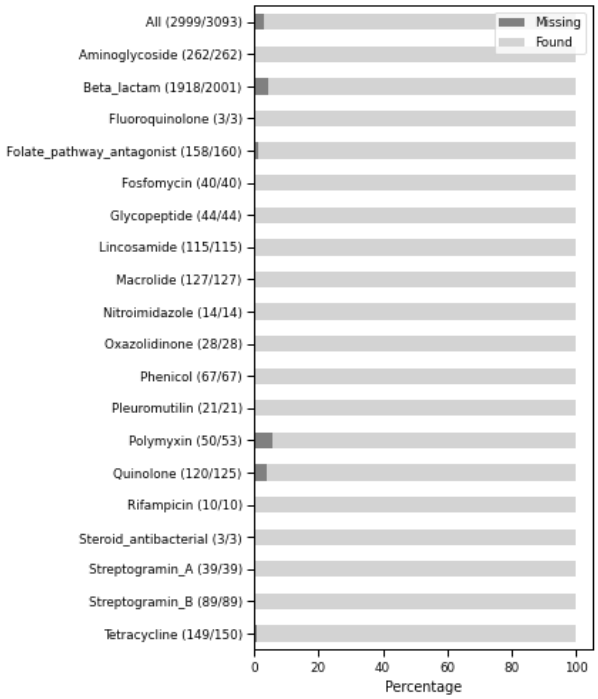



*Figure S2:* ***More than 96% of ARG templates had at least one aligned fragment.*** *The bars illustrate the percentage of ARGs per resistance class without and with at least one aligned fragment. The parenthesis after each class label contains the*

## Supplementary Material

Martiny, H. M., Munk, P., Brinch, C., Aarestrup, F. M., & Petersen, T. N. (2022). A curated data resource of 214K metagenomes for characterization of the global antimicrobial resistome. PLOS Biology 20(9): e3001792.

*number of genes found out of the total available templates. The data underlying this figure can be found at https://doi.org/10.5281/zenodo.6919377.*

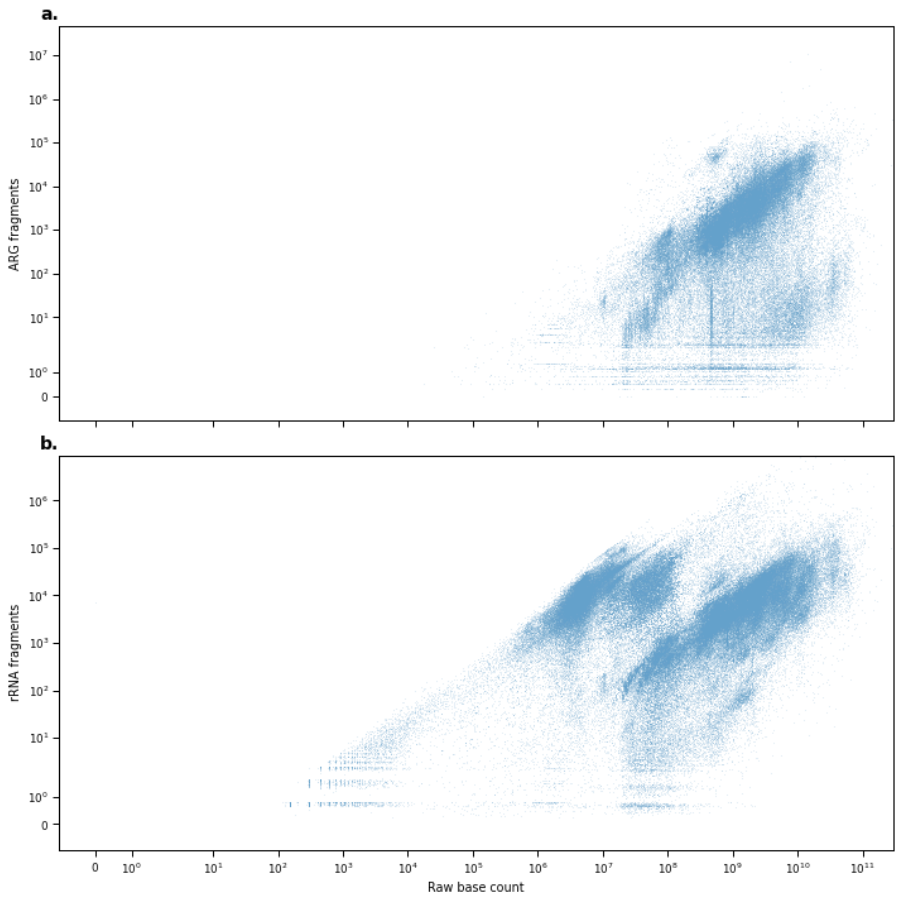



*Figure S3:* ***The sample-wise distribution of aligned a. ARG or b. rRNA fragments compared to raw sequencing base counts.*** *The data underlying this figure can be found at https://doi.org/10.5281/zenodo.6919377.*

## Supplementary Material

Martiny, H. M., Munk, P., Brinch, C., Aarestrup, F. M., & Petersen, T. N. (2022). A curated data resource of 214K metagenomes for characterization of the global antimicrobial resistome. PLOS Biology 20(9): e3001792.

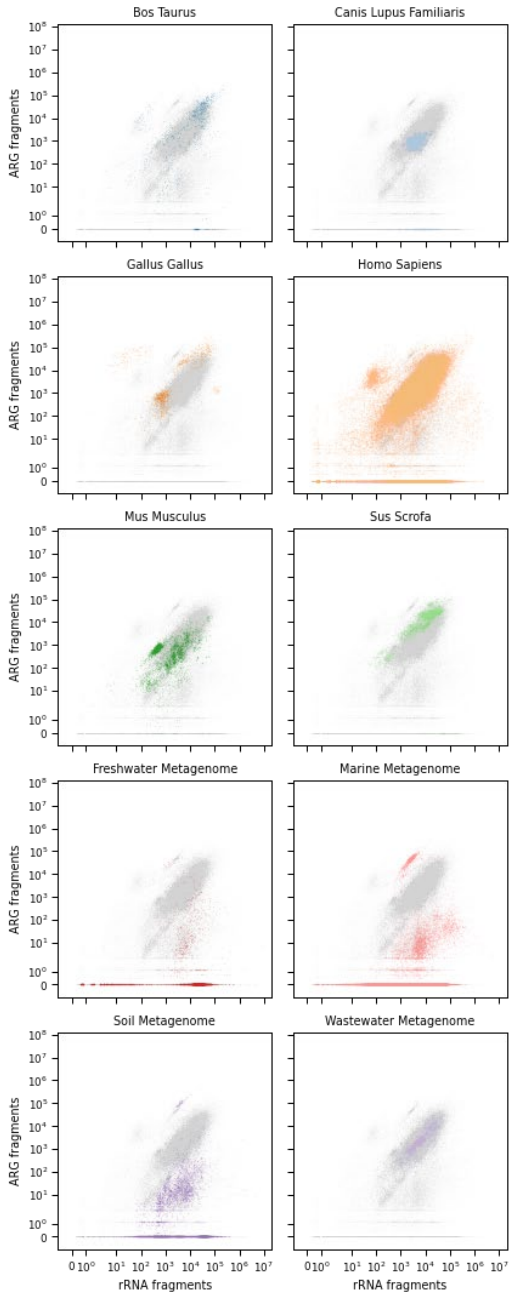



*Figure S4:* ***The sample-wise distribution of aligned rRNA fragments and ARG fragments****, colored by selected host and environmental sources. The data underlying this figure can be found at https://doi.org/10.5281/zenodo.6919377.*

## Supplementary Material

Martiny, H. M., Munk, P., Brinch, C., Aarestrup, F. M., & Petersen, T. N. (2022). A curated data resource of 214K metagenomes for characterization of the global antimicrobial resistome. PLOS Biology 20(9): e3001792.

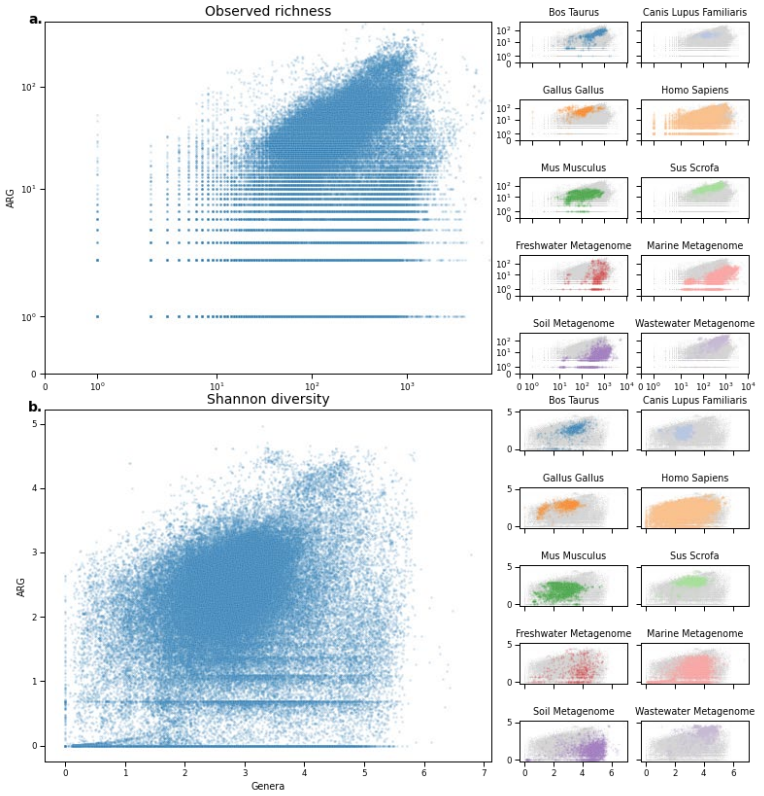



*Figure S5:* ***Additional distributions showing the relationship between ARGs and genera for all metagenomic samples. a.*** *The richness of genus groups (x-axis) vs. ARG richness (y-axis).* ***b.*** *The relationship between Shannon diversity index calculated on genus level (x-axis) and ARGs(y-axis). Right: Samples colored by selected host or environmental origins. The data underlying this figure can be found at https://doi.org/10.5281/zenodo.6919377.*

CHAPTER 5

# Manuscript II

## Global Distribution of *mcr* Gene Variants in 214K Metagenomic Samples

Hannah-Marie Martiny[1], Patrick Munk[1], Christian Brinch[1], Judit Szarvas[1], Frank M. Aarestrup[1], Thomas Nordahl Petersen[1]

1. National Food Institute, Technical University of Denmark, Kongens Lyngby, Denmark.

Published in mSystems.

DOI: https://doi.org/10.1128/msystems.00105-22

AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE

# Global Distribution of *mcr* Gene Variants in 214K Metagenomic Samples

**Hannah-Marie Martiny,[a] Patrick Munk,[a] Christian Brinch,[a] Judit Szarvas,[a] Frank M. Aarestrup,[a] Thomas Nordahl Petersen[a]**

[a]Research Group for Genomic Epidemiology, Technical University of Denmark, Kgs. Lyngby, Denmark

**ABSTRACT** Since the initial discovery of a mobilized colistin resistance gene (*mcr-1*), several other variants have been reported, some of which might have circulated a while beforehand. Publicly available metagenomic data provide an opportunity to reanalyze samples to understand the evolutionary history of recently discovered antimicrobial resistance genes (ARGs). Here, we present a large-scale metagenomic study of 442 Tbp of sequencing reads from 214,095 samples to describe the dissemination and emergence of nine *mcr* gene variants (*mcr-1* to *mcr-9*). Our results show that the dissemination of each variant is not uniform. Instead, the source and location play a role in the spread. However, the genomic context and the genes themselves remain primarily unchanged. We report evidence of new subvariants occurring in specific environments, such as a highly prevalent and new variant of *mcr-9*. This work emphasizes the importance of sharing genomic data for the surveillance of ARGs in our understanding of antimicrobial resistance.

**IMPORTANCE** The ever-growing collection of metagenomic samples available in public data repositories has the potential to reveal new details on the emergence and dissemination of mobilized colistin resistance genes. Our analysis of metagenomes deposited online in the last 10 years shows that the environmental distribution of *mcr* gene variants depends on sampling source and location, possibly leading to the emergence of new variants, although the contig on which the *mcr* genes were found remained consistent.

**KEYWORDS** antimicrobial resistance, metagenomics, microbiome

Antimicrobial resistance (AMR) is considered one of the most significant threats against human and animal health (1). Over the years, we have observed the emergence of a multitude of novel antimicrobial resistance genes (ARGs), and it is generally believed that such genes have emerged and evolved in the commensal flora for a long time prior to being detected in pathogenic isolates (2).

Colistin is an important antibiotic used as a last-resort choice to treat multidrug-resistant (MDR) and carbapenem-resistant bacteria (3). Before 2015, colistin resistance was believed to be only due to mutational and regulatory changes in chromosomal genes. A mobilized colistin resistance gene, *mcr-1*, was discovered in 2015 on a plasmid in *Escherichia coli* isolates from China (4), raising concern in the scientific community about the possibility of resistance spreading more rapidly by horizontal gene transfer by mobile genetic elements (MGEs) (4, 5). Immediately following the first report, a large number of studies were initiated in several countries around the world, and it was soon determined that *mcr-1* was already widespread and has now been detected in all continents (6–8). In initial reports, the most frequent isolates were sampled from livestock sources, followed by humans, meat, and food products (9). Since then, several new variants of *mcr* genes have also been identified, named *mcr-2* to *mcr-10* and sharing 81%, 32.5%, 34%, 36%, 83%, 35%, 31%, 36%, and 29.31% amino acid sequence

**Editor** Zackery Bulman, University of Illinois at Chicago

**Ad Hoc Peer Reviewer** Val Fernández-Lanza, Ramón y Cajal Institute for Health Research

**Copyright** © 2022 Martiny et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Hannah-Marie Martiny, hanmar@food.dtu.dk.

The authors declare no conflict of interest.

**Received** 3 February 2022
**Accepted** 28 February 2022
**Published** 28 March 2022

Downloaded from https://journals.asm.org/journal/msystems on 29 July 2022 by 192.38.90.185.

identity to *mcr-1*, respectively (10–17). Retrospective screening of bacterial isolates and available sequences of mainly pathogenic isolates showed a more widespread occurrence and prior evolution of *mcr* before its initial discovery (8, 18, 19).

However, investigating only pathogenic strains or cultivable bacteria will only provide limited insight into the potential reservoirs of such novel ARGs. As documented by our research group and others, investigating the entire microbiome provides additional information on the presence and diversity of ARGs (7, 20–22). Today most researchers conducting microbiome studies are depositing their raw data in global repositories, allowing other researchers to reanalyze the data and provide novel insight.

This study was conducted to investigate the occurrence and global dissemination of known *mcr* gene variants in publicly available metagenomic data sets. We downloaded 442 Tbp of raw reads from 214,095 metagenomic data sets and determined the presence and abundance of 9 *mcr* gene variants. We found that only a small subset of the metagenomic data sets was positive for at least one of the *mcr* genes but that the abundance gradually increased as a function of time. The distribution of each variant varied by region and sampling source, but the genomic background of each gene was consistent across different environments. However, several subvariants are observed with conserved single nucleotide polymorphisms (SNPs) across multiple samples. Despite the sparsity of the data once stratified by the presence of *mcr* genes, our analysis suggests that multiple factors have likely influenced the dissemination of colistin resistance and that screening publicly available metagenomic samples can, together with single isolates, further deepen our understanding of the distribution of *mcr* gene variants.

## RESULTS

**Data set.** After retrieval, quality checking, and trimming of the raw sequencing reads of the 214,095 metagenomic data sets, we aligned the reads against ARGs and 16S rRNA sequences using the assembly-free method KMA. The resulting counts of read fragments aligned to different reference sequences were used to analyze the distribution and abundance of *mcr* genes. The abundance of an *mcr* gene was calculated as the fragment count of that gene over the total amount of bacterial fragments for a sample or a group, whereas the fragment count for ARGs was the only one used for statistical analyses.

Out of the 214,095 metagenomes, we found that 2.09% (4,465) of them contained read fragments aligning to at least 1 of the 9 *mcr* gene variants. The average number of reads per *mcr*-positive sample was 27 million reads, and on average, 0.003% of the reads were aligned to *mcr* genes. Among the variants in the *mcr* family, *mcr-1* and *mcr-9* were the most frequent, with 25.91% and 57.47% of the *mcr*-aligned reads aligning to these variants, and disseminated across 10 and 13 sampling years, 21 and 56 countries, and 23 and 61 hosts, respectively. The rarest variants were *mcr-2*, *mcr-6*, and *mcr-8* with read frequencies of 0.03%, 0.01%, and 0.08%, respectively, and their metagenomic origins were more restricted (Table 1). Overall, different log-ratio abundance levels seemed to be different across the sampling years in different countries and hosts (Fig. S1).

**Level of *mcr* variants over time.** The *mcr*-positive metagenomic samples were collected between 2003 and 2019, with the exception of 2005, in which no *mcr* fragments were detected (Fig. 1). Only two metagenomes sampled in 2003 contained *mcr* fragments, and a single *mcr*-positive metagenome was from 2004. Onward, the percentage of positive samples fluctuated, with the lowest value of 0.5% in 2008 and the highest of 6.4% in 2019 (Fig. S1). All the variants were frequently found in samples from 2016 to 2017, except *mcr-6*, which was only found in 2012 (Table 1).

We found that the log ratio abundance of aligned read fragments fluctuated for the nine variants in each sampling year (Fig. 1). The oldest positive metagenomes were sampled in 2003 and 2004 and contained only *mcr-3* and *mcr-5*. From 2006, the other variants began to emerge. *mcr-1* was detected first in 2009 at a low log abundance, and increased levels were observed between 2011 and 2019. Similarly, *mcr-9* could be detected in small amounts in metagenomes from 2007. In 2012 and 2013, *mcr-9* was



Downloaded from https://journals.asm.org/journal/msystems on 29 July 2022 by 192.38.90.185.

**TABLE 1** Read alignment of each *mcr* variant across different sample types

| | *mcr-1* | *mcr-2* | *mcr-3* | *mcr-4* | *mcr-5* | *mcr-6* | *mcr-7* | *mcr-8* | *mcr-9* |
|---|---|---|---|---|---|---|---|---|---|
| Origin[a] | | | | | | | | | |
| Read frequency (%) | 25.91 | 0.03 | 10.33 | 0.98 | 1.86 | 0.01 | 3.32 | 0.08 | 57.47 |
| No. of yrs | 10 | 6 | 14 | 13 | 13 | 1 | 13 | 11 | 13 |
| No. of countries | 21 | 6 | 59 | 42 | 43 | 1 | 27 | 14 | 56 |
| No. of hosts/reservoirs | 22 | 6 | 49 | 54 | 40 | 1 | 43 | 8 | 60 |
| Yr[b] | | | | | | | | | |
| 2010 | | 16.67 | | | | | | | |
| 2012 | | | | | 9.03 | 100.00 | | 20.00 | |
| 2013 | | | | 17.04 | | | | | |
| 2014 | | | | | | | | | 23.89 |
| 2015 | 37.66 | | | | | | | | |
| 2016 | | | 58.77 | 20.89 | 61.11 | | 30.68 | 16.00 | 22.11 |
| 2017 | 22.86 | 16.67 | 14.36 | | | | 17.40 | | |
| Country[c] | | | | | | | | | |
| Angola | | 16.67 | | | | | | | |
| Cambodia | 6.36 | | | | | | | | |
| China | 68.96 | | | | | | 21.9 | | |
| Denmark | | | 40.11 | 12.16 | 36.24 | | | | |
| France | | | | | | 100.0 | | | |
| Kenya | | | | | | | | 11.10 | |
| Netherlands | | 16.67 | | | | | | | |
| USA | | | 13.43 | 39.38 | 18.78 | | 23.63 | 11.11 | 45.18 |
| Host/reservoir[d] | | | | | | | | | |
| *Homo sapiens* | 32.51 | 33.33 | | 23.71 | | | | 19.35 | 56.46 |
| Panda | 22.59 | | | | | | | | |
| Activated sludge metagenome | | | | | 5.93 | | | | |
| Freshwater metagenome | | | 10.16 | | | | 18.24 | | |
| Marine metagenome | | 22.22 | | | | | | 48.39 | |
| Microbial mat metagenome | | | | | | 100.0 | | | |
| Wastewater metagenome | | | 65.77 | 26.20 | 60.81 | | 25.88 | | 13.44 |

[a]Read frequencies and counts of unique sampling origins, i.e., the number of years, countries, and hosts/reservoirs.
[b]The top two sampling years for the given variant was the most abundant in abundant in is shown in percentage of *mcr*-mapped reads.
[c]The top two sampling countries as described in footnote b.
[d]The top two sampling hosts and reservoirs as described in footnote b.

the most abundant variant, with 81% and 86% of the read fragments aligning to this gene. In 2007, only 3% of the *mcr* read fragments aligned to *mcr-7*, but more and more fragments for each year were assigned to the *mcr-7* gene and peaked in 2019 with 95% of the *mcr* fragments aligned being to it.

Significant levels of different *mcr* genes were observed for sampling years 2011, 2013, 2014, 2015, 2016, and 2017 ($P$ value $< 0.05$, Fig. 4a). Even though the variance of *mcr* levels within the sampling years was high, several variants stood out as having higher or lower levels in specific years compared to other years. In 2011, *mcr-3* had a higher abundance than expected and continued to be high in 2013 to 2014, together with *mcr-1* and *mcr-5*. *mcr-9* was lower in those years. In 2016, the metagenomic picture changed as *mcr-3* and *mcr-5* had decreased levels, while *mcr-1*, *mcr-4*, *mcr-7*, and *mcr-9* were increased.

**Geographical distribution of *mcr* gene variants.** The 9 variants were spread across 95 different sampling locations (Fig. S1), although samples from different world regions were often different in which variant they were positive for (Table 1). A higher abundance of *mcr* gene variants was observed in the Americas, Asia, and Europe, and decreased abundances were observed in Africa. The highest total log-ratio abundances of *mcr* fragments could be found in metagenomes from Australia, Lake Huron (USA), and Cambodia, and lowest levels, in Kiribati, Greece, and the Caribbean Sea (Fig. 2).

The individual variants were not equally distributed worldwide; instead, it seemed like specific variants were restricted to specific regions (Fig. 2). For example, the variant *mcr-1* was less widely spread worldwide (Americas, Asia, and Europe) than *mcr-9*



Downloaded from https://journals.asm.org/journal/msystems on 29 July 2022 by 192.38.90.185.

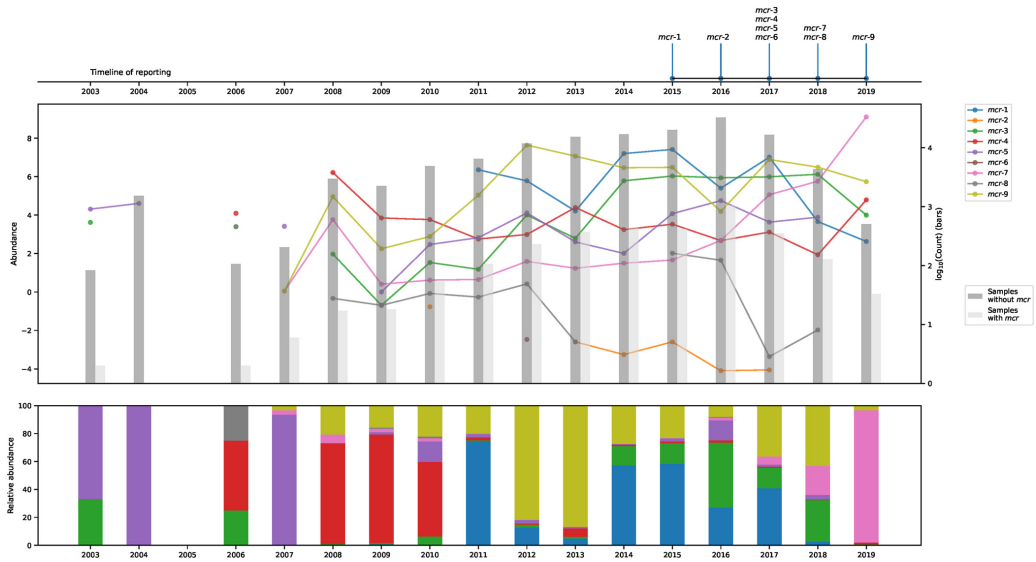



**FIG 1** Discovery and the change of *mcr* genes over time. (Top) Timeline showing when each gene was first reported in the literature. (Middle) Changes in log abundance of aligned *mcr* read fragments over time for each gene are shown, as well as the number of samples with or without an *mcr* hit from each year as bars. (Bottom) The frequency of each gene compared to the total *mcr* amount. Data were normalized with gene lengths to generate the charts.

(Africa, Americas, Asia, Europe, Oceania, Atlantic Ocean, and the Pacific Ocean). No metagenomic locations contained all types of variants. In the Australian metagenomes, *mcr-9* was the most dominant gene, whereas *mcr-4* had high abundance levels in Lake Huron (USA), and *mcr-1* and *mcr-9* had high abundance levels in Cambodian metagenomes. The only location of *mcr-6* was France.

Of the 95 sampling locations, 15 had significant abundances of at least one of the *mcr* variants ($P$ value $< 0.05$); however, the variance in the samples from most of the locations was high and did not have a large effect size compared to other locations (Fig. 4c). Metagenomic locations that showed consistency within the group and were found to be different from the rest of the locations had lower levels of single variants —*mcr-1* in Bulgaria, *mcr-3* in Iceland, *mcr-5* in Malaysia, and *mcr-9* in Cambodia.

**Host- and reservoir-specific *mcr* abundances.** We found *mcr* genes present in 125 different sampling hosts and reservoirs, but with the various variants having different log-ratio fragment abundances (Fig. S2) and the two most frequent types differing for each variant (Table 1). All 6 metagenomes from *Pomacea canaliculate* (golden apple snail) and the 11 *Danio rerio* (zebrafish) metagenomes contained *mcr* fragments. For two of the largest sampling groups, we found 897 out of 1,803 (49.75%) wastewater metagenomes and 13,831 out of 102,211 (1.35%) human-derived samples to be *mcr* positive.

Out of the 125 hosts, only 20 of them showed significant levels of *mcr* gene fragments. These all had higher levels of colistin resistance genes (Fig. 3). The dispersion within most of the 20 hosts was high, and their log-ratio levels were not significantly different from those of the other hosts, except a few (Fig. 4e). The zebrafish samples had lower levels of *mcr-7* than expected, whereas golden apple snail metagenomes had higher levels. Panda metagenomes had elevated levels of *mcr-9* but had slightly smaller amounts of *mcr-3*, *mcr-4*, and *mcr-5*. Metagenomes from pigs (*Sus scrofa* and pig gut) had increased levels of both *mcr-1* and *mcr-9*. Human metagenomes did not have large effect sizes but contained slightly less *mcr-1* and *mcr-9* than expected.

**Diversity of *mcr*-positive metagenomes.** By performing compositional PCA analysis on CLR values, we can visualize the variance in *mcr* read proportions in biplots

Downloaded from https://journals.asm.org/journal/msystems on 29 July 2022 by 192.38.90.185.

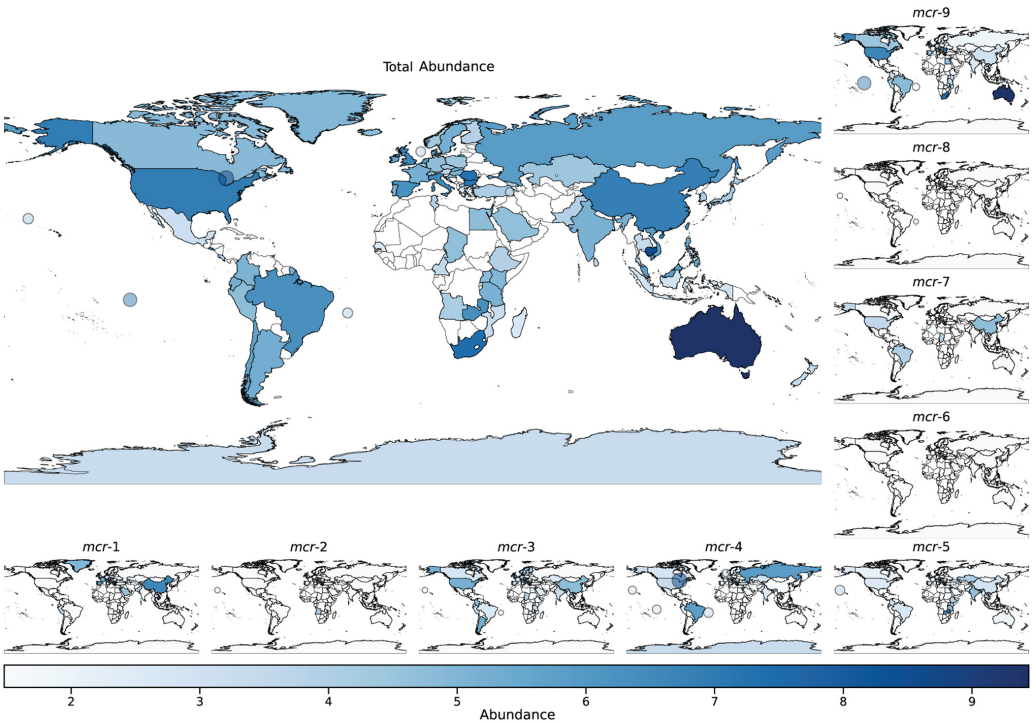



**FIG 2** Global levels of *mcr* genes. The large map shows the total log abundance levels of all *mcr* genes, whereas the nine smaller plots show the individual gene log-scaled abundances worldwide. A circle represents a collection of samples from water containing *mcr* genes. White color indicates an absence of results, not that a specific location does not have any *mcr* genes. The circle markers illustrate water environments.

showing which type of samples make the level of a resistance gene significant (Fig. 4). The biplots highlight a clear separation of metagenomes that contain *mcr-1* or *mcr-9* and show that these samples also differ a lot from each other. None of the samples from the different years are similar, which means that high levels of one of the variants cannot be explained simply due to a specific collection year (Fig. 4b). Instead, we can see that several Panda metagenomes came from China in 2016, which most likely contributed to the higher levels of *mcr-1* in 2016 (Fig. 4d). Likewise, human metagenomes clearly show a geographical separation mainly driven by *mcr-1* being abundant in China and *mcr-9* in the United States and Australia (Fig. 4f, Fig. S3), which could explain that even if these metagenomes contain significant levels of *mcr* genes, we could not observe large effect sizes. Excluding these two most abundant genes suggests, however, that the differences are mainly driven by source and not by year or geographical location (Fig. S4).

**Distribution of *mcr* variants in pathogenic bacterial genomes.** As several studies have performed retrospective screening of pathogenic bacterial isolates, we decided to compare the metagenomic *mcr* abundances to the prevalence in pathogenic single isolates. Out of 912,469 isolates screened by the NCBI Pathogen Detection Pipeline, only 7,934 (0.87%) were shown to carry at least 1 of the *mcr* genes. The majority of the *mcr*-positive isolates contained either *mcr-1* (51.08%) or *mcr-9* (40.38%), while *mcr-6* and *mcr-7* were not detected at all (Table S1).

The congruence of relative counts in isolates and relative abundance levels in metagenomes varied depending on which allele and what kind of sample grouping it



Downloaded from https://journals.asm.org/journal/msystems on 29 July 2022 by 192.38.90.185.

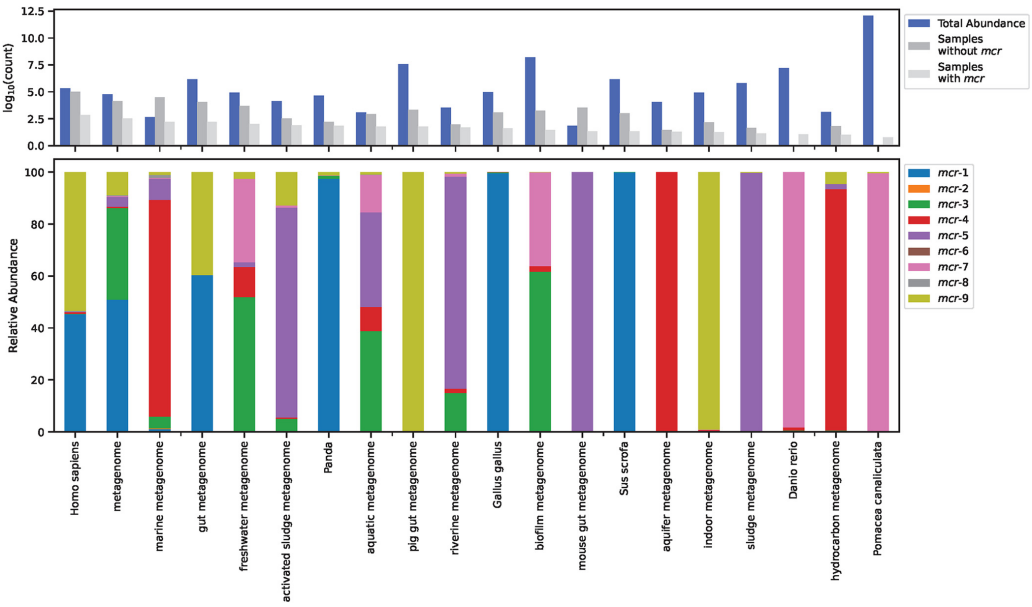



**FIG 3** Distribution of *mcr* genes in selected sampling hosts. Hosts were selected based on the host showing significant CLR values according to the ALDEx2 analysis (Fig. 4). (Top) Bar plot showing both the number of samples without and with an *mcr* hit and the overall *mcr* level for each host measured by log-abundance values. (Bottom) The abundance of individual *mcr* genes relative to total *mcr* levels. Data were normalized with gene lengths before plotting. To see the distribution of *mcr* genes for all sampling hosts available, refer to Fig. S2.

was. Grouped by the sampling location, the *mcr-1* gene appeared to be more widespread according to the isolates, whereas *mcr-3* had a larger global distribution based on the metagenomes. Similarly, for human samples, *mcr-1* had a higher prevalence in isolates, whereas metagenomes showed a higher abundance of *mcr-9*-aligned read fragments (Fig. S5).

**Genomic background of *mcr* genes.** The dissemination of colistin resistance genes between different reservoirs and countries in different sampling years was further investigated by creating assemblies of metagenomic samples with 95% coverage of at least one variant. We assembled 869 metagenomes, where we found 1,939 different contigs carrying *mcr* genes (range, 1 to 20 *mcr* contigs per metagenome). The most frequent gene present on these contigs was *mcr-9*, followed by *mcr-3* and *mcr-5* (Fig. S6a). To identify structural patterns between different metagenomic origins, we analyzed the genetic signatures in regions up- and downstream of an *mcr* gene (the flanks) with a minimum size of 1,000 bp and a maximum of 21,000 bp to include most of the elements found in the flanks (Fig. S6b). As most contigs were shorter than 1,000 bp (Fig. S6a), only 138 contigs passed the size criteria. All 20 contigs containing plasmid replicons in their flanks carried *mcr-1* genes, whereas the 63 flanks with MGEs were on contigs with different *mcr* variants (Fig. S6c).

Six distinct clusters became apparent upon calculating the distance between the flanks surrounding the *mcr* genes (Fig. 5). We find that the presence of specific MGEs seemingly correlated with the presence of an *mcr* variant on the contig, as ISA*pl1* occurs only on *mcr-1* contigs, and IS*903*, on *mcr-9* contigs. Five of the six clusters are all flanks around the same variant, with two being *mcr-1* clusters, while the sixth contains flanks surrounding four different *mcr* gene variants.

In the first *mcr*-1 cluster, an IncX4 plasmid replicon was present either upstream or downstream in most members. These contigs were found in metagenomes sampled in

Downloaded from https://journals.asm.org/journal/msystems on 29 July 2022 by 192.38.90.185.

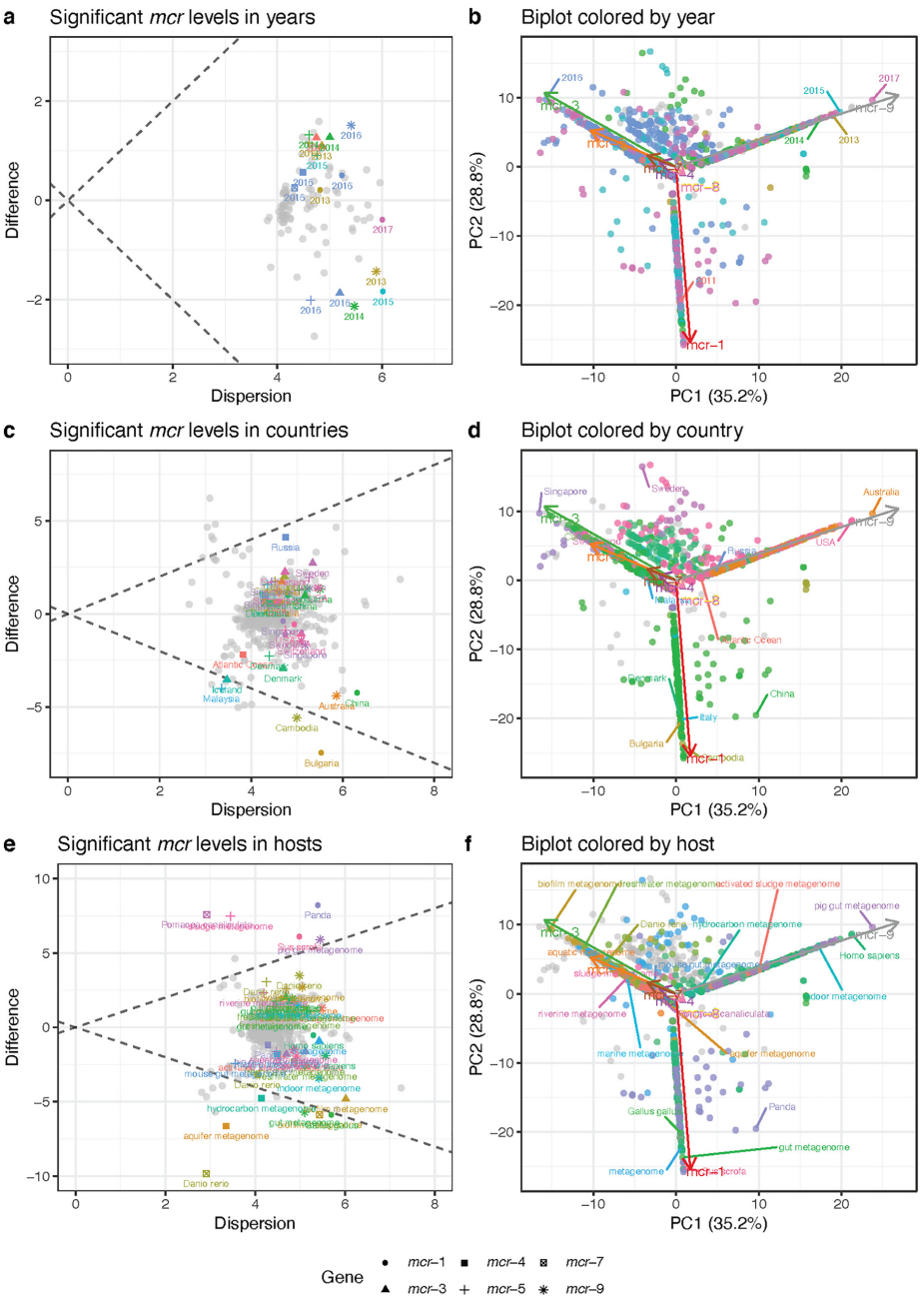



**FIG 4** Analysis of significant *mcr* levels in sampling years, countries, and hosts. (a, c, and e, left column) Visualizations of within-group dispersion of CLR values of individual *mcr* genes compared to the between-group difference in CLR values for (a) sampling

(Continued on next page)



Downloaded from https://journals.asm.org/journal/msystems on 29 July 2022 by 192.38.90.185.

**FIG 5** Clustering of *mcr* contigs reveals that the genomic context remains conserved. The k-mer distance tree for flanks of *mcr* contigs with flank sizes between 1,000 bp and 21,000 bp is drawn in the left panel, with the metagenomic origins (year, country, host) added as colored tiles in the middle panel. The genomic background on the right is a schematic illustration of the size of each flank region in gray centered on the middle of the *mcr* gene and plasmid replicons and mobile genetic elements (MGEs) colored in the flanks.

2016 and 2017 from a diverse background, indicating that IncX4 plasmids are involved in multiple transmission events in different settings. The second *mcr-1* cluster differs from the first in that we see an absence of IncX4 and IncI2 in a few contigs instead. The cluster can instead be best characterized by the presence of the insertion element IS*Apl1* in half of the flanks, which mainly originate from gut metagenomes from Cambodia in 2014, while those without the insertion element are from Chinese samples from 2015 to 2017.

Four of the *mcr*-9 contigs contain IS*102*, IS*26*, or IS*Kpn43*, while the remaining carry IS*903*. Since all these contigs contain at least one insertion sequence, it suggests that they were highly mobilized between different metagenomes from human and environmental sources between 2008 and 2018.

The mixed cluster shows a surprising clustering of flanks on contigs of uneven lengths, with different MGEs and replicons present, carrying four different *mcr* gene variants—*mcr-1*, *mcr-4*, *mcr-8*, and *mcr-9*. This indicates that despite the contigs carrying different *mcr* genes, there are similarities in their broader genomic context, despite it not being obvious how they are connected considering their various sample types.

**Metagenomic evidence of new *mcr* subvariants.** The varied origins of the collected metagenomes can be used to investigate how conserved known *mcr* gene variants are in different sources, as well as provide evidence of the presence of new *mcr* subvariants. Overall, most of the *mcr* reference sequences could be recovered from the metagenomic samples, although a large proportion seems only to be fragmented

**FIG 4** Legend (Continued)
year, (c) location, and (e) host. (b, d, and f, right column) Compositional biplots of the first two principal components (PC) capturing 64% of the variation in the data set, where samples are colored according to significant (b) years, (d) countries, and (f) hosts. Gray filled markers are samples that were nonsignificant. CLR: centered-log ratio transformed values of the proportion of *mcr* aligned read fragments.



Downloaded from https://journals.asm.org/journal/msystems on 29 July 2022 by 192.38.90.185.

**TABLE 2** Coverage of mcr templates according to KMA[a]

| | No. of | | Template coverage (%) | | | | Depth of coverage (×) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Gene | Samples | Known subvariants | Avg. | Min. | Mdn. | Max. | Avg. | Min. | Mdn. | Max. |
| *mcr-1* | 418 | 12/14 | 71.238 | 1.05 | 85.980 | 100.18 | 43.877 | 0.01 | 2.350 | 1315.16 |
| *mcr-2* | 12 | 2/2 | 34.694 | 1.05 | 1.240 | 100 | 2.435 | 0.01 | 0.060 | 11.27 |
| *mcr-3* | 1,204 | 25/25 | 37.170 | 1.05 | 30.595 | 100.8 | 2.822 | 0.01 | 0.580 | 209.51 |
| *mcr-4* | 565 | 5/6 | 35.112 | 1.11 | 21.590 | 100 | 1.684 | 0.01 | 0.350 | 67.77 |
| *mcr-5* | 1,100 | 2/2 | 40.030 | 1.1 | 31.990 | 100 | 1.705 | 0.01 | 0.490 | 118.32 |
| *mcr-6* | 4 | 1/1 | 64.982 | 1.24 | 84.755 | 89.18 | 3.177 | 0.01 | 2.335 | 8.03 |
| *mcr-7* | 384 | 1/1 | 28.138 | 1.17 | 6.665 | 100.06 | 8.103 | 0.01 | 0.120 | 209.91 |
| *mcr-8* | 32 | 1/1 | 18.987 | 1.06 | 1.325 | 100 | 1.919 | 0.01 | 0.010 | 30.31 |
| *mcr-9* | 2,148 | 1/1 | 50.690 | 1.17 | 39.570 | 102.53 | 23.161 | 0.01 | 0.690 | 3,985.21 |

[a]The table contains an overview of the found number of *mcr* subvariants out of how many were known in the metagenomic samples, as well as summary statistics of template coverage and depth of coverages. Avg., average; Min., minimum; Mdn., median; Max., maximum.

sequences (Table 2). We constructed consensus sequences that had at least 90% template coverage, mean coverage depth of 5, and query identity of at least 90% and kept single nucleotide polymorphisms (SNPs) that had a minimum depth of 5 and 90% frequency. Of the 968 sequences constructed, 27.38% had at least one SNP difference in their template (Table 3). The majority of consensus sequences recovered from the metagenomes, whether they were known or new potential subvariants, could only be recovered in a few samples. Although, there are a few groups that stand out. We found 33 different subvariants of *mcr-3* genes, 32 of *mcr-7.1*, and a highly prevalent subvariant of *mcr-9.1* (Table 3, Fig. 6). Since these sequences were constructed from metagenomic samples with KMA, we call SNP variants for potential new subvariants.

The number of *mcr-3* subvariants in our version of the ResFinder database is 25, making it the variant with most subvariants. We found evidence of 33 new subvariants, though most only appear in a small number of samples, except for a subvariant of *mcr-3.6* called *mcr-3.6.v1* (Fig. S7a) and a subvariant of *mcr-3.15* called *mcr-3.15.v2* (Fig. S7b). Both *mcr-3.6.v1* and *mcr-3.15.v2* were detected in genomes of *Aeromonas* species (Table S2).

None of the *mcr-7.1* constructed sequences was an exact match to the reference sequence, and instead, we saw various numbers of SNPs (Fig. 6, Fig. S8a). While the frequencies of each new possible *mcr-7.1* variant in the metagenomes were not high, there appear to be several SNPs that were well conserved, for example, the two SNPs A1020G and A1275T present in 29 and 30 of the variants, respectively (Fig. S8a). Many of the 32 possible subvariants of *mcr-7.1* were found in water sources (e.g., zebrafish, freshwater, and wastewater) sampled over a period of 4 years (2016 to 2019) (Fig. 6). Unfortunately, none of the possible *mcr-7.1* subvariants had complete BLAST matches (Table S2).

We saw a high occurrence of a new subvariant sequence of *mcr-9.1*, named *mcr-9.1.v4*, which contained two SNPs, A1619G and A1620G (Fig. S8b). *mcr-9.1.v4* appears to originate in human or gut samples, a similar distribution to that of the template

**TABLE 3** Overview of SNP variant calling on consensus sequences

| Gene[a] | Total sequences[b] | SNP variants (%) | Unique SNP subvariant sequences[c] |
|---|---|---|---|
| *mcr-1* | 170 | 1.18 | 2 |
| *mcr-2* | 4 | 100.00 | 4 |
| *mcr-3* | 127 | 47.20 | 33 |
| *mcr-4* | 33 | 36.40 | 5 |
| *mcr-5* | 58 | 1.72 | 1 |
| *mcr-6* | 0 | 0 | 0 |
| *mcr-7* | 39 | 100.00 | 32 |
| *mcr-8* | 3 | 66.70 | 2 |
| *mcr-9* | 534 | 27.20 | 6 |
| Total | 968 | 27.38 | 62 |

[a]The number of consensus sequences per *mcr* gene.
[b]The percentage of consensus sequences found that are SNP variants.
[c]The number of unique SNP variant sequences recovered.

Downloaded from https://journals.asm.org/journal/msystems on 29 July 2022 by 192.38.90.185.

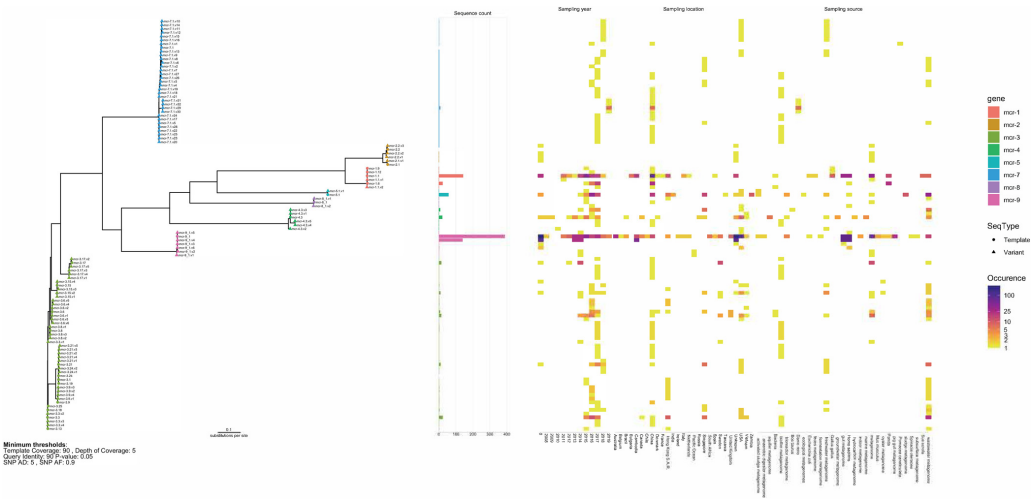



**FIG 6** Phylogenetic tree of consensus sequences. Sequences were aligned with MAFFT and clustered with FastTree. On the right is the occurrence of each sequence variant in different sampling origins (year, location, and source).

*mcr-9.1* (Fig. 6). *mcr-9.1.v4* had 20 BLAST hits, where the two most common species carrying this gene were *Enterobacter hormaechei* and *Salmonella enterica* subsp. *enterica* (Table S2).

## DISCUSSION

The growing collection of metagenomic sequencing data available in public repositories has the potential to provide a much more detailed picture of the emergence, evolution, and spread of ARGs. We downloaded and analyzed 214,095 host-derived and environmental metagenomes to characterize the global distribution of 9 mobilized colistin resistance genes that have been identified since 2015. Among the downloaded samples, we found that 4,465 metagenomes (2%) contained *mcr* reads distributed differently across the sampling period (2003 to 2019) and geographical and host origins. We found that all the nine different gene variants were present in metagenomes sampled years before their discovery (Fig. 1), confirming the notion of the resistance genes circulating in the environment long before being reported (18, 19, 23). This confirms the value of publicly sharing raw next-generation sequencing (NGS) data to promote new, better, and more comprehensive analyses of existing data.

To date, *mcr-1* is the most studied *mcr* gene variant, and the dissemination has been described in detail. Multiple studies agree that even though *mcr-1* has been detected in a few isolates from the 1980s (18), it has been appearing with increasing frequencies in samples between 2011 and 2017 and was decreasing in later years (8, 18, 19, 24). We see a similar trend in the frequencies of *mcr-1*-positive metagenomes, although the levels seem to increase starting from 2008 to the highest levels in 2015, the year of *mcr-1* first being reported (4) (Fig. 1).

In just a few years after discovering *mcr-1*, multiple other *mcr* variants were reported in different world regions, with *mcr-1* and *mcr-9* being the most disseminated genes (6). Despite *mcr-9* being the newest member (17), we observe that it was the most abundant gene variant in publicly available metagenomes, with *mcr-1* being the second most abundant. The two variants are not equally distributed across sampling sites, as *mcr-1* appears to be more geographically restricted to Europe and Asia, whereas *mcr-9* has reached a wider area (Fig. 2). In human metagenomes, *mcr-1* and *mcr-9* dominate, whereas other hosts and environmental origins display a considerable



Downloaded from https://journals.asm.org/journal/msystems on 29 July 2022 by 192.38.90.185.

variation in *mcr* variants. Despite earlier reports of the presence of *mcr-1* of both animal (4) and environmental (25) origins, we see only very few environmental origins of the gene (Fig. S2). Only a few hosts have significantly different levels of *mcr* gene abundances than expected, where *mcr-1* and *mcr-9* tend to be higher in pigs and pandas, and *mcr-3*, *mcr-4*, *mcr-5*, and *mcr-7* are lower in other hosts. *Mcr-2*, *mcr-6*, and *mcr-8* only appear in very few metagenomes, with *mcr-6* being the rarest variant. Since the first report of *mcr-6* (14), it has only been detected in very few places around the world and all in 2014 and 2015 (14, 26), but we can here report the presence of *mcr-6* in very small amounts in a metagenome from France sampled in 2012 (Fig. 1 and 2). Overall, there appears to be a connection between the abundance of a variant and the sampling source and location, but due to the sparse nature of our data set, we have not been able to determine the relative contribution of these factors to the observed *mcr* levels.

When observing the trends of the aligned *mcr* read fragment abundances in the data set, one should keep in mind that this collection is restrained by what was available in ENA at the time of download. The type of metagenomic data sets available is dependent on the ongoing research trends in the different scientific communities, which can cause a bias toward specific hosts or environments, such as the panda samples, by being overrepresented in the repository. Furthermore, there are challenges due to improved experimental protocols and sequencing platforms becoming available, possibly causing mapping bias. On the other hand, the evidence of the number of read fragments that match a specific gene should not be discarded too easily regardless of the sample origin. We applied compositional methods that can handle the nature of various read counts to ensure that the observed abundance levels of the different *mcr* alleles were not simply due to chance.

The NCBI Pathogen Detection Project is another example of a surveillance program that routinely screens available public data, in this case, genomes of single isolates. This data collection also has the same biases as those highlighted for our metagenomic collection, where our comparison of the two resources showed that each resource is in some cases better at capturing the prevalence of specific *mcr* alleles than the other. Essentially, our study highlights the benefit of using metagenomic data sets in addition to single isolates to monitor the distribution of AMR.

Interestingly, we observed that the *mcr* contigs from the assembled metagenomes were well conserved across reservoirs and locations except for *mcr-1* contigs. This suggests that most of the *mcr* alleles have only been mobilized once and then spread globally and between reservoirs. In contrast, *mcr-1* is known to be present in a variety of genomic backgrounds (8), which we also observe as the flanking regions of our *mcr-1* contigs grouped together in three distinct clusters (Fig. 5). *Mcr-1* is commonly found on IncI2, IncHI2, and IncX4 plasmids with ISA*pl1* (8, 25), although we only observed ISA*pl1* on two IncI2 plasmids, a possible loss of ISA*pl1* near IncX4 replicons, and we observed that no IncHi2 plasmids were present on *mcr-1* contigs. The absence of ISA*pl1* in one of the *mcr-1* clusters could indicate a loss of mobility due to either their difference in sampling years or a shift in hosts. IS*26* has been observed downstream of *mcr-9* (27), which we only observed once, and instead, we see that IS*903* occurs on both sides of *mcr-9* in the examined contigs. The metagenomic origin of *mcr-9* contigs is highly diverse, suggesting that the presence of multiple different insertion sequences has been a contributing factor in their mobilization between 2008 and 2018.

Even with the diverse genomic context of *mcr-1*, only very few of the *mcr-1* consensus sequences we constructed contained any SNPs, indicating that despite the different mobilization factors, the different *mcr-1* subvariants remain well conserved (Fig. 6). On the contrary, the sequences of *mcr-3* subvariants were highly prone to contain SNPs, as shown by our report of 33 potential new members, where several of them could be matched to genes in known species with BLAST (Table S2). Similarly, the diverse origin of *mcr-9* contigs is also reflected in the fact that an unknown subvariant



Downloaded from https://journals.asm.org/journal/msystems on 29 July 2022 by 192.38.90.185.

of *mcr-9.1*, which we are calling *mcr-9.1.v4*, was detected in 100 different genomes, with the species *Enterobacter hormaechei* being the most common (Table S2). We hesitate to call these SNP variants new variants, as more work needs to be done to test the expression levels and susceptibility of the organism carrying one of these potential subvariants, although there is strong evidence for the *mcr-9.1.v4* variant already being widely distributed.

As we collected data by downloading publicly available metagenomic samples, we present a data set with uneven coverage of sampling locations and sources. This bias heavily influences our ability to provide an in-depth understanding of the mobilization, emergence, and spread of the *mcr* genes. Regardless of this, we have shown the potential of using raw sequencing reads generated by other researchers to improve our knowledge. It is, however, important that all such generated data are shared publicly to allow for future exploration and improved understanding of the global microbial biology (28). Since the start of this project, another *mcr* variant was discovered named *mcr-10* (29), but we decided not to include the gene due to the massive computation task of mapping 442 Tb of raw sequencing reads. Nevertheless, it will be indeed interesting to figure out when *mcr-10* first appeared and characterize its dissemination as well, which we hope to do for this and for other ARGs in the future.

## MATERIALS AND METHODS

**Data collection.** Metagenomic data sets were collected from the public data repository the European Nucleotide Archive (ENA) (30). We queried the ENA API for samples uploaded between 1 January 2010 and 1 January 2020 that were shotgun sequenced and had at least 100,000 sequencing reads. In total, we downloaded 214,095 sequencing runs from 146,732 samples from 6,307 projects corresponding to 442 Tbp of raw reads.

**Reference sequence databases.** The AMR gene database ResFinder (31) (downloaded 25 January 2020) contains 3,085 sequences. The 16S rRNA SILVA (32) gene database (version 138, downloaded 16 January 2020) contains 2,225,272 sequences.

**Preprocessing and mapping sequencing reads.** The raw FASTQ reads were quality checked using FastQC v.0.11.15 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimmed with BBduk2 36.49 (33) to remove low-quality sequences and adaptors. BBduk2 settings were set as follows: minimum read length set to 50 bp, k=19, kmin=11, tbo flag on 11, the Phred quality threshold at 20 (99% accuracy), and only right trimming (ktrim=r). Assignment of trimmed reads to reference sequences was done with global alignment using KMA 1.2.21 (34) with the following alignment parameters: 1, -2, -3, -1 for a match, mismatch, gap opening, and gap extension. Also, a value of 7 for read pairing and a minimum relative alignment score of 0.75 were used. We used ResFinder to assess the number of acquired AMR reads in each sample and Silva to determine the bacterial content. On average, it took 5.7 s per metagenome for ResFinder mapping and 232.7 s for Silva mapping on a node equipped with dual 20 core Xeon Gold 6230 CPUs clocked at 2.1 Ghz using the Danish National Supercomputer for Life Sciences (https://www.computerome.dk).

**Compositional data analysis.** The collected metagenomic data have large variability in how the samples were collected, how DNA was extracted, and how it was sequenced. Furthermore, the probability of observing a gene also depends on the sequencing depth. To account for some of the variability, we use read fragment counts as the gene counts for mapping against ResFinder genes, and they were adjusted by individual gene lengths. Bacterial 16S read fragment counts from Silva mapping were aggregated to a total sum for each sample and divided by a million.

Abundance tables of *mcr* genes were created by transforming the composition $x$ of *mcr-1 to mcr-9* length-adjusted counts $n_i$ ($i = 1 \ldots 9$) and the summed per million bacterial component $n_B$ by using the bacterial component as the reference and log-transforming the ratios:

$$x = [n_1, n_2, \ldots, n_9, n_B]$$

$$\text{Abundance}(x) = \left(\log\frac{n_1}{n_B}, \log\frac{n_2}{n_B}, \ldots, \log\frac{n_9}{n_B}\right)$$

For the statistical analysis performed on the mapping results, we treated the *mcr* read fragment counts as compositional. If we do not consider the observed counts as being relative to each sample, statistical tests can produce faulty results. Instead, if we apply the methods of compositional data analysis, this is avoided. As proposed by Aitchison (35), we log-ratio transform the counts to make the data symmetric, linear, and in a log-ratio coordinate space.

However, before applying log-transformations, counts of zero needed to be treated. Since a zero does not necessarily mean that a gene is absent from a sample and the logarithm of zero is an undefined value, we infer the proportion $p_i$ of reads of an ARG $i$ within a sequenced sample directly from the observed read count $n_i$. If we assume that each $n_i$ was sampled from a Poisson process, $n_i \sim \text{Poisson}(\lambda_i)$,



Downloaded from https://journals.asm.org/journal/msystems on 29 July 2022 by 192.38.90.185.

and the vector of counts follows a multinomial distribution $\{[n_1, n_2, \ldots] \mid n\} \sim \text{Multinomial}(p_1, p_2, \ldots \mid n)$, where $n = \sum_i n_i$ and $p_i = \frac{\lambda_i}{\sum_k \lambda_k}$. The posterior distribution of $[p_1, p_1, \ldots$ is given as the product of the multinomial likelihood with a Dirichlet$\left(\frac{1}{2}, \frac{1}{2}, \ldots\right)$ prior. These inferred proportions will never be precisely zero, even if the observed count is zero because of the multivariate distribution (36).

We used the centered log-ratio (CLR) transformation on the zero-replaced composition consisting of *mcr* read proportions *p*, excluding the bacterial component:

$$p = [p_1, p_2, \ldots, p_9]$$

$$\text{CLR}(p) = \left(\log \frac{p_1}{g_m(p)}, \ldots, \log \frac{p_9}{g_m(p)}\right)$$

where $g_m(p) = \left(\prod_{i=1}^{D} p_i\right)^{\frac{1}{D}}$, $D = 9$ is the geometric mean of the composition. The CLR values were used as the input for differential abundance tests and principal-component analysis, as described below.

**Data visualization.** Graphics visualizing abundance and relative abundances were created with Python 3.8 with Matplotlib 3.3.2 (37) and seaborn 0.11.0 (38). Bar plots showing relative abundances were created by closing the composition to 100. Geographical maps showing gene abundances were created using Shapely 1 3166.7.1 (39) and Cartopy 0.18.0 (40) to translate labels from the metadata into geographical shapes with the Natural Earth data set.

**Statistical analysis.** We carried out a differential abundance analysis on samples containing *mcr* fragments with ALDEx2 (41) 1.18.0 in R. We aimed to identify which experimental groups showed a difference in their abundance of *mcr* gene read fragments compared to other groups. ALDEx2 tests for significant differences of CLR abundance between categorical sample groups used Welch's *t* test followed by a Benjamini-Hochberg false-discovery rate (FDR) correction (42). We report significant groups of either sample locations, host, or collection year where the FDR is $<0.05$ and differential abundances were represented in an effect plot (43) displaying the within- and between-group variation in CLR values.

Principal-component analysis (PCA) was applied to the centered, scaled by total variance, and CLR transformed data set of *mcr* read proportions (44) to reduce the dimensionality of the data. The eigenvectors and eigenvalues from PCA were used to create a biplot, highlighting the significant sample groups found in the differential abundance analysis.

**Comparison of metagenomic abundance levels to prevalence in pathogen isolates.** The NCBI Pathogen Detection Project (45) routinely screens new isolates to identify AMR genes with the tool AMRFinderPlus (46), which reports whether a gene was found or not in an assembled genome. We downloaded the annotation results of 912,469 assembled genomes from NCBI's Pathogen Detection Resource (https://ftp.ncbi.nlm.nih.gov/pathogen/Results/, accessed 5August 2021); 7,934 (0.87%) of the single isolates contained at least 1 of the 9 *mcr* variants. We reported the frequency of the number of isolates carrying each *mcr* variant. Furthermore, we grouped the isolates by either sampling year, location, or host and reported the relative count of each variant to the relative abundance levels in the metagenomes.

**Metagenomic assembly of *mcr* samples.** We assembled metagenomes where at least one of the *mcr* genes had a minimum coverage of 95% by trimmed reads, according to KMA. The trimmed reads were assembled with MetaSPAdes 3.14.0 (47) with at least 1.2 terabytes of memory, 40 threads per node, and a maximum runtime of 1 week. Out of 1,014 metagenomes, 145 were not assembled, as they did not complete within the chosen time frame of a week. Contigs carrying the nine different *mcr* gene variants were identified with blastn 11.0 (48) with a percentage identity of $\geq$95.

**Flank analysis of metagenomic assemblies.** The metagenomic contigs carrying *mcr* genes were used in the flank analysis. Flanks were created by masking the *mcr* gene in the contig and cutting out up- and downstream regions of increasing sizes between 1,000 bp and 30,000 bp by intervals of 1,000 bp with BEDTools (49). The presence of plasmids in the flanks was identified with PlasmidFinder 2.1 (50), and mobile elements, with MobileElementFinder 1.0.3 (51). The distance between the flanks was calculated as the Szymkiewicz-Simpson dissimilarity with KMA (34). Hierarchical clustering on the flank distances was done with Ward's method (52) to create a dendrogram plotted with ggtree 2.0.4 (53) and ggplot 3.3.3 (54). This approach is similar to the workflow of the tool Flanker by Matlock et al. (55), except that we cluster with KMA.

**Variant analysis of *mcr* genes.** We investigated the presence of SNPs in KMA-produced consensus sequences that matched the following minimum requirements: template coverage, $\geq$98%; depth of coverage, $\geq$5; query identity, $\geq$90%; and *P* value, $\leq$0.05. SNPs were kept if they passed the following filters, checked with bcftools 1.13 (56): a minimum allele depth of 5 (AD) and a minimum allele frequency of 0.90 (AF). Sequences were aligned with MAFFT v7.490 (57), and phylogenetic trees were created with FastTree v2.1.1 (58) using a nucleotide substitution model. Trees were visualized with ggtree. A visual summary of SNPs in sequence alignments was created with snipit (https://github.com/aineniamh/snipit).

We screened all unique sequences that had at least 1 SNP difference to their template against complete and draft genome sequences in GenBank with BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi, accessed 24 January 2022). Matches were identified if they had 100% identity to the template.

**Data availability.** Source data for generating abundance figures and running statistical tests and flank and variant analysis can be found in the supplementary files at https://doi.org/10.5281/zenodo.5946866, and the supporting code is available at https://github.com/hmmartiny/mcr_metagenomes.



Downloaded from https://journals.asm.org/journal/msystems on 29 July 2022 by 192.38.90.185.

## SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

**FIG S1**, TIF file, 2.9 MB.

**FIG S2**, TIF file, 0.6 MB.

**FIG S3**, TIF file, 0.3 MB.

**FIG S4**, TIF file, 0.6 MB.

**FIG S5**, TIF file, 2.6 MB.

**FIG S6**, TIF file, 0.7 MB.

**FIG S7**, TIF file, 0.4 MB.

**FIG S8**, TIF file, 0.7 MB.

**TABLE S1**, DOCX file, 0.01 MB.

**TABLE S2**, DOCX file, 0.02 MB.

## ACKNOWLEDGMENTS

This work was supported by the European Union's Horizon H2020 grant VEO (874735) and the Novo Nordisk Foundation (grant NNF16OC0021856: Global Surveillance of Antimicrobial Resistance).

F.M.A. and T.N.P. designed the project. H.-M.M. performed data acquisition, assembly, and analysis. J.S. and T.N.P. provided guidance with sequence downloading and mapping, P.M. and C.B. provided guidance with compositional analysis, P.M. provided guidance with flank analysis, and F.M.A., T.N.P., and P.M. provided guidance with SNP variant analysis. H.-M.M. wrote the draft. All authors have read, commented on, and accepted the final version.

## REFERENCES

1. WHO. 2014. Antimicrobial resistance: global report on surveillance. World Health Organization, Geneva, Switzerland.
2. Davies J, Davies D. 2010. Origins and evolution of antibiotic resistance. Microbiol Mol Biol Rev 74:417–433. https://doi.org/10.1128/MMBR.00016-10.
3. Falagas ME, Kasiakou SK. 2005. Colistin: the revival of polymyxins for the management of multidrug-resistant gram-negative bacterial infections. Clin Infect Dis 40:1333–1341. https://doi.org/10.1086/429323.
4. Liu Y-Y, Wang Y, Walsh TR, Yi L-X, Zhang R, Spencer J, Doi Y, Tian G, Dong B, Huang X, Yu L-F, Gu D, Ren H, Chen X, Lv L, He D, Zhou H, Liang Z, Liu J-H, Shen J. 2016. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. Lancet Infect Dis 16:161–168. https://doi.org/10.1016/S1473-3099(15)00424-7.
5. Gharaibeh MH, Shatnawi SQ. 2019. An overview of colistin resistance, mobilized colistin resistance genes dissemination, global responses, and the alternatives to colistin: a review. Vet World 12:1735–1746. https://doi.org/10.14202/vetworld.2019.1735-1746.
6. Ling Z, Yin W, Shen Z, Wang Y, Shen J, Walsh TR. 2020. Epidemiology of mobile colistin resistance genes mcr-1 to mcr-9. J Antimicrob Chemother 75:3087–3095. https://doi.org/10.1093/jac/dkaa205.
7. Cuadrat RRC, Sorokina M, Andrade BG, Goris T, Dávila AMR. 2020. Global ocean resistome revealed: exploring antibiotic resistance gene abundance and distribution in TARA Oceans samples. Gigascience 9:1–12. https://doi.org/10.1093/gigascience/giaa046.
8. Wang R, van Dorp L, Shaw LP, Bradley P, Wang Q, Wang X, Jin L, Zhang Q, Liu Y, Rieux A, Dorai-Schneiders T, Weinert LA, Iqbal Z, Didelot X, Wang H, Balloux F. 2018. The global distribution and spread of the mobilized colistin resistance gene mcr-1. Nat Commun 9:1179. https://doi.org/10.1038/s41467-018-03205-z.
9. Luo Q, Wang Y, Xiao Y. 2020. Prevalence and transmission of mobilized colistin resistance (mcr) gene in bacteria common to animals and humans. Biosaf Heal 2:71–78. https://doi.org/10.1016/j.bsheal.2020.05.001.
10. Xavier BB, Lammens C, Ruhal R, Butaye P, Goossens H, Malhotra-Kumar S. 2016. Identification of a novel plasmid-mediated colistin-resistance gene, mcr-2, in Escherichia coli, Belgium, June 2016. Euro Surveill 21. https://doi.org/10.2807/1560-7917.ES.2016.21.27.30280.
11. Yin W, Li H, Shen Y, Liu Z, Wang S, Shen Z, Zhang R, Walsh TR, Shen J, Wang Y. 2017. Novel plasmid-mediated colistin resistance gene mcr-3 in Escherichia coli. mBio 8:e00543-1. https://doi.org/10.1128/mBio.01166-17.
12. Carattoli A, Villa L, Feudi C, Curcio L, Orsini S, Luppi A, Pezzotti G, Magistrali CF. 2017. Novel plasmid-mediated colistin resistance mcr-4 gene in Salmonella and Escherichia coli, Italy 2013, Spain and Belgium, 2015 to 2016. Eurosurveillance 22:1–5. https://doi.org/10.2807/1560-7917.ES.2017.22.31.30589.
13. Borowiak M, Fischer J, Hammerl JA, Hendriksen RS, Szabo I, Malorny B. 2017. Identification of a novel transposon-associated phosphoethanolamine transferase gene, mcr-5, conferring colistin resistance in d-tartrate fermenting Salmonella enterica subsp. enterica serovar Paratyphi B. J Antimicrob Chemother 72:3317–3324. https://doi.org/10.1093/jac/dkx327.
14. AbuOun M, Stubberfield EJ, Duggett NA, Kirchner M, Dormer L, Nunez-Garcia J, Randall LP, Lemma F, Crook DW, Teale C, Smith RP, Anjum MF. 2017. mcr-1 and mcr-2 variant genes identified in Moraxella species isolated from pigs in Great Britain from 2014 to 2015. J Antimicrob Chemother 72:2745–2749. https://doi.org/10.1093/jac/dkx286.
15. Yang YQ, Li YX, Lei CW, Zhang AY, Wang HN. 2018. Novel plasmid-mediated colistin resistance gene mcr-7.1 in Klebsiella pneumoniae. J Antimicrob Chemother 73:1791–1795. https://doi.org/10.1093/jac/dky111.
16. Wang X, Wang Y, Zhou Y, Li J, Yin W, Wang S, Zhang S, Shen J, Shen Z, Wang Y. 2018. Emergence of a novel mobile colistin resistance gene, mcr-8, in NDM-producing Klebsiella pneumoniae. Emerg Microbes Infect 7:1–9. https://doi.org/10.1038/s41426-018-0124-z.
17. Carroll LM, Gaballa A, Guldimann C, Sullivan G, Henderson LO, Wiedmann M. 2019. Identification of novel mobilized colistin resistance gene mcr-9 in a multidrug-resistant, colistin-susceptible Salmonella enterica serotype Typhimurium isolate. mBio 10:e00853-19. https://doi.org/10.1128/mBio.00853-19.
18. Shen Z, Wang Y, Shen Y, Shen J, Wu C. 2016. Early emergence of mcr-1 in Escherichia coli from food-producing animals. Lancet Infect Dis 16:293. https://doi.org/10.1016/S1473-3099(16)00061-X.
19. Skov RL, Monnet DL. 2016. Plasmid-mediated colistin resistance (mcr-1 gene): three months later, the story unfolds. Euro Surveill 21:30155. https://doi.org/10.2807/1560-7917.ES.2016.21.9.30155.
20. Hendriksen RS, Munk P, Njage P, van Bunnik B, McNally L, Lukjancenko O, Röder T, Nieuwenhuijse D, Pedersen SK, Kjeldgaard J, Kaas RS, Clausen PTLC, Vogt JK, Leekitcharoenphon P, van de Schans MGM, Zuidema T, de



Downloaded from https://journals.asm.org/journal/msystems on 29 July 2022 by 192.38.90.185.

Roda Husman AM, Rasmussen S, Petersen B, Amid C, Cochrane G, Sicheritz-Ponten T, Schmitt H, Alvarez JRM, Aidara-Kane A, Pamp SJ, Lund O, Hald T, Woolhouse M, Koopmans MP, Vigre H, Petersen TN, Aarestrup FM, Global Sewage Surveillance Project Consortium. 2019. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nat Commun 10:1124. https://doi.org/10.1038/s41467-019-08853-3.
21. Brinch C, Leekitcharoenphon P, Reyes A, Duarte ASR, Svendsen CA, Jensen JD. 2020. Long-term temporal stability of the resistome in sewage from Copenhagen. mSystems 5:e00841-20. https://doi.org/10.1128/mSystems.00841-20.
22. Forsberg KJ, Reyes A, Wang B, Selleck EM, Sommer MOA, Dantas G. 2012. The shared antibiotic resistome of soil bacteria and human pathogens. Science 337:1107–1111. https://doi.org/10.1126/science.1220761.
23. Dalmolin T, Lima-Morales D, Barth A. 2018. Plasmid-mediated colistin resistance: what do we know? J Infectiol 1:16–22. https://doi.org/10.29245/2689-9981/2018/2.1109.
24. Jiang Y, Zhang Y, Lu J, Wang Q, Cui Y, Wang Y, Quan J, Zhao D, Du X, Liu H, Li X, Wu X, Hua X, Feng Y, Yu Y. 2020. Clinical relevance and plasmid dynamics of mcr-1-positive Escherichia coli in China: a multicentre case-control and molecular epidemiological study. Lancet Microbe 1:e24–e33. https://doi.org/10.1016/S2666-5247(20)30001-X.
25. Anyanwu MU, Jaja IF, Nwobi OC. 2020. Occurrence and characteristics of mobile colistin resistance (Mcr) gene-containing isolates from the environment: a review. Int J Environ Res Public Health 17:1028. https://doi.org/10.3390/ijerph17031028.
26. Khanawapee A, Kerdsin A, Chopjitt P, Boueroy P, Hatrongjit R, Akeda Y, Tomono K, Nuanualsuwan S, Hamada S. 2020. Distribution and molecular characterization of Escherichia coli harboring mcr genes isolated from slaughtered pigs in Thailand. Microb Drug Resist 27:971–979. https://doi.org/10.1089/mdr.2020.0242.
27. Tyson GH, Li C, Hsu C-H, Ayers S, Borenstein S, Mukherjee S, Tran T-T, McDermott PF, Zhao S. 2020. The mcr-9 gene of salmonella and Escherichia coli is not associated with colistin resistance in the United States. Antimicrob Agents Chemother 64:e00573-20. https://doi.org/10.1128/AAC.00573-20.
28. Aarestrup FM, Woolhouse MEJ. 2020. Using sewage for surveillance of antimicrobial resistance. Science 367:630–632. https://doi.org/10.1126/science.aba3432.
29. Wang C, Feng Y, Liu L, Wei L, Kang M, Zong Z. 2020. Identification of novel mobile colistin resistance gene mcr-10. Emerg Microbes Infect 9:508–516. https://doi.org/10.1080/22221751.2020.1732231.
30. Leinonen R, Akhtar R, Birney E, Birney L, Cerdeno-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, LundGibson R, Hoad G, Jang M, Pakseresht N, Plaister S, Radhakrishnan R, Reddy K, Sobhany S, Ten Hoopen P, Vaughan R, Zalunin V, Cochrane G. 2011. The European nucleotide archive. Nucleic Acids Res 39:D28–D31. https://doi.org/10.1093/nar/gkq967.
31. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 67:2640–2644. https://doi.org/10.1093/jac/dks261.
32. Quast C, Pruesse E, Yilmaz P, Yilmaz J, Schweer T, Yarza P, Peplies J, Glöckner F. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:590–596. https://doi.org/10.1093/nar/gks1219.
33. Bushnell B. 2014. BBMap: a fast, accurate, splice-aware aligner. https://www.osti.gov/servlets/purl/1241166.
34. Clausen PTLC, Aarestrup FM, Lund O. 2018. Rapid and precise alignment of raw reads against redundant databases with KMA. BMC Bioinformatics 19:1–8. https://doi.org/10.1186/s12859-018-2336-6.
35. Aitchison J. 1982. The statistical analysis of compositional data. J R Stat Soc Series B Stat Methodol 44:139–160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x.
36. Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB. 2013. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. PLoS One 8:e67019. https://doi.org/10.1371/journal.pone.0067019.
37. Hunter JD. 2007. Matplotlib: a 2D graphics environment. Comput Sci Eng 9:90–95. https://doi.org/10.1109/MCSE.2007.55.
38. Waskom M, Botvinnik O, Gelbart M, Ostblom J, Hobson B, Lukauskas S, Gemperline DC, Augspurger T, Halchenko Y, Warmenhoven J, Cole JB, de Ruiter J, Vanderplas J, Hoyer S, Pye C, Miles A, Swain C, Meyer K, Martin M, Bachant P, Quintero E, Kunter G, Villalba S, Fitzgerald C, Evans CG, Williams ML, O'Kane D, Yarkoni T, Brunner T. 2020. mwaskom/seaborn: v0.11.0. https://doi.org/10.5281/zenodo.592845.
39. Gillies S, and others. 2007. A. Shapely: manipulation and analysis of geometric objects. https://github.com/shapely/shapely.
40. Met Office, 2021. Cartopy: a cartographic python library with a Matplotlib interface. http://scitools.org.uk/cartopy.
41. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. 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 2:15–13. https://doi.org/10.1186/2049-2618-2-15.
42. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc 57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
43. Gloor GB, Macklaim JM, Fernandes AD. 2016. Displaying variation in large datasets: plotting a visual summary of effect sizes. J Comput Graph Stat 25:971–979. https://doi.org/10.1080/10618600.2015.1131161.
44. Calle ML. 2019. Statistical analysis of metagenomics data. Genomics Inform 17:e6. https://doi.org/10.5808/GI.2019.17.1.e6.
45. Pathogen Detection. https://www.ncbi.nlm.nih.gov/pathogens/. Accessed 5 August 2021.
46. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges T, Haft DH, Hoffmann M, Pettengill JB, Prasad AB, Tillman GE, Tyson GH, Klimke W. 2021. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep 11:12728. https://doi.org/10.1038/s41598-021-91456-0.
47. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. MetaSPAdes: a new versatile metagenomic assembler. Genome Res 27:824–834. https://doi.org/10.1101/gr.213959.116.
48. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421.
49. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033.
50. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller Aarestrup F, Hasman H. 2014. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother 58:3895–3903. https://doi.org/10.1128/AAC.02412-14.
51. Johansson MHK, Bortolaia V, Tansirichaiya S, Aarestrup FM, Roberts AP, Petersen TN. 2021. Detection of mobile genetic elements associated with antibiotic resistance in *Salmonella enterica* using a newly developed web tool: MobileElementFinder. J Antimicrob Chemother 76:101–109. https://doi.org/10.1093/jac/dkaa390.
52. Ward JH. 1963. Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244. https://doi.org/10.1080/01621459.1963.10500845.
53. Yu G, Smith DK, Zhu H, Guan Y, Lam TT. 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol 8:28–36. https://doi.org/10.1111/2041-210X.12628.
54. Wickham H. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag, New York, NY.
55. Matlock W, Lipworth S, Constantinides B, Peto TEA, Walker AS, Crook D, Hopkins S, Shaw LP, Stoesser N. 2021. Flanker: a tool for comparative genomics of gene flanking regions. Microb Genom 7:000634. https://doi.org/10.1099/mgen.0.000634.
56. Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987–2993. https://doi.org/10.1093/bioinformatics/btr509.
57. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010.
58. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2: approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490.

Downloaded from https://journals.asm.org/journal/msystems on 29 July 2022 by 192.38.90.185.

## Supplementary Material

Martiny, H. M., Munk, P., Brinch, C., Szarvas, J., Aarestrup, F. M., & Petersen, T. N. (2022). Global Distribution of mcr Gene Variants in 214K Metagenomic Samples. mSystems, 7(2), e00105-22

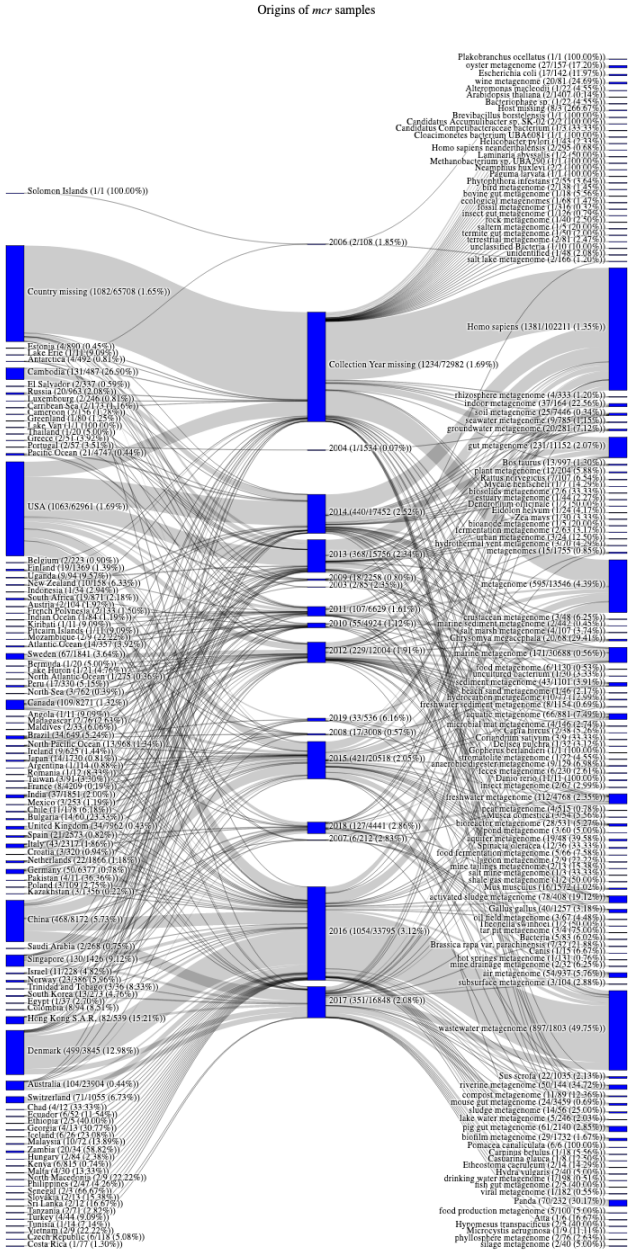



*Figure S1: Sample origins of metagenomes containing mcr* genes *and the number of metagenomes without in parentheses.*

## Supplementary Material

Martiny, H. M., Munk, P., Brinch, C., Szarvas, J., Aarestrup, F. M., & Petersen, T. N. (2022). Global Distribution of mcr Gene Variants in 214K Metagenomic Samples. mSystems, 7(2), e00105-22

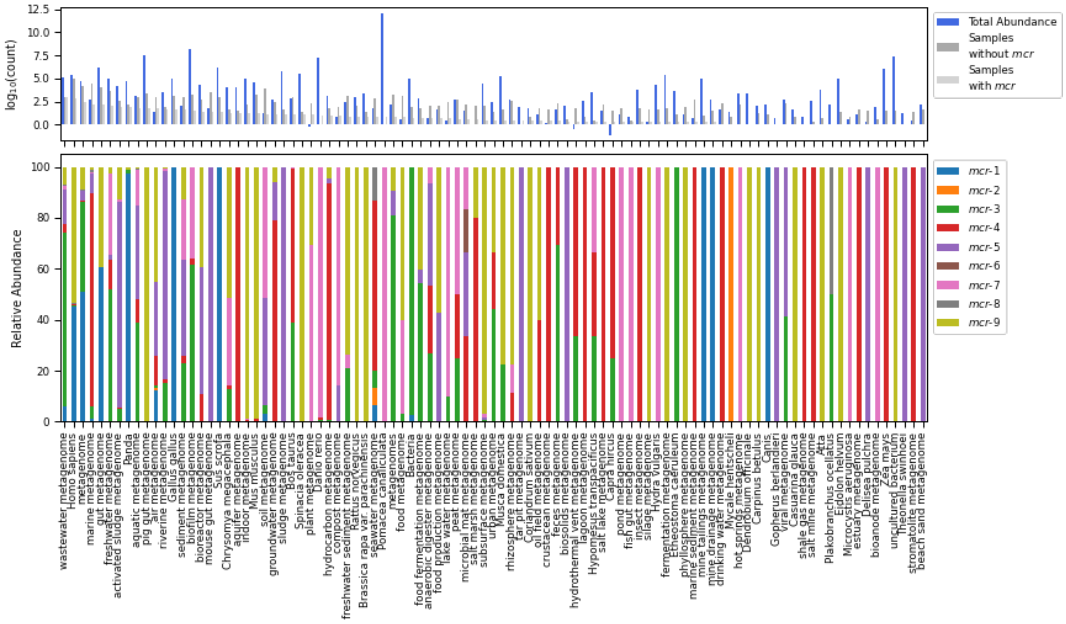



*Figure S2:* ***Distribution of mcr genes in all sampling sources****.Top: The log count of metagenomes with and without mcr genes, as well as the total mcr abundance (blue) per host or environment. Bottom: The relative abundance of mcr read fragments aligned.*

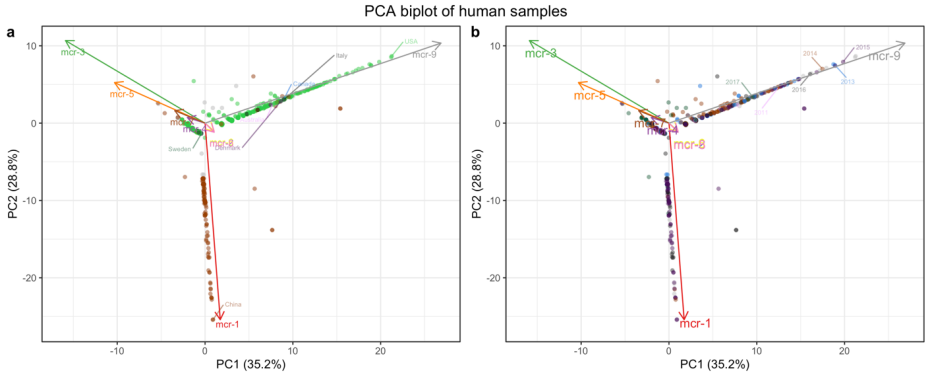



*Figure S3:* ***Biplot of human samples from PCA*** *on the full dataset colored by* ***a****. country and* ***b****. sampling year.*

## Supplementary Material

Martiny, H. M., Munk, P., Brinch, C., Szarvas, J., Aarestrup, F. M., & Petersen, T. N. (2022). Global Distribution of mcr Gene Variants in 214K Metagenomic Samples. mSystems, 7(2), e00105-22

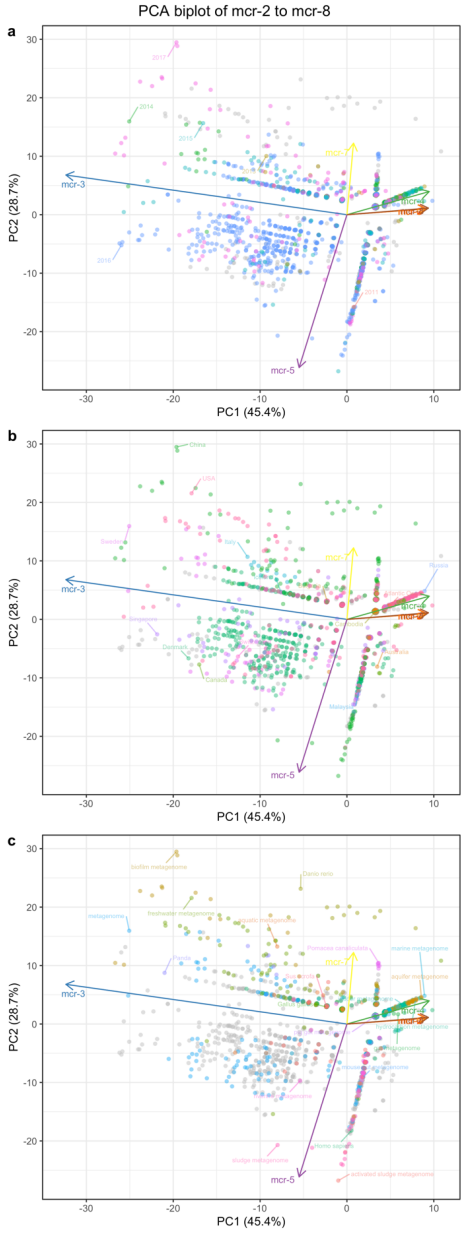



*Figure S4:* ***Biplots produced by doing compositional PCA on read counts of mcr-2 to mcr-8****, where samples are colored by* ***a.*** *sampling year,* ***b.*** *sampling location, and* ***c.*** *sampling host.*

## Supplementary Material

Martiny, H. M., Munk, P., Brinch, C., Szarvas, J., Aarestrup, F. M., & Petersen, T. N. (2022). Global Distribution of mcr Gene Variants in 214K Metagenomic Samples. mSystems, 7(2), e00105-22

| | *mcr-1* | *mcr-2* | *mcr-3* | *mcr-4* | *mcr-5* | *mcr-6* | *mcr-7* | *mcr-8* | *mcr-9* |
|---|---|---|---|---|---|---|---|---|---|
| Isolate frequency (%) | 51.08 | 0.25 | 6.83 | 0.74 | 1.26 | 0.00 | 0.00 | 1.18 | 40.38 |

*Table S1:* ***Prevalence of mcr gene variants in single isolates screened by the NCBI Pathogen Detection pipeline****. The percentage is the number of isolates with one of the mcr genes out of the total number of isolates screened.*

## Supplementary Material

Martiny, H. M., Munk, P., Brinch, C., Szarvas, J., Aarestrup, F. M., & Petersen, T. N. (2022). Global Distribution of mcr Gene Variants in 214K Metagenomic Samples. mSystems, 7(2), e00105-22

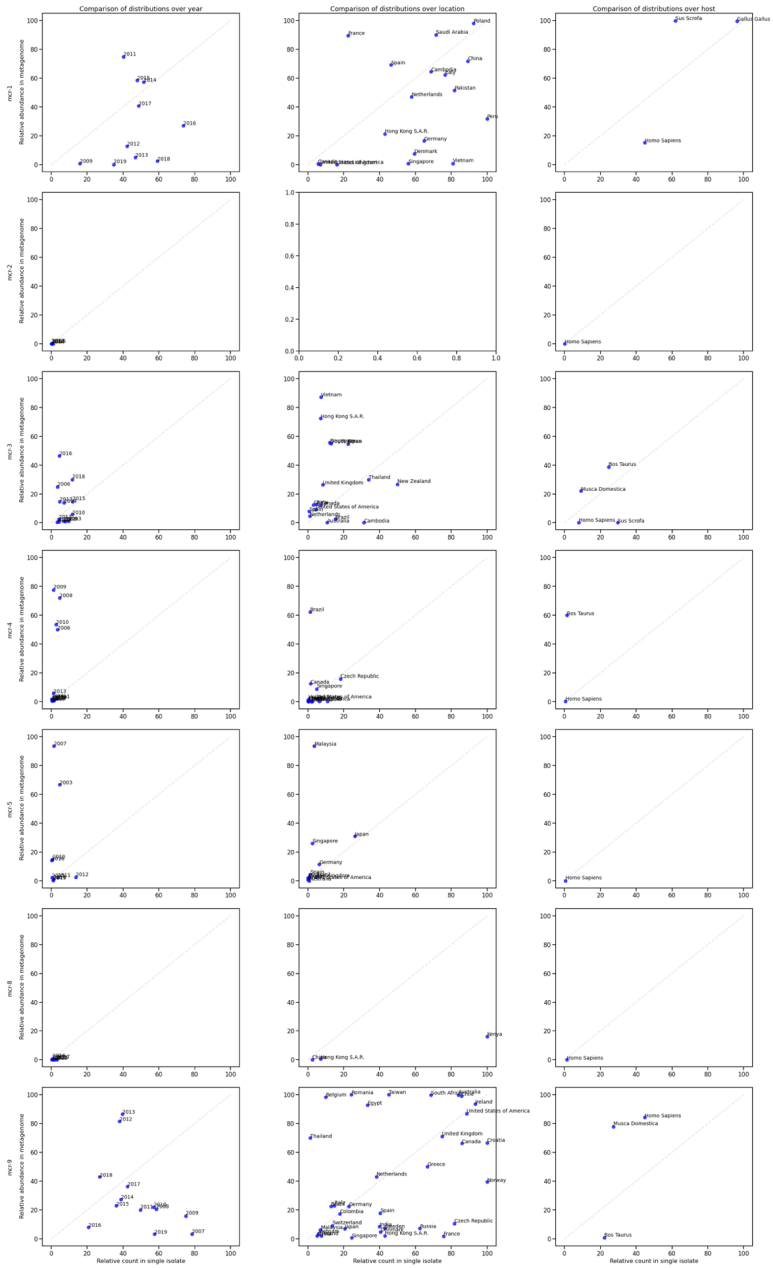



*Figure S5:* ***Pairwise comparison of mcr gene prevalence in single isolates and metagenomes for common sample origins****, where the prevalence is higher than zero for both measurements. The relative count is the percentage of mcr positive isolates carrying a variant, and the relative abundance is the frequency of metagenomic mcr reads aligned to a variant. The closer to the grey line a marker is, the more agreement there is between the two measurements.*

## Supplementary Material

Martiny, H. M., Munk, P., Brinch, C., Szarvas, J., Aarestrup, F. M., & Petersen, T. N. (2022). Global Distribution of mcr Gene Variants in 214K Metagenomic Samples. mSystems, 7(2), e00105-22

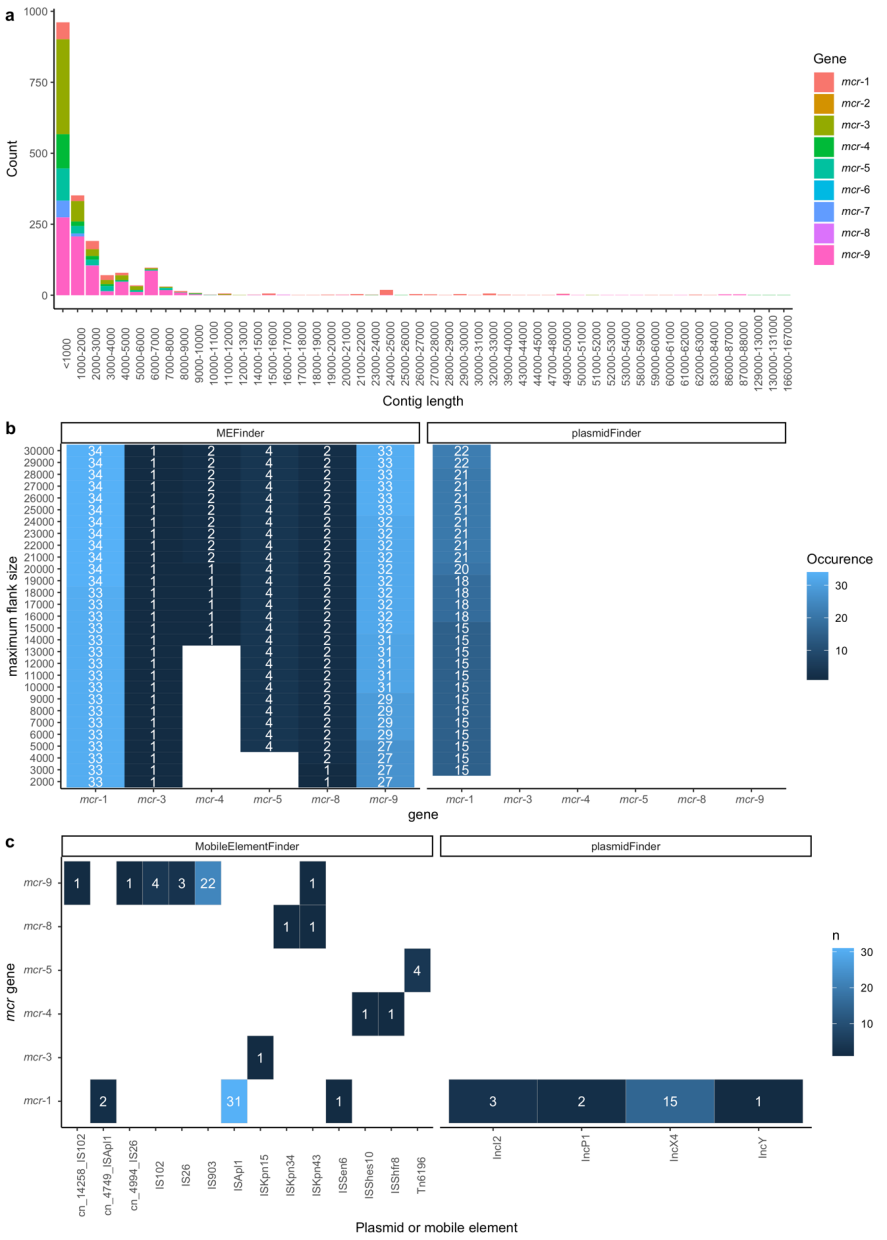



*Figure S6:* ***Distribution of mcr contig lengths and the occurrence of elements in flanks****.* ***a****. Histogram of mcr contig lengths, colored by which mcr gene present in the contig.* ***b****. Count of mobile elements or plasmids occurring in flanks of different sizes.* ***c.*** *Count of the occurrence of mobile genetic elements (left) and plasmids (right) in mcr contigs with flank sizes between 1000 and 21,000 bp. The number does not consider if two elements are on the same contig.*

## Supplementary Material

Martiny, H. M., Munk, P., Brinch, C., Szarvas, J., Aarestrup, F. M., & Petersen, T. N. (2022). Global Distribution of mcr Gene Variants in 214K Metagenomic Samples. mSystems, 7(2), e00105-22

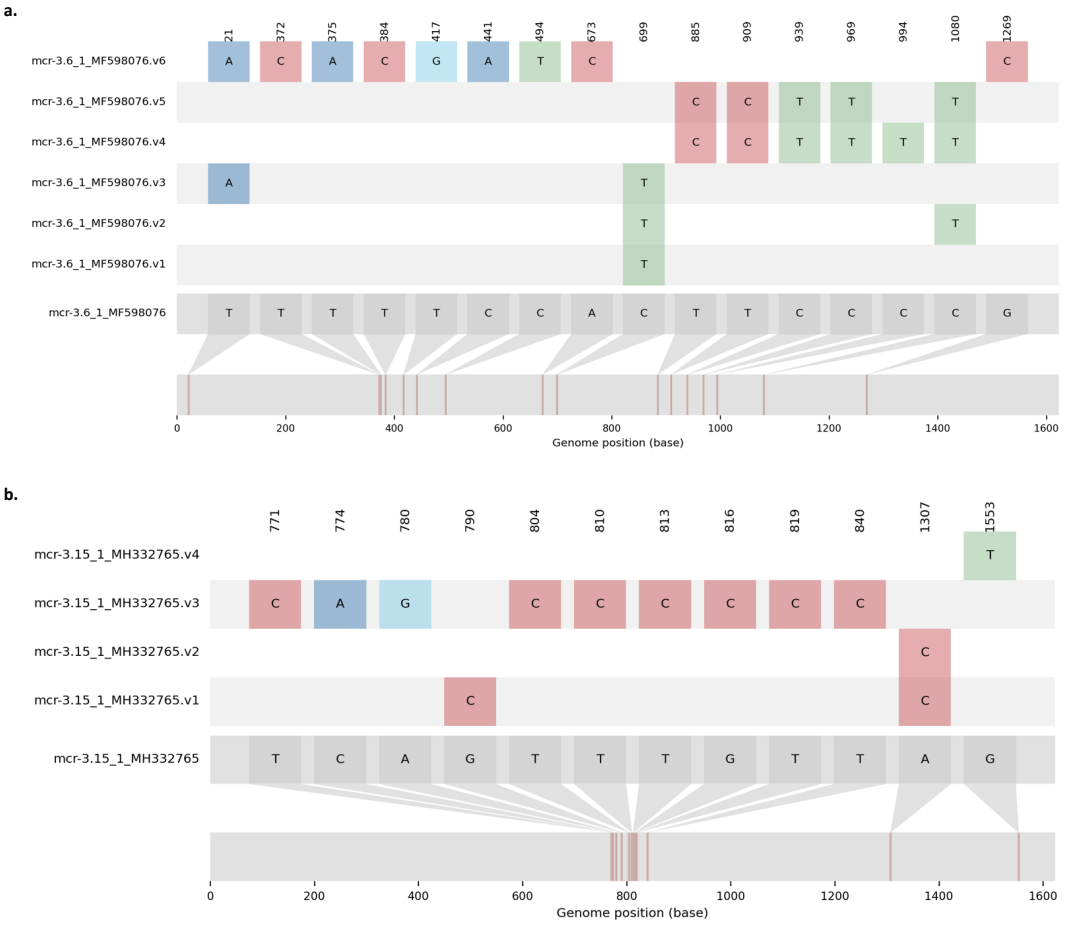
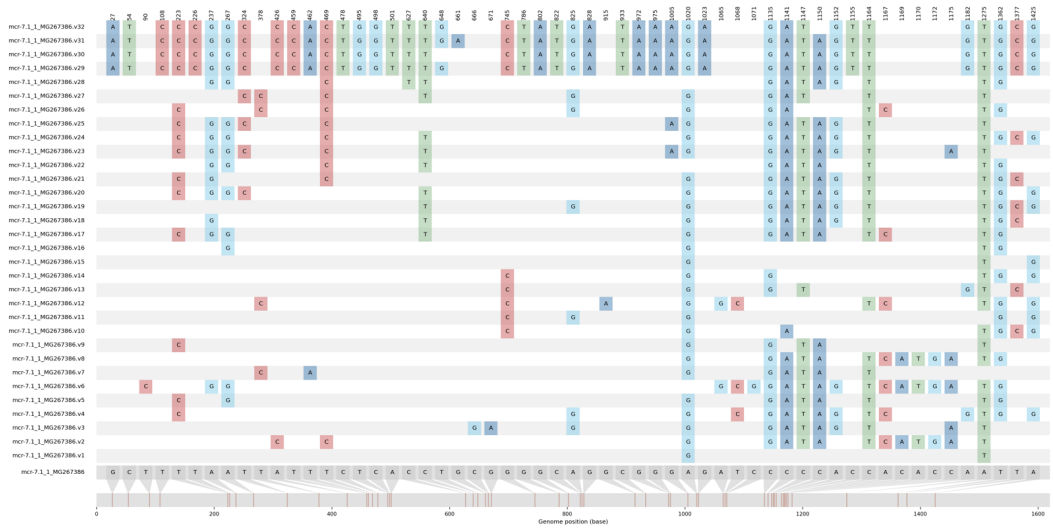



Figure S7: **Overview of SNPs found in a. mcr-3.6 consensus sequences and b. *mcr-3.15* consensus sequences** . The distinct sequences were aligned with MAFFT and visualized with snipit.

## Supplementary Material

Martiny, H. M., Munk, P., Brinch, C., Szarvas, J., Aarestrup, F. M., & Petersen, T. N. (2022). Global Distribution of mcr Gene Variants in 214K Metagenomic Samples. mSystems, 7(2), e00105-22

**a.**




**b.**

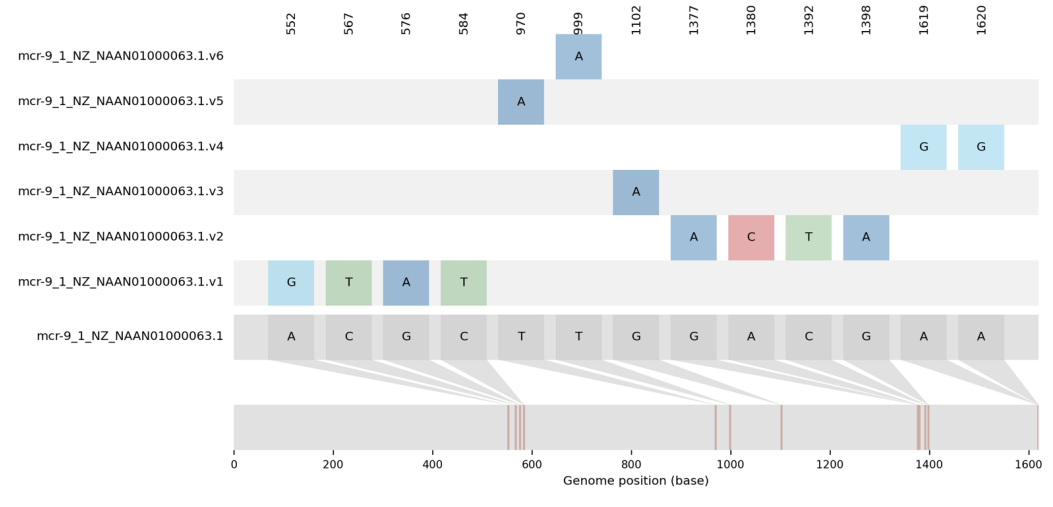



*Figure S8:* ***Overview of SNPs found in a. mcr-7.1 consensus sequences and b. mcr-9.1 sequences****. The distinct sequences were aligned with MAFFT and visualized with snipit.*

## Supplementary Material

Martiny, H. M., Munk, P., Brinch, C., Szarvas, J., Aarestrup, F. M., & Petersen, T. N. (2022). Global Distribution of mcr Gene Variants in 214K Metagenomic Samples. mSystems, 7(2), e00105-22

| | ***mcr*-1.1.v2** | ***mcr*-3.6.v1** | ***mcr*-3.6.v3** | ***mcr*-3.15.v2** | ***mcr*-4.3.v3** | ***mcr*-4.3.v4** | ***mcr*-4.3.v5** | ***mcr*-5.1.v1** | ***mcr*-8_1.v1** | ***mcr*-9_1.v4** |
|---|---|---|---|---|---|---|---|---|---|---|
| Total metagenomic hits | 1 | 14 | 1 | 9 | 8 | 1 | 1 | 1 | 1 | 139 |
| Total BLAST hits | 17 | 3 | 4 | 3 | 25 | 1 | 1 | 5 | 20 | 100 |
| **Species** | | | | | | | | | | |
| Acinetobacter baumannii | | | | | 5 | | | | | |
| Acinetobacter nosocomialis | | | | | 1 | | | | | |
| Aeromonas hydrophila | | | 1 | 2 | | | | | | |
| Aeromonas media | | | 3 | | | | | | | |
| Aeromonas sanarellii | | 1 | | | | | | | | |
| Aeromonas sp. 2692-1 | | | | 1 | | | | | | |
| Aeromonas veronii | | 2 | | | | | | | | |
| Citrobacter freundii | | | | | | | | | | 4 |
| Citrobacter sp. | | | | | | | | | | 1 |
| Citrobacter telavivensis | | | | | | | | | | 1 |
| Enterobacter asburiae | | | | | | | | | | 3 |
| Enterobacter cancerogenus | | | | | | | | | | 1 |
| Enterobacter cloacae | | | | | 3 | | | | | 3 |
| Enterobacter cloacae complex sp. | | | | | | | | | | 2 |
| Enterobacter hormaechei | | | | | | | | | | 31 |
| Enterobacter hormaechei subsp. xiangfangensis | | | | | | | | | | 1 |
| Enterobacter kobei | | | | | 1 | | | | | 2 |
| Enterobacter roggenkampii | | | | | | | | | | 1 |
| Enterobacter sp. BWH 37 | | | | | | | | | | 1 |
| Enterobacter sp. MGH 14 | | | | | | | | | | 1 |

**Supplementary Material**

Martiny, H. M., Munk, P., Brinch, C., Szarvas, J., Aarestrup, F. M., & Petersen, T. N. (2022). Global Distribution of mcr Gene Variants in 214K Metagenomic Samples. mSystems, 7(2), e00105-22

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Enterobacter sp. T2 | | | | | | | | | | 1 |
| Escherichia coli | 12 | | | | | | | 1 | | 9 |
| Expression vector pUC57-Kan-mcr-9 | | | | | | | | | | 1 |
| Hafnia paralvei | | | | | | | | | | 1 |
| Klebsiella grimontii | | | | | | | | | | 1 |
| Klebsiella oxytoca | | | | | | | | | | 1 |
| Klebsiella pneumoniae | 5 | | | | | | | | 18 | 2 |
| Klebsiella quasipneumoniae | | | | | | | | | 1 | |
| Leclercia adecarboxylata | | | | | 1 | | | | | 2 |
| Leclercia sp. | | | | | | | | | | 2 |
| Lelliottia amnigena | | | | | 1 | | | | | |
| mixed culture bacterium | | | | | 1 | | | | | |
| Salmonella enterica | | | | | | | | | | 2 |
| Salmonella enterica subsp. diarizonae serovar b,50:-:- | | | | | | | | | | 1 |
| Salmonella enterica subsp. enterica serovar 4,[5],12:i:- | | | | | | | | | | 3 |
| Salmonella enterica subsp. enterica serovar 4,5,12:i- | | | | | | | | | | 3 |
| Salmonella enterica subsp. enterica serovar Albany | | | | | | | | | | 2 |
| Salmonella enterica subsp. enterica serovar Brandenburg | | | | | | | | | | 1 |
| Salmonella enterica subsp. enterica serovar Dessau | | | | | | | | | | 2 |
| Salmonella enterica subsp. enterica serovar Heidelberg | | | | | | | | | | 4 |

**Supplementary Material**

Martiny, H. M., Munk, P., Brinch, C., Szarvas, J., Aarestrup, F. M., & Petersen, T. N. (2022). Global Distribution of mcr Gene Variants in 214K Metagenomic Samples. mSystems, 7(2), e00105-22

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Salmonella enterica subsp. enterica serovar Johannesburg | | | | | | | | | | 1 |
| Salmonella enterica subsp. enterica serovar Newport | | | | | | | | | | 1 |
| Salmonella enterica subsp. enterica serovar Saintpaul | | | | | | | | | | 4 |
| Salmonella enterica subsp. enterica serovar Schwarzengrund | | | | | | | | | | 2 |
| Salmonella enterica subsp. enterica serovar Westhampton | | | | | | | | | | 1 |
| Serratia marcescens | | | | | | | | | | 1 |
| Shewanella frigidimarina NCIMB 400 | | | | | 1 | | | | | |
| uncultured bacterium | | | | | 9 | 1 | 1 | 4 | 1 | |
| Acinetobacter baumannii | | | | | 5 | | | | | |
| Acinetobacter nosocomialis | | | | | 1 | | | | | |

*Table S 2:* ***New potential subvariants of mcr genes can be found in a multitude of different species.*** *Matches were found with BLAST and that had 100% identity.*

CHAPTER 6

# Manuscript III

## Utilizing co-abundances of antimicrobial resistance genes to identify potential co-selection in the resistome

Hannah-Marie Martiny[1], Patrick Munk[1], Christian Brinch[1], Frank M. Aarestrup[1], M. Luz Calle[2], Thomas Nordahl Petersen[1]

1. National Food Institute, Technical University of Denmark, Kongens Lyngby, Denmark.
2. Biosciences Department, University of Vic - Central University of Catalonia, Vic, Spain.

In preparation.

# Utilizing co-abundances of antimicrobial resistance genes to identify potential co-selection in the resistome

**Authors**:

Hannah-Marie Martiny[1]

Patrick Munk[1]

Christian Brinch[1]

Frank M. Aarestrup[1]

M. Luz Calle[2]

Thomas N. Petersen[1]

**Affiliations:**

[1]Research Group for Genomic Epidemiology, Technical University of Denmark, Kongens Lyngby, Denmark

[2]Biosciences Department, Faculty of Sciences and Technology, University of Vic - Central University of Catalonia, Vic, Spain

## Abstract

The rapid spread of antimicrobial resistance (AMR) is a threat to global health, and the nature of co-occurring antimicrobial resistance genes (ARGs) may cause collateral AMR effects once antimicrobial agents are used. Therefore, it is essential to identify which pairs of ARGs co-occur. Given the wealth of NGS data available in public repositories, we have investigated the correlation between ARG abundances in a collection of 214,095 metagenomic datasets. Using more than $6.76 \cdot 10^8$ read fragments aligned to ARGs to infer pairwise correlation coefficients, we found that more ARGs correlated with each other in human and animal sampling origins than in soil and water environments. Furthermore, we showed that the correlations serve as risk profiles of resistance co-occurring to critically important antimicrobials. Using these profiles, we found several key ARGs indirectly but

strongly selecting for ARGs of critical importance, such as tetracycline ARGs correlating with most forms of resistances. In conclusion, this study highlights the important ARG players indirectly involved in shaping the resistomes of various environments that can serve as monitoring targets in AMR surveillance programs.

## Introduction

Antimicrobial resistance (AMR) is one of the biggest threats to human and animal health[1,2], and it is widely acknowledged that the misuse of antimicrobials has accelerated the dissemination and prevalence of antimicrobial resistance genes (ARGs)[3]. Most attempts to reduce the burden of AMR have focused on reducing the use of single classes of antimicrobial agents considered of critical importance[4]. Despite considerable efforts in various settings, such as livestock, these regulations have not significantly reduced the spread of ARGs[5,6]. It is known that even after banning the use of specific antimicrobials, the ARGs conferring resistance will still be prevalent in the environment[7]. Furthermore, studies have shown that ARGs can be indirectly selected if they co-occur with an ARG conferring resistance to the antimicrobial, causing the selective pressure[6,8–10]. We have also recently observed that changes in the selective pressure of even a single antimicrobial agent influence several ARGs in pig metagenomes[10]. It is, however, not known whether this is due to co-selection due to the genetic linkage of ARGs, the presence of ARGs in the same bacterial clones or because different bacterial species with different ARGs are co-selected in a microbial network.

Most studies associating AMU and AMR have, however, focused on one antimicrobial agent at that time and how the use impacts the development of resistance to that agent and not to other unrelated antimicrobials. This needs to be understood better to stop these collateral damage effects[11,12]. Co-occurrence of microbes has been evaluated in soil[13,14] and marine environments[15,16], whereas ARG co-occurrences have been studied in sewage sludge[17], freshwater[18], marine[19], swine[7], and cattle[20]. However, most of these studies have only been evaluated on a smaller scale and not always using the same methods. There are currently a large collection of next-generation sequencing datasets from metagenomic samples available in public repositories, providing an excellent resource to quantify the prevalence of ARGs by analyzing the read abundances [21–24].

In this study, we have studied the co-abundance of sequencing reads aligned to ARGs to assess how resistance to one specific antimicrobial agent is linked to the abundance of another class of antimicrobial. With a collection of 214,095 metagenomic datasets[25], we examined the correlation between pairwise ARG read abundances with a compositional approach[26,27] using SparCC (Sparse Correlations for Compositional data)[28]. Our results demonstrated that many ARG pairs interact but are highly specific to the environment. We believe that these interactions provide a new foundation to understand how ARGs are being co-selected independently of the microbial context, and the findings can be used to design targeted interventions to limit the spread of AMR.

# Methods

## Data collection and pre-processing

We have previously described in detail the process of downloading and analyzing 214,095 metagenomic samples[23,25], but in brief: We downloaded raw sequencing reads corresponding to 442 Tbp from metagenomic samples deposited in the European Nucleotide Archive[29] that were uploaded between 2010-01-01 and 2020-01-01 and had at least 100,000 reads and were shotgun sequenced. The raw sequencing reads were quality-checked with FASTQC v.0.11.15 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimmed with BBduk2 36.49[30]. The trimmed reads were then globally aligned using KMA[31] 1.2.21 against two reference sequence databases: ResFinder[32] (downloaded 25-01-2020) and Silva[33] (version 38, downloaded 16-01-2020). ResFinder is a database of 3,085 acquired ARGs, whereas Silva is a 16S/18S rRNA database of 2,225,272 sequences.

## Homology-reduced ResFinder

Since there is a possibility that KMA might assign reads to closely related homologs differently across samples, we decided to homology reduce the 3,085 ARGs of ResFinder using USEARCH[34] 11.0.667 with 90% nucleotide identity, producing a total of 716 ARG groups. For each of these groups, the read counts for the genes in that group were first adjusted by the length of the gene and then aggregated together. These new group counts are used in the

correlation analysis described in the next section. For the rest of the manuscript, we refer to these 716 ARG clusters as simply ARGs.

## Calculating relative abundances

For a category label, we calculated the relative abundance of fragment counts assigned to different genes or classes as:

$$\text{Relative abundance}(x) = \frac{\kappa}{\sum n_i} n_i$$

where $x$ is the label, $n_i$ is the count of read fragments assigned to gene $i$, and $\kappa = 100$ is a scaling constant.

## Inferring pairwise correlations with SparCC

The SparCC algorithm[28] was used to obtain correlations using pairs of log-ratio transformed ARG read counts to infer linear Pearson correlations. SparCC obtains linear Pearson correlations and p-values through an iterative approach that adjusts for spurious correlations and lowers the false discovery rate. The ARG-ARG correlations were inferred as the average over 50 iterations, and one-sided pseudo p-values were obtained through a bootstrapping procedure of 100 rounds. In each bootstrapping round, the input count matrix was shuffled and correlations were averaged over 10 iterations to infer one-sided p-values to test whether the correlation for the observed data was statistically significant. Correlations values $\geq 0.6$ with p-values $\leq 0.01$ were selected for further analysis. We implemented SparCC to run on GPUs on the Danish National Supercomputer for Life Sciences (https://www.computerome.dk).

We ran SparCC on the entire dataset of the 214K metagenomic samples and subsets of samples grouped by sampling host and environment, where at least 800 samples existed. Due to inconsistent labeling of the sampling sources, we made new source groups, as shown in Table 1. We only consider genes for SparCC analysis that are present in at least 10 samples with a minimum read fragment count of 50. In total, 11 different correlation matrices were made, i.e., one for each of the source groups listed in Table 1. The correlation networks were visualized in R 4.1.0[35] with packages igraph[36], qgraph[37], and ggraph[38] using the Fruchterman-Reingold layout algorithm[39].

| Source group | Sampling source label(s) | Number of samples | Number of samples with ARGs |
|---|---|---|---|
| Air | Air metagenome | 914 | 870 |
| Chicken | *Gallus gallus* (1,219), chicken gut metagenome (4) | 1,223 | 1,215 |
| Cow | *Bos taurus* (872), cow dung metagenome (14) | 886 | 824 |
| Dog | *Canis lupus familiaris* | 3,439 | 3,182 |
| Freshwater | Freshwater metagenome | 4,494 | 585 |
| Human | *Homo sapiens* | 95,003 | 57,239 |
| Marine | Marine metagenome | 30,002 | 5,444 |
| Mouse | *Mus musculus* (1,462), mouse metagenome (50), mouse gut metagenome (2,435) | 3,947 | 3,909 |
| Pig | *Sus scrofa* (673), *Sus scrofa domesticus* (355), pig metagenome (2,129), pig gut metagenome (72) | 3,229 | 3,461 |
| Soil | Soil metagenome | 6,533 | 2,822 |
| All | | 214,095 | 119,206 |

*Table 1: Grouped labels for hosts and environments. The parenthesis after each sampling source label denotes the number of samples assigned to that label if the group consisted of multiple labels.*

## Network comparisons

The topology of the different networks is described using different metrics: the number of nodes ($N$) and edges ($E$), the global clustering coefficient, network density, edge density, and the number of components. The global clustering coefficient, or the graph transitivity, measures the density of node triplets in the network[40]. The network density is calculated as $2E/N(N-1)$ as given in Parente et al. (2018)[41]. Edge density is the number of edges over the number of possible edges[42]. The average correlation between ARGs of two antimicrobial classes was calculated using Fisher's z-transformation on the correlation values, averaging the z-scores and converting it back to a correlation score with the inverse Fisher transformation[43].

## Data and code availability

The matrix of ARG read counts is available at https://zenodo.org/record/6919377, and the code used to run the analysis and create figures are available at https://github.com/hmmartiny/global_resistome_correlations. Classifications of antimicrobial importance were retrieved from the 6$^{th}$ revision of critically important antimicrobials for human medicine from https://www.who.int/publications/i/item/9789241515528, accessed 2022-10-10.

## Results

This study investigated the correlation of pairwise ARG abundances across a highly diverse set of 214,095 metagenomic samples. The samples included in the 214K metagenomic collection come from a variety of backgrounds, such as the sampling period spanning between 2000 and 2020 (Figure S1a), primarily sequenced on Illumina platforms (Figure S1b), and hailing from all over the world (Figure S1c). We observed that not all samples contained sequence read fragments aligned to ARGs, so we only included those that did in the correlation analyses (n = 119,206; Table 1). Looking at the various sampling sources, read fragments aligned to ARGs conferring resistance to tetracycline and beta-lactams were very common in host-associated sources, whereas phenicol resistance was more frequently observed in environmental sources (freshwater, marine, and soil) (Table 2). *catA1* was the most dominant gene in the environmental samples, especially in marine samples, whereas various *tet* genes had high relative abundances in livestock or human samples: *tet(W)* in chicken and pig samples; *tet(Q)* in cow, human, and pig metagenomes. *blaTEM-52B* accounted for more than 30% of the read fragments assigned to ARGs in air metagenomes (Figure S2).

### Balancing the sparsity and network complexity

We selected ARGs to infer abundance-based correlations by considering how sparse the input count matrix was and how that sparsity affected the results of SparCC. Out of the 716 ARGs that could be included, we observed that without filtering the raw counts, SparCC found that most of the ARGs correlated with each other, even if the count of read fragments was low. Based on this observation, we decided to require that for an ARG to be included, it had to have a minimum count across a specific number of samples for a group. Applying these two filters, the number of correlations inferred for each sampling group decreased as compared to the no-filter results (Table S1). We decided to pick filter settings to balance the amount of sparsity allowed in the sample groups and the number of correlation coefficients, which ended up being a minimum fragment count of 50 across at least 10 samples.

## Analysis of large-scale metagenomic correlation networks

Based on our filter settings, we constructed a global network using the correlation coefficients for the entire collection of metagenomes, with each node representing an ARG and each edge representing a pairwise ARG connection (correlation $\geq 0.6$, p-value $< 0.01$, Figure S3). The global network, nicknamed 'all', contained 225 ARGs connected through 2,344 correlation edges (Figure 1a). As this all-network was hard to interpret due to the many highly interconnected ARGs, we also inferred pairwise ARG correlations in specific sampling groups (Table 1). The genes that were part of these networks were found to correlate with varying degrees of strength (Figure 1a-b). For example, the human network contained many correlation coefficients, but most were less than 0.8 (Figure 1c). Another example is the marine network, where only a few ARGs were found to correlate, but with values above 0.9 (Figure 1b-c). Despite the networks reflecting the composition of the various environments, we still observed overlaps between which ARGs were found to correlate. One example is that all the correlations inferred from the pig metagenomes also existed in the human metagenomic network (Figure 1d).

There was a limited number of correlations for ARGs encoding resistance to fluoroquinolones, steroid antibiotics (fusidic acid), colistin, and rifampicin. On the other hand, beta-lactam, tetracycline, and aminoglycoside ARGs had many correlations with each other and with other classes (Figure S4, Figure S5). Despite resistance to some antimicrobial classes being the most abundant, the ARGs did not always correlate to many others. For example, tetracycline ARGs were the most abundant in dog metagenomes, but no correlations were inferred for these ARGs. Similarly, in the human samples where aminoglycoside and beta-lactam ARGs were less abundant than tetracycline ARGs, aminoglycoside and beta-lactam resistance genes had a higher number of correlations coefficients were reported (Table 2, Figure S4).

On the level of ARG abundance, we observed that just because an ARG was highly abundant in a sampling group, it did not automatically mean that it correlated to many other ARGs. The highly abundant *catA1* gene in marine (97.4% of all ARG reads), freshwater (55.1%), and soil samples (84.7%) (Figure S2) did only correlate with one or two other genes in the water environmental networks, and none in the soil network. On the other hand, *catA1* did seem

to be correlated with 15 other genes in the pig correlation network despite not being highly abundant in that group of samples (Figure S5a). *mef(A)*_1 accounted for 15.9% of the reads aligned to ARGs in cow samples and 6.63% in pigs (Figure S2) and was also strongly correlating with other genes, which mainly conferred resistance to aminoglycosides, (fluoro)quinolones and tetracyclines (Figure S5b). *tet(L)_4* only accounted for 4.01% of the read fragments aligned to ARGs in metagenomes of chicken origins (Figure S2) but was shown to correlate in its abundance with the abundance of 8 other ARGs, e.g., with a correlation of 0.77 with *lnu(A)_1 (*Figure S5c).

## The hidden signals between ARGs profile the potential risk of co-selection in different environmental contexts

Using the correlations between ARGs in the various environments (Figure 1), we calculated the average correlation between ARGs of different antimicrobial classes (Figure S6). These average correlations can then serve as profiles to assess the risk of indirectly selecting ARGs that gives resistance to different ARGs through co- and cross-resistance. These risk profiles can then be used to judge the strength of interactions upon using one antimicrobial in each setting. Upon constructing these profiles, we observed that the strength and the number of correlations highly depend on the antimicrobial classes and the environmental context. We hypothesize that if an important (IA) or highly important antimicrobial (HIA) is used, first, the resistance to the antimicrobial class will likely flourish, and, secondly, through co- and cross-resistance, so will ARGs conferring resistance to other classes, including those that are critically important antimicrobials (CIA).

Figure 2 shows two risk profiles for ARG correlations for two highly critical important antimicrobials glycopeptide and macrolide. Correlations between ARGs of glycopeptide resistance to other resistance classes were much rarer than those connected with macrolide ARGs (Figure S6) and those correlations that were observed were relatively low (correlation<0.8, Figure 2). Different vancomycin resistance cassettes were responsible in different environments, namely *VanHAX*, *VanC2*, and *VanX_bc* in human samples and *VanHDX* and *VanC1XY* in pig samples (Figure S7a). *VanHAX* only has one correlation, whereas the remaining five correlate with many different ARGs.

On the contrary, ARGs conferring resistance to macrolide were much more interactive with other classes of resistance genes in all networks and had strong correlations with specific classes (correlation > 0.9, Figure 2, Figure S6). The macrolide ARGs co-abundant with other ARGs were many but separated into distinct network clusters (Figure S7b). For example, *mef(A)* and *msr(D)* were usually found together in different environments, both in small and large clusters.

While Figure 2 highlights how CIA ARGs interact, it is just as important to investigate what ARGs of less critically important antimicrobials correlate with. As seen in Figure 3, ARGs for pleuromutilin resistance (IA) and for tetracycline resistance (HIA) were found to interact with many other classes, including those that are critically important. For example, pleuromutilin ARGs are few but well connected (Figure 3, Figure S6), as seen with the connections with *lsa(E)* and *cfr(C)* (Figure S8a). Tetracycline ARGs correlated with the abundance of multiple ARGs, such as those conferring resistance to lincosamides, macrolides, and phenicols (Figure 3). While there were many ARGs for tetracycline resistance, they often correlated to the same ARGs (Figure S8b).

## Discussion

Considering the complexity of microbiomes, studying how microbial composition shapes the distribution of ARGs is a challenging task but one that could shed light on how ARGs indirectly select one another. However, with the high-throughput sequencing technologies and many metagenomic datasets available in public repositories, it is now much more feasible to extract the patterns of how ARGs co-occur without knowing their microbial origin. Using our recently published collection of 214K metagenomic datasets[25], we have inferred correlations of pairwise ARG abundances to profile which types of resistance influence the shape of resistome and which genes are the key players. To the best of our knowledge, this is the first study to relate ARG abundances on such a large and broad scale.

Our correlation networks revealed that not all ARGs are connected, as we observed that the interactions were largely shaped by the composition of the environmental resistome. Nonetheless, we did not always observe strong correlations if an ARG was highly abundant. The all-network with correlation for all metagenomic samples was the most complex, which

we speculate is due to both the wide variety of sampling sources and the overrepresentation of human metagenomes (Table 1). We found the differences in the animal and environmental networks much more interesting, as they reflect the dynamics of ARG abundances under exposure to different antimicrobials (Figure 1). On the level of which antimicrobial class an ARG confers resistance to, we could also observe that some resistance classes had more and stronger correlations (Figure S6). There are several cases of our observed co-abundant ARGs that have been reported in the literature, and some of the strongest correlation pairs have been detected due to the ARGs sitting together in the same genome. For example, the gene cassette *vanHAX* has been found together with *msr(C)* and *aac(6')* in genomes of human isolates[52] and *mef(A)* linked with *tet(O)*[53] and *mdf(A)* with *blaTEM, aph(6), sul2, and, tet(A)*[54] (Figure S7).

As highlighted in Figure 2 and Figure 3, we argue that the correlations can serve as a way to profile the risk of co-selection of ARGs occurring in a setting if exposed to an antimicrobial. Some of these relationships between ARGs of different antimicrobial classes have been observed in other studies.

Antimicrobials have been classified differently to reflect their importance to human medicine, of which glycopeptides and macrolides are CIA with the highest priorities. Glycopeptide and macrolide resistance has previously been linked genetically[6,8] in pigs, where we also can report the presence of correlations between ARGs of glycopeptide and macrolide resistances not only in pigs but also in human and mouse environments (Figure 2). Pleuromutilin and tetracycline antimicrobials have less critical importance of antimicrobials [44], but many correlations for ARGs of pleuromutilin and tetracycline exist in various networks (Figure 3) suggest that there are still risks associated with the use of these two. Tetracycline resistance has been found to occur together with resistance to macrolides[45–47], aminoglycosides[45,48], folate pathway antagonists[49,50], lincosamide[10], and beta-lactams[48], to name a few studies. This high connectivity of tetracycline ARGs seems in line with our results, as this specific group of ARGs was connected to almost all classes of antimicrobials in most of our networks (Figure 3, Figure S8b). The variety of connections to antimicrobials of less importance to human health should be more in focus, as our results

show that there are risks of critical antimicrobial resistances emerging from the enrichment of less essential resistance genes.

Following this line of thought of focusing on ARGs that gives resistance to less critically important antimicrobials, a recent study by Tarek and Garner (2022)[51] proposed to create a monitoring framework based on isolated components, or clusters, in correlation networks. They constructed a correlation network encompassing ARG abundances in samples from wastewater treatment plants and argued that representative ARG members from each cluster in their network should be monitoring targets. Our networks suggest that limiting to only a few ARGs would fail to capture the complete picture in some environments, such as human microbiomes (Figure 1). However, as our risk profiles show, limiting the set of monitoring targets by only focusing on what happens to ARG abundances during exposure to one type of antimicrobials would be a way to implement a monitoring system (Figures 2-3, Figure S7-Figure S8). If a monitoring system is implemented, it would need to be updated regularly to show the changes in antimicrobial usage and ARGs co-occurrences since the correlations we have inferred in this study only reflects the current and past usage of antimicrobials.

In order to use correlation networks for surveillance of AMR, more work is needed to confirm that the observed interactions do indeed exist in nature[55]. We have defined interactions as being indirect since we need to determine the degree that co-abundance is explained by physical linkage on a genome or plasmid or sharing of a cell. To investigate the direction of correlation, for example, whether *vanHAX* influences *msr(C)* or the other way around (Figure S6a), the SPIEC-EASI[56] method could be used to infer such directional dependences. A directional correlation could be included in a risk profile.

We have only included ARGs of different resistance classes that are co-abundant in our analysis, but the abundance of bacterial reads could also have been analyzed. By including the bacterial counts, it would likely be possible to understand how the dynamics of microbiomes and resistomes are linked. Similarly, redoing the alignment procedure to include the count of reads aligned to other drivers of resistance, e.g., mobile genetic elements, could also lead to the discovery of new patterns. There has been much work on

linking the prevalence of ARGs with antimicrobial usage (AMU) in various settings[6,8,10], and incorporating AMU could confirm our speculation that many of the co-abundances are driven by selective pressures of various antimicrobials. Pooling the various data types together in this kind of analysis would be a complicated task that nevertheless would shed light on how the abundance and selective pressures contribute to the shape of resistomes.

By utilizing the wealth of information on ARG abundances available in a collection of 214K metagenomic datasets, we have studied the co-abundance of ARGs to discover how these interactions shape the prevalence of resistances in different environments. The inferred correlation networks provide insights into how two resistance types indirectly and species-independent select for each other in different habitats. Our results further highlight that there are instances of genes of one type of resistance often co-occurring with many other types of resistance and that the environmental context plays an important role, revealing them as important targets in surveillance programs to limit their impact on global health.

## References


1. O'Neill, J. & The Review on Antimicrobial Resistance (Chaired by Jim O'Neill). Tackling Drug-Resistant Infections Globally: Final Report and Recommendations. *Rev. Antimicrob. Resist.* 1–80 (2016).
2. Murray, C. J. *et al.* Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* **399**, 629–655 (2022).
3. Holmes, A. H. *et al.* Understanding the mechanisms and drivers of antimicrobial resistance. *Lancet* **387**, 176–187 (2016).
4. Collignon, P., Powers, J. H., Chiller, T. M., Aidara-Kane, A. & Aarestrup, F. M. World health organization ranking of antimicrobials according to their importance in human medicine: a critical step for developing risk management strategies for the use of antimicrobials in food production animals. *Clin. Infect. Dis.* **49**, 132–141 (2009).
5. Jensen, H. H. & Hayes, D. J. Impact of Denmark's ban on antimicrobials for growth promotion. *Curr. Opin. Microbiol.* **19**, 30–36 (2014).
6. Aarestrup, F. M. *et al.* Effect of abolishment of the use of antimicrobial agents for growth promotion on occurrence of antimicrobial resistance in fecal enterococci from

food animals in Denmark. *Antimicrob. Agents Chemother.* **45**, 2054–2059 (2001).
7. Sun, J. *et al.* Development of aminoglycoside and β-lactamase resistance among intestinal microbiota of swine treated with lincomycin, chlortetracycline, and amoxicillin. *Front. Microbiol.* **5**, 580 (2014).
8. Aarestrup, F. M. Characterization of glycopeptide-resistant Enterococcus faecium (GRE) from broilers and pigs in Denmark: Genetic evidence that persistence of GRE in pig herds is associated with coselection by resistance to macrolides. *J. Clin. Microbiol.* **38**, 2774–2777 (2000).
9. Looft, T. *et al.* In-feed antibiotic effects on the swine intestinal microbiome. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 1691–1696 (2012).
10. Andersen, V. D. *et al.* Predicting effects of changed antimicrobial usage on the abundance of antimicrobial resistance genes in finisher' gut microbiomes. *Prev. Vet. Med.* **174**, 104853 (2020).
11. Peterson, L. R. Squeezing the antibiotic balloon: The impact of antimicrobial classes on emerging resistance. *Clin. Microbiol. Infect. Suppl.* **11**, 4–16 (2005).
12. Paterson, D. L. 'Collateral damage' from cephalosporin or quinolone antibiotic therapy. *Clin. Infect. Dis.* **38**, 341–345 (2004).
13. Barberán, A., Bates, S. T., Casamayor, E. O. & Fierer, N. Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J.* **6**, 343–351 (2012).
14. Yuan, M. M. *et al.* Climate warming enhances microbial network complexity and stability. *Nat. Clim. Chang.* **11**, 343–348 (2021).
15. Zhang, Z. *et al.* The large-scale spatial patterns of ecological networks between phytoplankton and zooplankton in coastal marine ecosystems. *Sci. Total Environ.* **827**, 154285 (2022).
16. Li, L. G., Xia, Y. & Zhang, T. Co-occurrence of antibiotic and metal resistance genes revealed in complete genome collection. *ISME J.* **11**, 651–662 (2017).
17. Ju, F. *et al.* Antibiotic resistance genes and human bacterial pathogens: Co-occurrence, removal, and enrichment in municipal sewage sludge digesters. *Water Res.* **91**, 1–10 (2016).
18. Pan, X., Lin, L., Zhang, W., Dong, L. & Yang, Y. Metagenome sequencing to unveil the resistome in a deep subtropical lake on the Yunnan-Guizhou Plateau, China. *Environ.*

*Pollut.* **263**, 114470 (2020).

19. Wu, J. *et al.* Prevalence and distribution of antibiotic resistance in marine fish farming areas in Hainan, China. *Sci. Total Environ.* **653**, 605–611 (2019).

20. Liu, J. *et al.* The fecal resistome of dairy cattle is associated with diet during nursing. *Nat. Commun.* **10**, (2019).

21. Munk, P. *et al.* Abundance and diversity of the faecal resistome in slaughter pigs and broilers in nine European countries. *Nat. Microbiol.* **3**, 898–908 (2018).

22. Hendriksen, R. S. *et al.* Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nat. Commun.* **10**, (2019).

23. Martiny, H.-M. *et al.* Global distribution of mcr gene variants in 214,095 metagenomic samples. *mSystems* (2022). doi:10.1128/msystems.00105-22

24. Karkman, A., Do, T. T., Walsh, F. & Virta, M. P. J. Antibiotic-Resistance Genes in Waste Water. *Trends Microbiol.* **26**, 220–228 (2018).

25. Martiny, H.-M., Id, P. M., Id, C. B., Aarestrup, F. M. & Petersenid, T. N. A curated data resource of 214K metagenomes for characterization of the global antimicrobial resistome. *PLOS Biol.* **20**, e3001792 (2022).

26. Aitchison, J. The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. Ser. B* **44**, 139–160 (1982).

27. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.* **8**, 1–6 (2017).

28. Friedman, J. & Alm, E. J. Inferring Correlation Networks from Genomic Survey Data. *PLoS Comput. Biol.* **8**, 1–11 (2012).

29. Leinonen, R. *et al.* The European nucleotide archive. *Nucleic Acids Res.* **39**, 44–47 (2011).

30. Bushnell, B. BBMap. (2014).

31. Clausen, P. T. L. C., Aarestrup, F. M. & Lund, O. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics* **19**, 1–8 (2018).

32. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **67**, 2640–2644 (2012).

33. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41**, 590–596 (2013).

34. Edgar, R. C. & Bateman, A. Search and clustering orders of magnitude faster than

BLAST. *Bioinformatics* **26**, 2460–2461 (2010).

35. Team, R. C. R: A Language and Environment for Statistical Computing. (2021).

36. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **Complex Sy**, 1695 (2006).

37. Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D. & Borsboom, D. qgraph: Network Visualizations of Relationships in Psychometric Data. *J. Stat. Softw.* **48**, 1–18 (2012).

38. Pedersen, T. L. ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. (2021).

39. Fructherman, T. M. . & Reingold, E. M. Graph Drawing by Force-directed Placement. *Softw. - Pract. Exp.* **21**, 1129–1164 (1991).

40. Barrat, A., Barthélemy, M., Pastor-Satorras, R. & Vespignani, A. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 3747 (2004).

41. Parente, E., Zotta, T., Faust, K., De Filippis, F. & Ercolini, D. Structure of association networks in food bacterial communities. *Food Microbiol.* **73**, 49–60 (2018).

42. Wasserman, S. & Faust, K. *Social network analysis: Methods and applications.* (Cambridge University Press, 1994).

43. Silver, N. C. & Dunlap, W. P. Averaging Correlation Coefficients: Should Fisher's z Transformation Be Used? *J. Appl. Psychol.* **72**, 146–148 (1987).

44. Group, W. H. O. A., Surveillance, I. & Resistance, A. *WHO | WHO list of Critically Important Antimicrobials (CIA)*.

45. Arredondo, A., Blanc, V., Mor, C., Nart, J. & León, R. Tetracycline and multidrug resistance in the oral microbiota: differences between healthy subjects and patients with periodontitis in Spain. *J. Oral Microbiol.* **13**, (2021).

46. Nielsen, H. U. K. *et al.* Tetracycline and macrolide co-resistance in Streptococcus pyogenes: Co-selection as a reason for increase in macrolide-resistant S. pyogenes? *Microb. Drug Resist.* **10**, 231–238 (2004).

47. Ardic, N., Ozyurt, M., Sareyyupoglu, B. & Haznedaroglu, T. Investigation of erythromycin and tetracycline resistance genes in methicillin-resistant staphylococci. *Int. J. Antimicrob. Agents* **26**, 213–218 (2005).

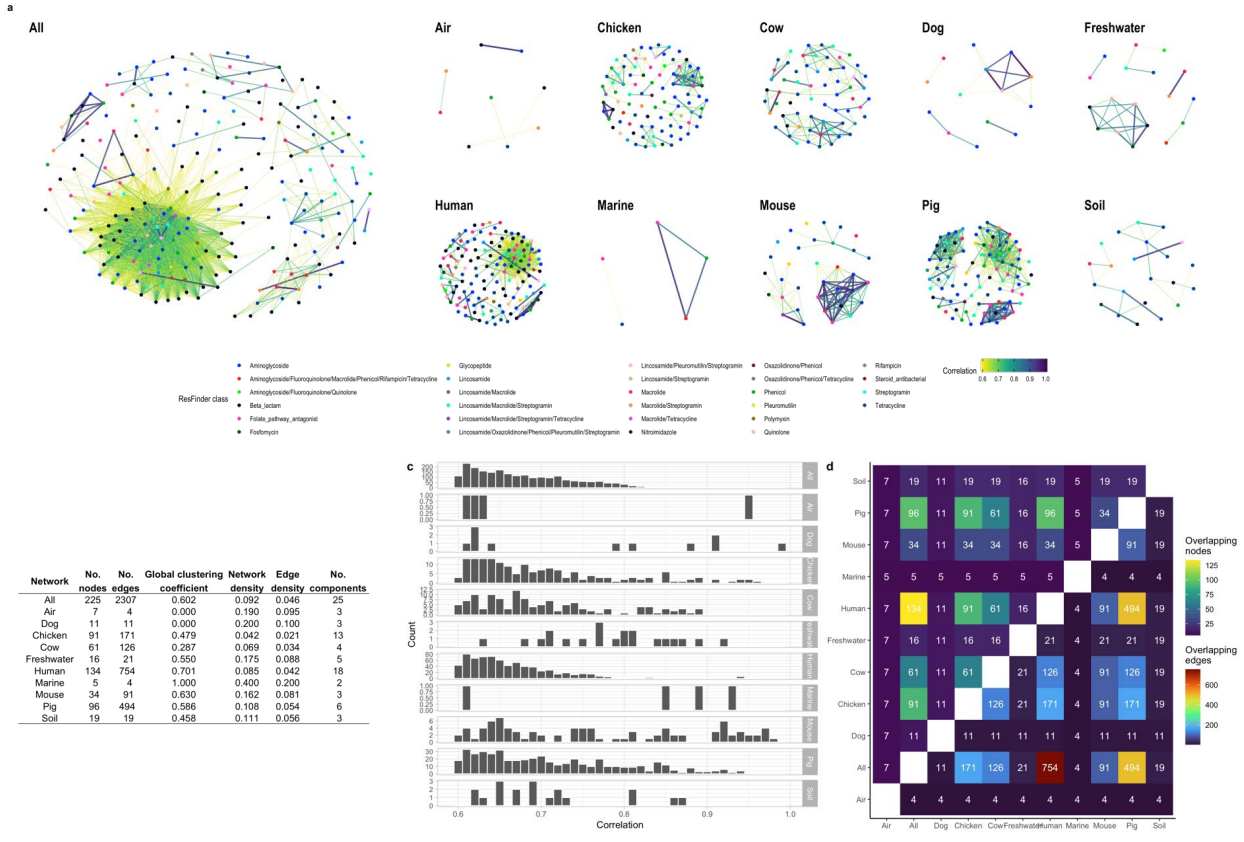
48. Stepanauskas, R. *et al.* Coselection for microbial resistance to metals and antibiotics in freshwater microcosms. *Environ. Microbiol.* **8**, 1510–1514 (2006).

49. Gao, P., Munir, M. & Xagoraraki, I. Correlation of tetracycline and sulfonamide antibiotics with corresponding resistance genes and resistant bacteria in a conventional municipal wastewater treatment plant. *Sci. Total Environ.* **421–422**, 173–183 (2012).
50. Zhang, M. Q., Yuan, L., Li, Z. H., Zhang, H. C. & Sheng, G. P. Tetracycline exposure shifted microbial communities and enriched antibiotic resistance genes in the aerobic granular sludge. *Environ. Int.* **130**, 104902 (2019).
51. Tarek, M. H. & Garner, E. A proposed framework for the identification of indicator genes for monitoring antibiotic resistance in wastewater: Insights from metagenomic sequencing. *Sci. Total Environ.* 158698 (2022). doi:10.1016/J.SCITOTENV.2022.158698
52. Freitas, A. R. *et al.* High-Resolution Genotyping Unveils Identical Ampicillin-Resistant Enterococcus faecium Strains in Different Sources and Countries: A One Health Approach. *Microorganisms* **10**, (2022).
53. Giovanetti, E., Brenciani, A., Lupidi, R., Roberts, M. C. & Varaldo, P. E. Presence of the tet(O) gene in erythromycin- and tetracycline-resistant strains of Streptococcus pyogenes and linkage with either the mef(A) or the erm(A) gene. *Antimicrob. Agents Chemother.* **47**, 2844–2849 (2003).
54. Trongjit, S. & Chuanchuen, R. Whole genome sequencing and characteristics of Escherichia coli with coexistence of ESBL and mcr genes from pigs. *PLoS One* **16**, 1–17 (2021).
55. Blanchet, F. G., Cazelles, K. & Gravel, D. Co-occurrence is not evidence of ecological interactions. *Ecol. Lett.* **23**, 1050–1063 (2020).
56. Kurtz, Z. D. *et al.* Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Comput. Biol.* **11**, e1004226 (2015).

| Resistance class | All | Air | Dog | Chicken | Cow | Freshwater | Human | Marine | Mouse | Pig | Soil |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Aminoglycoside (144) | 9.12 | 17.64 | 8.42 | 11.51 | 5.24 | 16.70 | 8.51 | 0.08 | 10.12 | 14.08 | 1.58 |
| Aminoglycoside/Fluoroquinolone/Macrolide/Phenicol/Rifampicin/Tetracycline (1) | 1.08 | 0.63 | 0.52 | 0.12 | 0.07 | 0.34 | 1.08 | 0.20 | 1.20 | 0.17 | 0.02 |
| Aminoglycoside/Fluoroquinolone/Quinolone (1) | 0.19 | <0.01 | <0.01 | 0.09 | - | 0.46 | 0.15 | <0.01 | 0.01 | <0.01 | <0.01 |
| Beta_lactam (227) | 20.55 | 41.60 | 20.07 | 0.79 | 1.64 | 8.44 | 18.50 | 1.85 | 3.02 | 5.59 | 1.89 |
| Folate_pathway_antagonist (37) | 3.20 | 2.92 | 0.97 | 2.19 | 0.41 | 5.45 | 3.13 | 0.36 | 4.33 | 1.03 | 0.82 |
| Fosfomycin (28) | 1.07 | 4.36 | 0.01 | 0.20 | <0.01 | 1.29 | 1.49 | <0.01 | 0.22 | 0.03 | 0.19 |
| Glycopeptide (27) | 0.29 | <0.01 | - | 0.01 | <0.01 | 0.01 | 0.58 | <0.01 | 0.72 | 0.02 | 0.01 |
| Lincosamide (9) | 3.33 | 1.88 | 29.00 | 3.42 | 8.29 | 0.36 | 2.02 | <0.01 | 4.01 | 11.61 | 0.71 |
| Lincosamide/Macrolide (1) | <0.01 | <0.01 | - | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | - | - | - |
| Lincosamide/Macrolide/Streptogramin (44) | 12.33 | 6.12 | 1.35 | 14.51 | 2.26 | 0.32 | 17.71 | 0.03 | 8.19 | 9.52 | 5.95 |
| Lincosamide/Macrolide/Streptogramin/Tetracycline (1) | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | 0.07 | <0.01 | <0.01 |
| Lincosamide/Oxazolidinone/Phenicol/Pleuromutilin/Streptogramin (3) | 0.49 | 0.41 | 0.01 | 1.36 | 0.50 | <0.01 | 0.45 | <0.01 | 1.64 | 1.25 | 0.04 |
| Lincosamide/Pleuromutilin/Streptogramin (7) | 0.27 | 0.48 | 0.18 | 0.09 | 0.02 | 0.02 | 0.34 | <0.01 | <0.01 | 0.21 | 0.05 |
| Lincosamide/Streptogramin (2) | 0.28 | 0.12 | <0.01 | 0.02 | <0.01 | 0.01 | 0.38 | <0.01 | 0.06 | <0.01 | <0.01 |
| Macrolide (24) | 3.04 | 1.86 | 5.78 | 0.36 | 16.71 | 1.90 | 2.53 | <0.01 | 4.22 | 6.83 | 0.25 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Macrolide/Streptogramin (4) | 1.10 | 1.69 | 1.01 | 0.12 | 0.39 | 1.18 | 1.43 | <0.01 | 1.83 | 0.48 | 0.11 |
| Macrolide/Tetracycline (1) | <0.01 | <0.01 | - | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | - | <0.01 | <0.01 |
| Nitroimidazole (9) | 0.32 | <0.01 | 0.04 | 0.01 | 0.05 | <0.01 | 0.56 | <0.01 | 0.02 | 0.14 | <0.01 |
| Oxazolidinone/Phenicol (1) | 0.02 | 0.03 | - | 0.10 | <0.01 | <0.01 | <0.01 | <0.01 | - | 0.09 | <0.01 |
| Oxazolidinone/Phenicol/ Tetracycline (1) | 0.07 | 0.02 | <0.01 | 1.54 | <0.01 | <0.01 | <0.01 | <0.01 | 0.03 | 0.04 | <0.01 |
| Phenicol (33) | 9.21 | 6.19 | 0.36 | 6.10 | 4.33 | 55.89 | 3.33 | 97.37 | 0.14 | 0.81 | 85.04 |
| Pleuromutilin (1) | 0.06 | 0.02 | - | <0.01 | <0.01 | 0.39 | 0.06 | <0.01 | 0.01 | 0.07 | <0.01 |
| Polymyxin (10) | 1.19 | 0.30 | <0.01 | 0.14 | <0.01 | 1.85 | 1.26 | <0.01 | 0.18 | 0.14 | 0.12 |
| Quinolone (17) | 0.11 | <0.01 | - | 0.01 | - | 0.18 | 0.06 | <0.01 | - | <0.01 | <0.01 |
| Rifampicin (6) | 0.06 | 0.09 | - | <0.01 | <0.01 | <0.01 | 0.03 | <0.01 | <0.01 | <0.01 | <0.01 |
| Steroid_antibacterial (2) | 0.15 | <0.01 | - | 3.30 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | 0.11 | 0.02 |
| Streptogramin (11) | 32.48 | 13.61 | 32.27 | 53.99 | 60.05 | 5.20 | 36.39 | 0.06 | 59.97 | 47.76 | 3.18 |
| Tetracycline (64) | <0.01 | - | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 | <0.01 |

*Table 2: Relative abundance of read fragments aligned to each resistance class (row) for each sampling source (column). The relative abundance is the number of read fragments aligned to the group out of all fragments aligned to ARGs. For each resistance class, the parenthesis shows the number of ARGs belonging to that category.*




| Network | No. nodes | No. edges | Global clustering coefficient | Network density | Edge density | No. components |
|---|---|---|---|---|---|---|
| All | 225 | 2307 | 0.602 | 0.092 | 0.046 | 25 |
| Air | 7 | 4 | 0.000 | 0.190 | 0.095 | 3 |
| Dog | 11 | 11 | 0.000 | 0.200 | 0.100 | 3 |
| Chicken | 91 | 171 | 0.479 | 0.042 | 0.021 | 13 |
| Cow | 61 | 126 | 0.287 | 0.069 | 0.034 | 4 |
| Freshwater | 16 | 21 | 0.550 | 0.175 | 0.088 | 5 |
| Human | 134 | 754 | 0.701 | 0.085 | 0.042 | 18 |
| Marine | 5 | 4 | 1.000 | 0.400 | 0.200 | 2 |
| Mouse | 34 | 91 | 0.630 | 0.162 | 0.081 | 3 |
| Pig | 96 | 494 | 0.586 | 0.108 | 0.054 | 6 |
| Soil | 19 | 19 | 0.458 | 0.111 | 0.056 | 3 |

*Figure 1: The resistome networks consisting of correlations (edges) between pairs of ARGs (nodes).* ***a****. Each correlation network is visualized, where each ARG node is colored by resistance class, and edges are colored by the correlation coefficient value.* ***b****. Metrics of the correlation networks reveal how interconnected and complex the networks are. No. stands for number of.* ***c****. Distribution of included correlation coefficients in each network. Figure S3 shows the distribution of all inferred correlations and their p-values.* ***d****. A heatmap showing how similar the content of the two networks is. In the upper half of the heatmap, the number of overlapping ARG nodes is shown. An overlap means that a specific ARG has a correlation in both networks, ignoring what it is co-abundant with. In the lower half of the heatmap, the number of overlapping correlation edges is shown. An overlapping edge is defined as whether the correlation coefficient is reported in both networks, regardless of the value of the coefficient.*

*Figure 2:* ***Correlation profiles for ARGs conferring resistance to the critically important antimicrobial classes glycopeptides (left) and macrolides (right).*** *Each column shows the average correlation from, e.g., macrolide ARGs to ARGs for other antimicrobial classes. The circle is colored by the average correlation, where a white circle indicates no statistically significant correlations of ARGs observed between the two antimicrobial classes.*

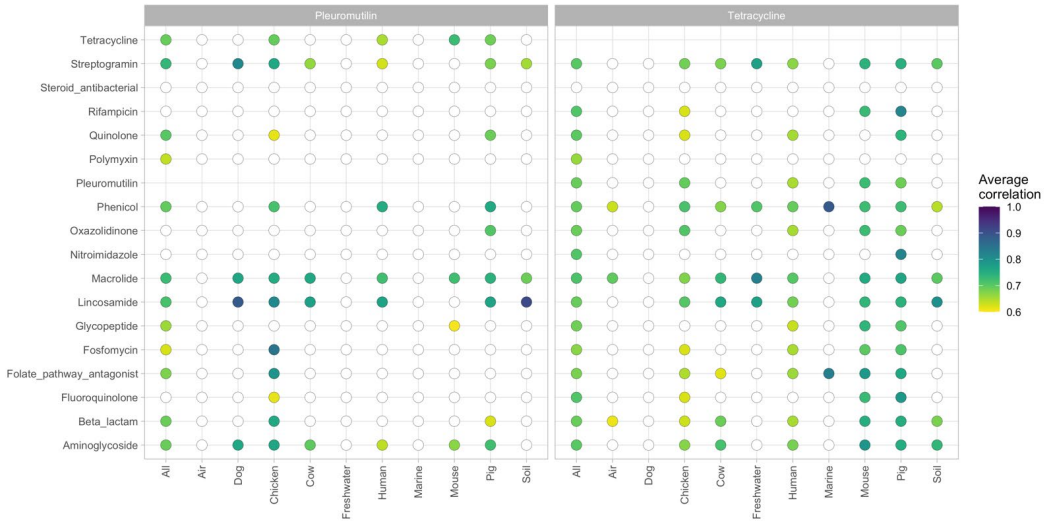



*Figure 3:* ***Correlation profiles for ARGs conferring resistance to important pleuromutilins (left) and highly important tetracyclines (right).*** *Each column shows the average correlation from, e.g., tetracycline ARGs to ARGs for other antimicrobial classes. The circle is colored by the average correlation, where a white circle indicates no statistically significant correlations of ARGs observed between the two antimicrobial classes.*

# Supplementary materials

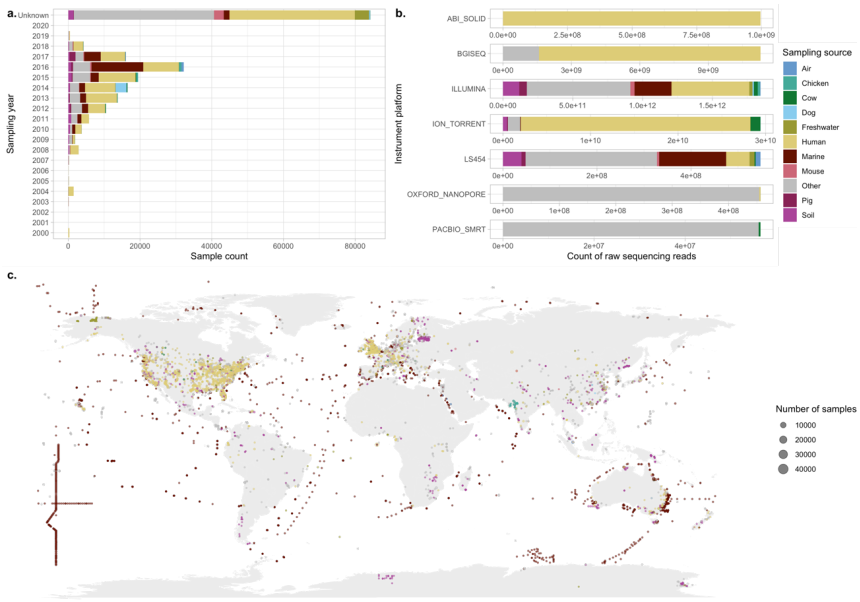



*Figure S1:* ***Metagenomic origins colored by the sampling group****. a. Overview of sampling year for metagenomic samples colored by their sampling origin. 100 samples were taken before 2000, and 84,238 did not have a valid sampling date. b. Overview of the amount of sequencing reads available for each sampling origin. c. Sampling locations of metagenomic samples used in the correlation analysis were split by their sampling source. Number of samples with no coordinates available; All: 83,361; Air: 16; Dog: 3,159; Chicken: 570; Cow: 262; Freshwater: 61; Human: 40,003; Marine: 363; Mouse: 2,842; Pig: 320; Soil: 477. The 'Other' label refers to those that are not in one of the source-specific networks but are included in the 'All' network.*

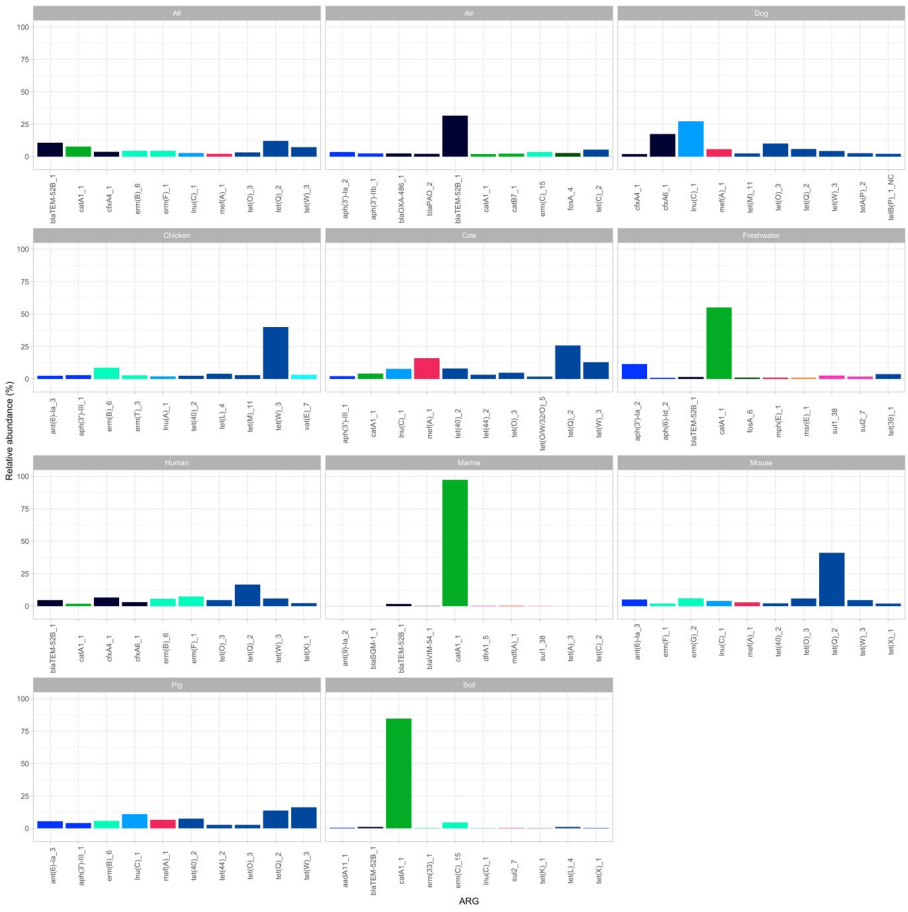



*Figure S2:* ***Top 10 most abundant ARGs per sampling group.*** *A barplot showing the ARGs having the highest 10 relative abundances per sampling source colored by resistance class.*

| | No filters | | | minF | minS | | minF | minS | | minF | minS | | minF | minS | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 25 | 10 | | 50 | 10 | | 50 | 20 | | 50 | 50 | |
| **Source** | **ARG cols** | **#nodes** | **#edges** | **ARG cols** | **#nodes** | **#edges** | **ARG cols** | **#nodes** | **#edges** | **ARG cols** | **#nodes** | **#edges** | **ARG cols** | **#nodes** | **#edges** |
| All | 716 | 443 | 34,815 | 459 | 275 | 4,786 | 425 | 257 | 2,970 | 313 | 165 | 320 | 313 | 165 | 320 |
| Air | 436 | 565 | 66,612 | 34 | 14 | 10 | 19 | 9 | 6 | 0 | - | - | 0 | - | - |
| Chicken | 391 | 634 | 92,563 | 145 | 123 | 289 | 123 | 191 | 226 | 69 | 56 | 87 | 69 | 56 | 87 |
| Cow | 445 | 614 | 92,881 | 94 | 75 | 242 | 85 | 68 | 198 | 37 | 25 | 57 | 37 | 25 | 57 |
| Dog | 158 | 72 | 85 | 45 | 25 | 35 | 45 | 25 | 35 | 45 | 25 | 35 | 45 | 25 | 35 |
| Freshwater | 579 | 589 | 43,054 | 56 | 40 | 74 | 36 | 24 | 33 | 1 | - | - | 1 | - | - |
| Human | 692 | 474 | 53,408 | 307 | 179 | 1,445 | 278 | 158 | 902 | 201 | 91 | 120 | 201 | 91 | 120 |
| Marine | 668 | 13 | 8 | 24 | 7 | 5 | 10 | 7 | 5 | 4 | - | - | 4 | - | - |
| Mouse | 326 | 549 | 80,254 | 55 | 45 | 122 | 41 | 35 | 107 | 25 | 21 | 70 | 25 | 21 | 70 |
| Pig | 408 | 634 | 87,754 | 136 | 114 | 864 | 118 | 99 | 606 | 61 | 52 | 175 | 61 | 52 | 175 |
| Soil | 614 | 510 | 22,415 | 72 | 53 | 85 | 47 | 28 | 37 | 3 | - | - | 3 | - | - |

*Table S1:* ***Filtering the counts impacts the number of input ARG columns, as well as the resulting number of correlations.*** *minF is the minimum number of fragments available for an ARG in one sample, and minS is the minimum number of samples to fulfill the minF requirement for the column to be included in the inference procedure*

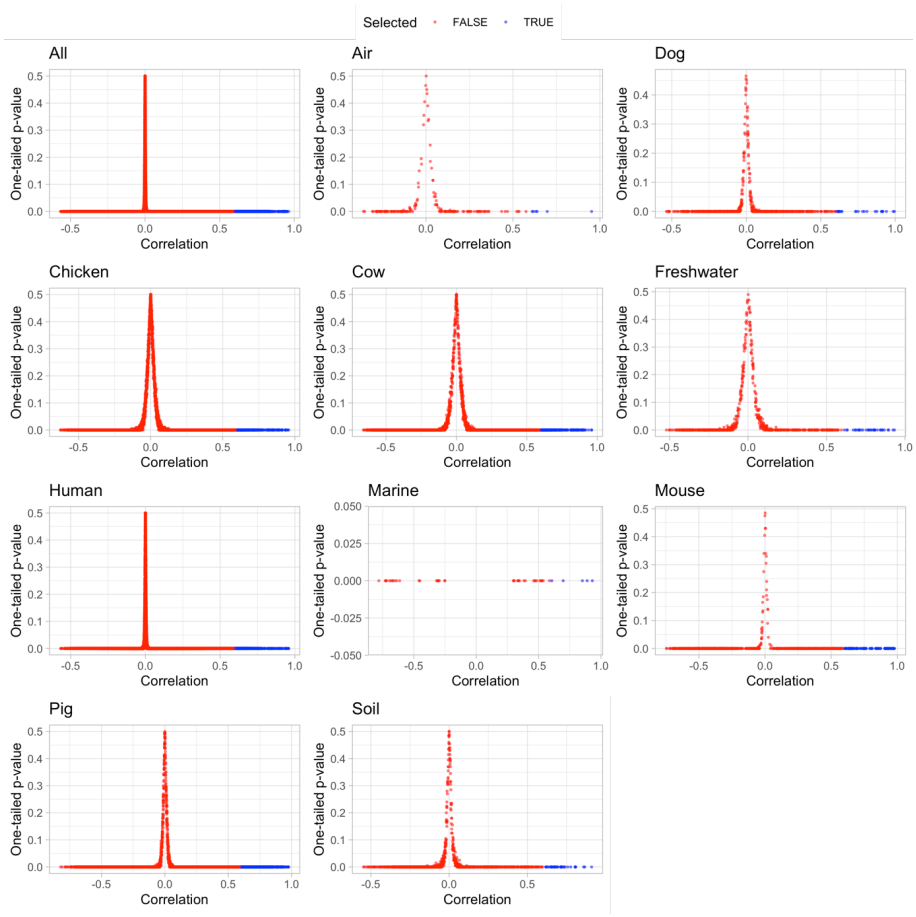



*Figure S3:* ***Distribution of correlations and p-values produced by SparCC for each data grouping****. Each point is a correlation between two ARGs and is colored by whether the point was selected if the p-value $< 0.01$ and correlation $\geq 0.6$.*

*Figure S4:* ***The relative abundance, number of ARGs, and the number of correlations for each resistance class in each sampling group.*** *The relative abundance in green shows the percentage of read fragments for each class. In grey is the number of ARGs for each resistance class. Orange coloring indicates the number of correlation coefficients inferred for ARGs in a resistance class.*

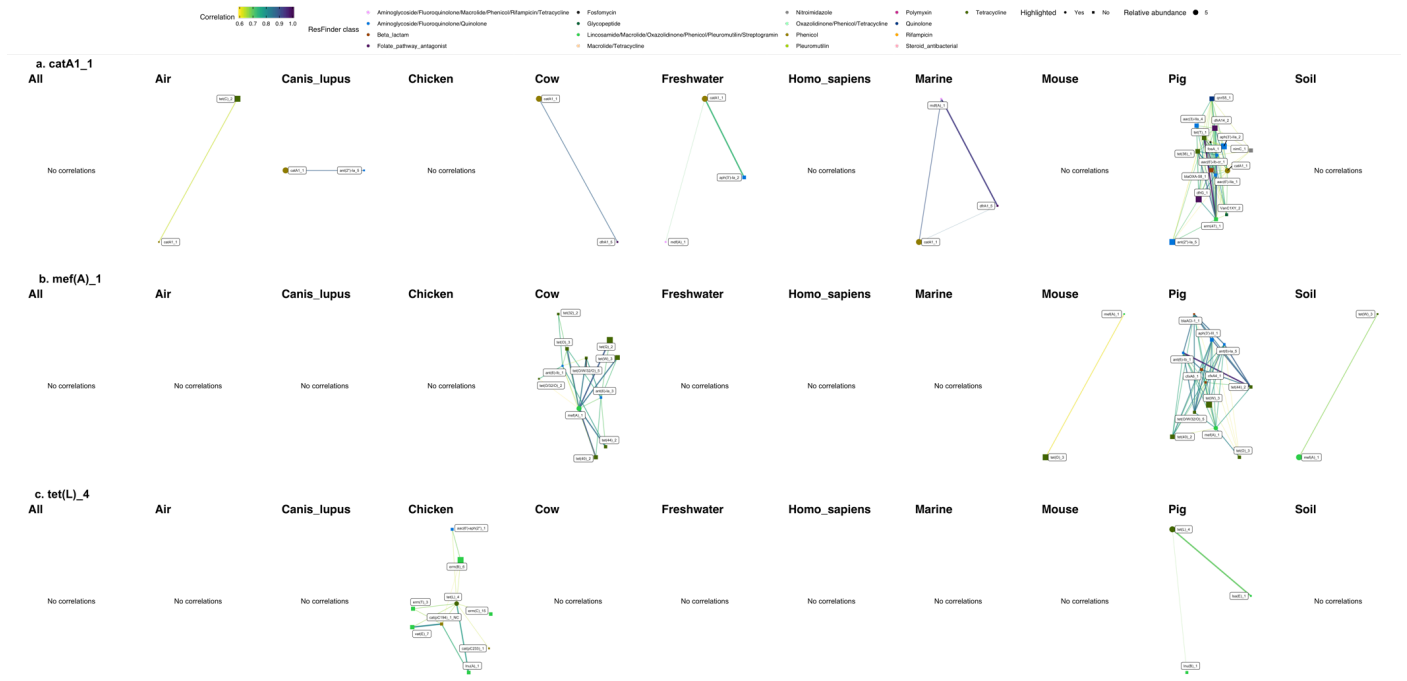



*Figure S5:* ***Networks illustrating how one ARG interacts differently depending on the sampling environments****. The three genes catA1_1, mef(A)_1, and tet(L)_4 are selected to illustrate that the genes correlate more or less with other genes in their abundance in some habitats.*

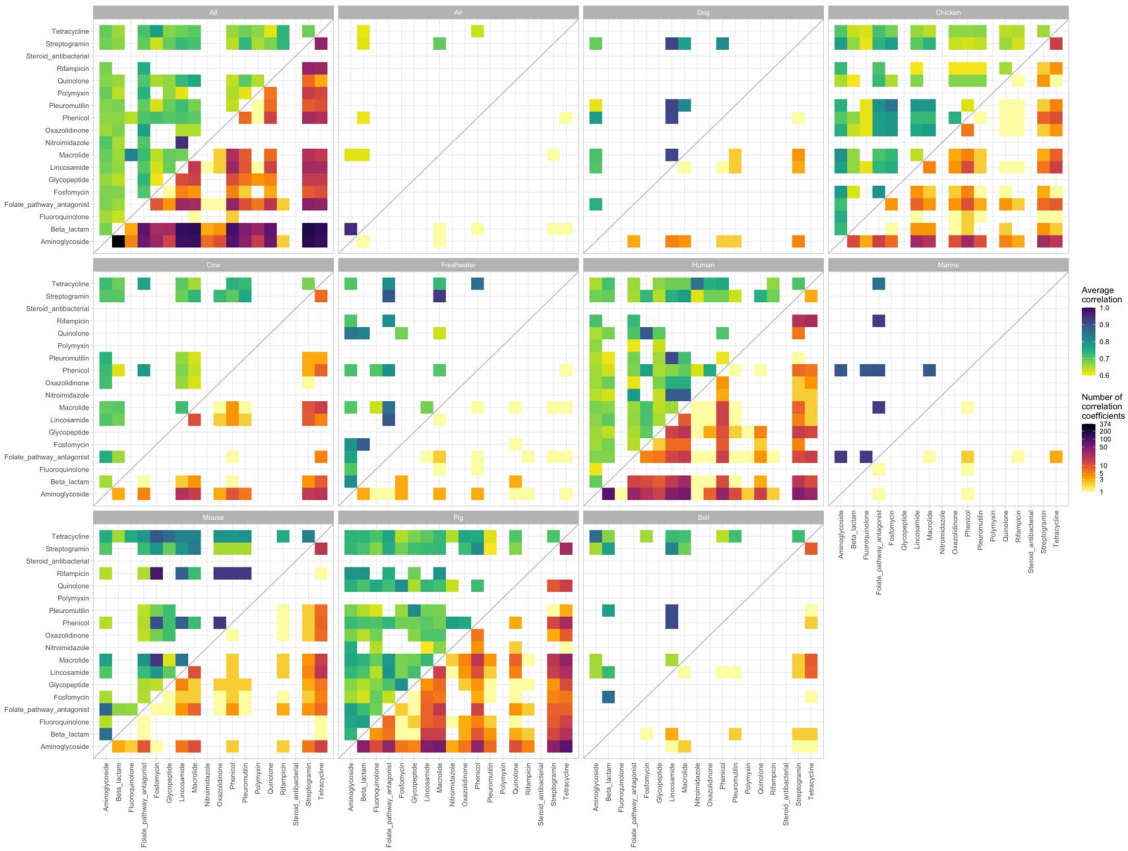



*Figure S6: The average correlation between ARGs of different resistance classes in the upper triangle of the heatmaps with the lower half shows the correlation coefficients between the two classes. Note that a correlation coefficient between two ARGs might be present in more than one tile, as some ARGs confer resistance to multiple classes of antimicrobials.*

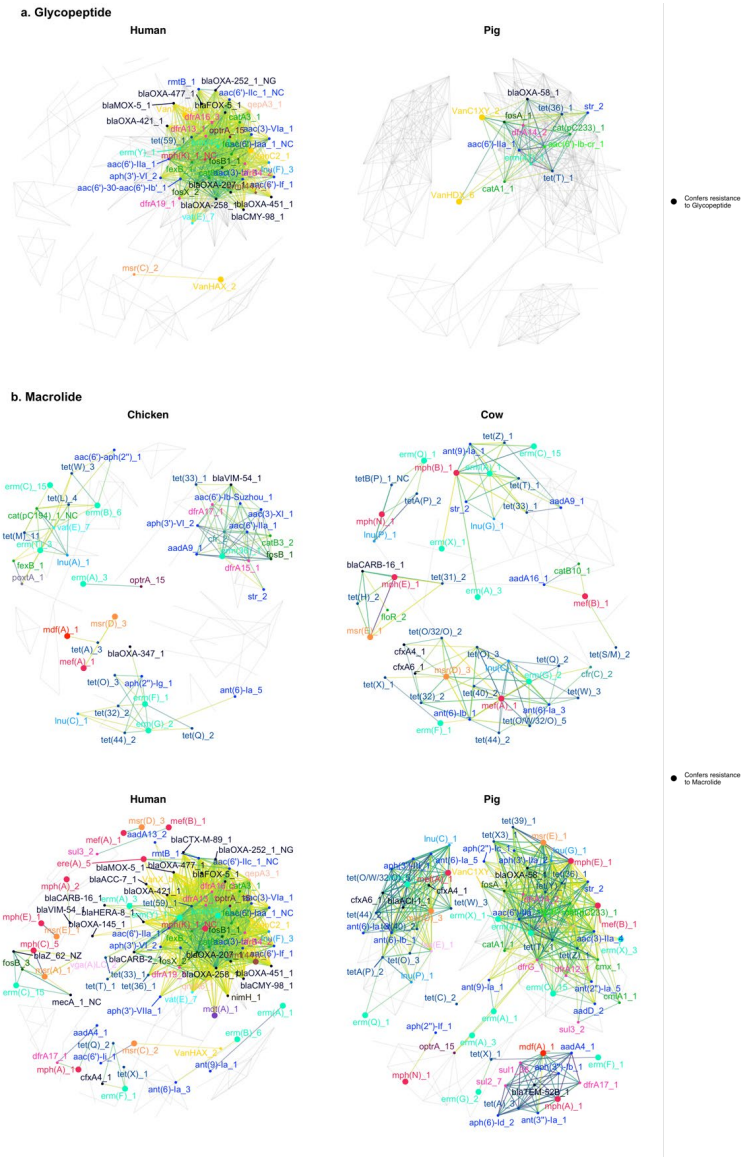



*Figure S7*: ***Highlights of interactions between ARGs of a. glycopeptide resistance and b. macrolide resistance***. *The ARG node size shows whether the corresponding ARG gives resistance to the antimicrobial class in focus. Only correlation edges and ARGs nodes are colored if they correlate with the highlighted ARGs; otherwise, they are colored grey. The coloring schemes for nodes and edges are given in Figure 1a.*

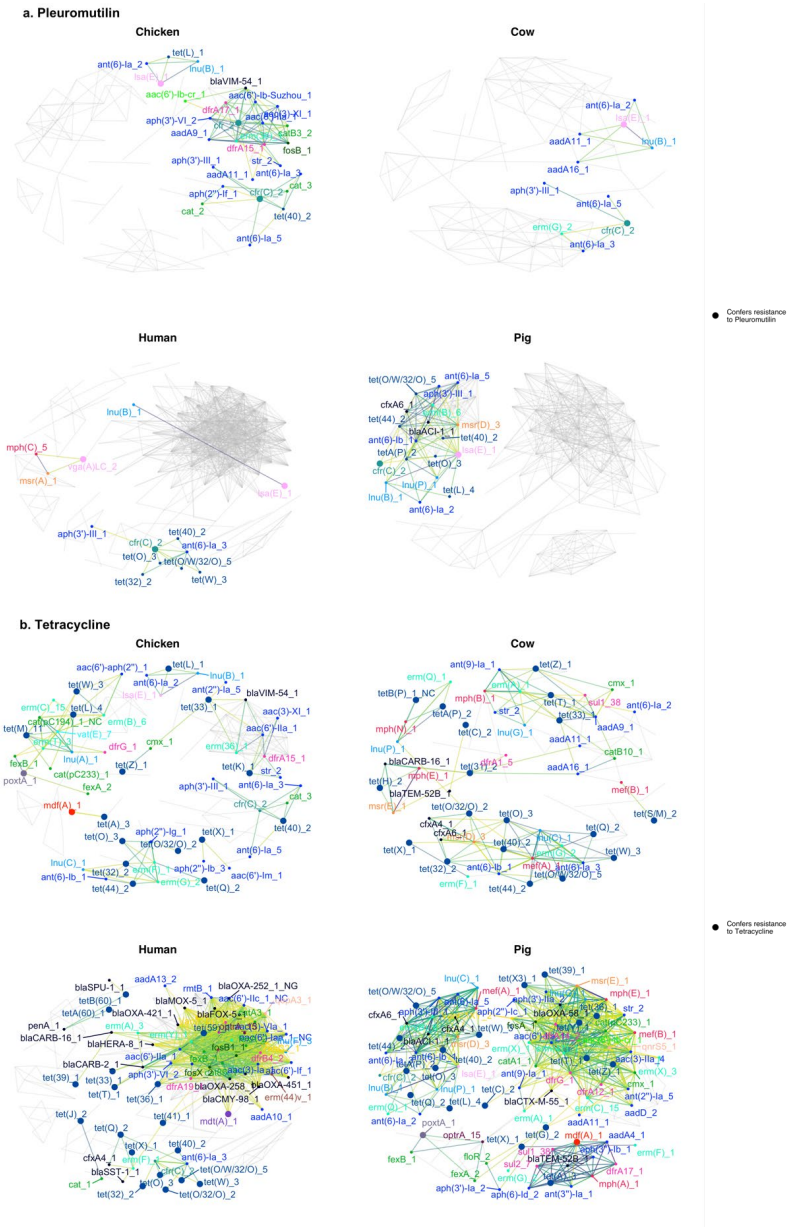



*Figure S8:* ***Highlights of interactions between ARGs of a. pleuromutilin resistance and b. tetracycline resistance.*** *The ARG node size shows whether the corresponding ARG gives resistance to the antimicrobial class in focus. Only correlation edges and ARGs nodes are colored if they correlate with the highlighted ARGs; otherwise, they are colored grey. The coloring schemes for nodes and edges are given in Figure 1a.*

# Part III

# Conclusion

CHAPTER 7

# Conclusion

The threat of AMR is not only a human health concern but also impacts animal and environmental health. It is a global problem that needs to be tackled with targeted solutions, one of them being establishing the surveillance of ARGs embracing the three pillars of the One Health approach. To date, most AMR studies have only focused on the prevalence of ARGs in specific genomes or environmental settings, which does not incorporate that ARGs spread across borders, environments, and hosts. Due to the many advances in NGS technologies and their frequent use in research, there is a vast amount of sequencing datasets available in public repositories ready to be reanalyzed. By analyzing the composition of both the microbiome and resistome in a multitude of environments with publicly available metagenomic datasets, new patterns of how ARGs emerge, disseminate, and evolve can be studied in much more detail.

In **Manuscript I**, the study presented laid the foundation for using publicly available host and environmental samples for worldwide surveillance of AMR. The steps for retrieving and aligning sequencing reads and curating the metadata of the 214,095 metagenomic samples shared between 2010 and 2020 on ENA were described in detail. A great effort was put into standardizing the metadata labels, underlining the issues regarding the information accompanying the samples. Metadata needs to be as error-free as possible in order to place the genetic information in the correct context. Despite these issues, the analysis of the $442 \cdot 10^{12}$ basepairs of sequencing reads revealed that there were geographical, temporal, and environmental differences in ARG abundances. These differences suggest that new patterns of AMR dissemination can be discovered by digging deeper into the data collection. In the spirit of following the FAIR principles, we have shared the mapping results to allow other researchers to take advantage of our data.

For the analysis of public metagenomic datasets to be truly powerful, the sample metadata needs to be as correct as possible. Unfortunately, there are still issues with how such attributions are recorded, primarily due to different iterations of how this kind of data is entered during the data-sharing process. A significant proportion of my time was spent fixing metadata, but many of the 214K metagenomic samples were still not usable. For example, if a metagenomic sample is annotated as a 'gut metagenome', we have no idea which animal was sampled. There is also the obstacle that we rely on the sequencing projects of the broader scientific community, which will

inherently introduce biases in what kind of samples are shared and how the samples have been sequenced. While these biases cannot be entirely removed, they should be acknowledged when putting the results of downstream analyses into a global context.

For the research presented in **Manuscript II**, the goal was to investigate the occurrence of the family of *mcr* genes in the 214K metagenomes to showcase the potential of the collection built in the first manuscript. The *mcr* genes confer mobilized resistance to the antibiotic colistin, which is only used when all other treatment options fail. After the discovery of *mcr*-1 in 2015, several reports of new *mcr* gene members and their widespread dissemination have been reported. With the variety of sampling origins in our collection, we decided to investigate the abundance of *mcr* sequencing reads to pinpoint differences and similarities in samples of different origins. Our results did confirm the global spread of the *mcr* genes and the notion that the colistin ARGs had circulated in the environment for a while before being discovered. What is more interesting is we found that *mcr*-9, reported in 2019, was the most abundant gene and that there was evidence of an unknown *mcr*-9 variant, showcasing how novel findings can be extracted from metagenomic data.

Since the sequencing reads generated represent a random sample of the environment, we do not know the true abundance of bacteria and genes. For example, we observed a low abundance of *mcr*-6 fragments in only one location, and while that did confirm that this gene is very rare, the low abundance needs to be regarded with a bit of skepticism; if only a few read fragments aligned, is that sufficient evidence to conclude that the ARG is there and it is not only an error stemming from either the sequencing protocol or alignment? In other words, such results need to be investigated in more detail, for example, by comparing them with what other surveillance programs find. As we saw by comparing *mcr* abundances in our metagenomes with the NCBI Pathogen Detection Project results, some ARGs were not captured by only looking at pathogen genomes, and others were not found in the metagenomic data. Each of the two resources was better at capturing some *mcr* genes than the other, emphasizing that combining both surveillance efforts will strengthen our understanding of ARG.

With **Manuscript III**, we wanted to investigate the co-abundance of ARG pairs in multiple environmental contexts to assess collateral damages happening under exposure to an antimicrobial. Using the read abundances of all ARGs in the full metagenomic collection, we concluded that the amount of pairwise ARG correlations inferred and the strength of these highly depended on the sampling origin. Moreover, the correlations on the gene levels were averaged for each pair of antimicrobial resistance classes to construct risk profiles. These profiles showed what happens in an environment if suddenly one group of ARGs increases in their abundance. In other words, we argued that our risk profiles show what can happen in a resistome if it suddenly comes under selective pressure by an antimicrobial drug. For example, we observed that a small group of ARGs conferring resistance to glycopeptides was responsible for many of the glycopeptide interactions, whereas tetracycline ARGs were

many and had even more correlations to other kinds of antimicrobial class resistances.

Interpreting correlation is a complex task, which was not simplified by the number of samples and correlations inferred by SparCC. Initially, we did not apply any filters on the raw read counts. However, when the resulting correlations were investigated in more detail, we saw a clear tendency that often spurious correlations were reported if an ARG had a low count in one sample out of many. As reflected in the paper, the reasoning behind our filter settings was to balance the sparsity of the samples and the number of correlations inferred per sampling group. In some sampling groups, the correlation networks were still highly complex, likely reflecting the number of samples available and how much the sampling environment has been affected by antimicrobials. Tetracycline ARGs have been reported to co-occur with many other classes of resistance, which we also observed across most of the networks. However, what is more concerning is that ARGs that confer resistance to antimicrobials deemed less critical to human health, e.g., pleuromutilin or tetracycline, did indirectly select ARGs of more important antimicrobials. For example, tetracycline ARGs had correlations with at least one macrolide ARG in most networks. These correlations indicated that even by switching to an antimicrobial, where there is less concern of resistance developing, the ARGs of that antimicrobial might send a ripple through the environment, causing ARGs of other antimicrobials to rise. There is still extra work needed to confirm these effects, such as verifying the correlations with other data sources, and experimentally verifying what happens under changing uses of antimicrobials.

The primary approach of this PhD has been to perform data-driven research, as opposed to the more traditional hypothesis-driven research. While this kind of exploratory research is like looking for a needle in a haystack, mining such large quantities of data can confirm and challenge existing notions on the distribution of ARGs. A significant strength of this metagenomic collection is that the samples span the entire world across multiple environments and years, which would not have been possible to build in the more traditional approach of going out and collecting samples ourselves. The keen observer might have noticed that the 214K collection only contains metagenomes that were shared on ENA up until 2020-01-01. The pool of NGS datasets on ENA has continued to grow, meaning that the collection could be expanded with the metagenomes uploaded since then. Not only has the number of metagenomic samples available grown but so have the reference sequence databases. For the work carried out during this PhD to become a truly global AMR surveillance program, a routine for retrieving newly shared metagenomic datasets should be implemented, and a process for remapping everything when there are enough new genes to look for. Updating the collection with both new samples and reference genes would highlight the value of open and reproducible research, which is a goal of mine to do in the future.

The curated data collection of 214K metagenomes contains a wealth of information ready to be explored in even more detail than what the three included manuscript has

done. With the help of DTU press release[1], the work has already gained some attention online, which will hopefully lead to even more studies using the ARG abundances and further encourage even more researchers to share their sequencing datasets. In conclusion, this PhD has demonstrated the value of reanalyzing public sequencing data for exploring the composition of microbiomes and resistomes.

[1] https://www.food.dtu.dk/english/news/nyhed?id=d27f6275-71de-4c5d-8621-9e8beb79c122

# Bibliography


[1] Rustam I. Aminov. "A brief history of the antibiotic era: Lessons learned and challenges for the future." In: *Frontiers in Microbiology* 1.DEC (2010), p. 134. ISSN: 1664302X. DOI: 10.3389/FMICB.2010.00134/BIBTEX.

[2] Alexander Fleming. "On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B. influenzae." In: *British journal of experimental pathology* 10.3 (1929), p. 226.

[3] Tomoo Saga and Keizo Yamaguchi. "History of Antimicrobial Agents and Resistant Bacteria." In: *JMAJ* 52.2 (2009).

[4] Kate Gould. "Antibiotics: from prehistory to the present day." In: *Journal of Antimicrobial Chemotherapy* 71.3 (Mar. 2016), pp. 572–575. ISSN: 0305-7453. DOI: 10.1093/JAC/DKV484. URL: https://academic.oup.com/jac/article/71/3/572/2364412.

[5] Matt Hutchings, Andrew Truman, and Barrie Wilkinson. "Antibiotics: past, present and future." In: *Current Opinion in Microbiology* 51 (Oct. 2019), pp. 72–80. ISSN: 18790364. DOI: 10.1016/J.MIB.2019.10.008. URL: https://doi.org/10.1016/j.mib.2019.10.008.

[6] Albert Schatz, Elizabeth Bugle, and Selman A. Waksman. "Streptomycin, a Substance Exhibiting Antibiotic Activity Against Gram-Positive and Gram-Negative Bacteria." In: *Proceedings of the Society for Experimental Biology and Medicine* 55.1 (Nov. 1944), pp. 66–69. ISSN: 15353699. DOI: 10.3181/00379727-55-14461. URL: https://journals.sagepub.com/doi/abs/10.3181/00379727-55-14461.

[7] Kim Lewis. "Platforms for antibiotic discovery." In: *Nature Reviews Drug Discovery* 12.5 (2013), pp. 371–387. ISSN: 14741776. DOI: 10.1038/nrd3975.

[8] Peter M. Wright, Ian B. Seiple, and Andrew G. Myers. "The evolving role of chemical synthesis in antibacterial drug discovery." In: *Angewandte Chemie - International Edition* 53.34 (2014), pp. 8840–8869. ISSN: 15213773. DOI: 10.1002/anie.201310843.

[9] Jim O'Neill and The Review on Antimicrobial Resistance (Chaired by Jim O'Neill). "Tackling Drug-Resistant Infections Globally: Final Report and Recommendations." In: *Review on antimicrobial resistance* May (2016), pp. 1–80.

[10] Christopher JL Murray, Kevin Shunji Ikuta, Fablina Sharara, Lucien Swetschinski, Gisela Robles Aguilar, Authia Gray, Chieh Han, Catherine Bisignano, Puja Rao, Eve Wool, et al. "Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis." In: *The Lancet* 399.10325 (2022), pp. 629–655.

[11] Marsha C Wibowo, Zhen Yang, Maxime Borry, Alexander Hübner, Kun D Huang, Braden T Tierney, Samuel Zimmerman, Francisco Barajas-Olmos, Cecilia Contreras-Cubas, Humberto García-Ortiz, et al. "Reconstruction of ancient microbial genomes from the human gut." In: *Nature* 594.7862 (2021), pp. 234–239.

[12] Vanessa M D'Costa, Christine E King, Lindsay Kalan, Mariya Morar, Wilson WL Sung, Carsten Schwarz, Duane Froese, Grant Zazula, Fabrice Calmels, Regis Debruyne, et al. "Antibiotic resistance is ancient." In: *Nature* 477.7365 (2011), pp. 457–461.

[13] Nicholas Waglechner and Gerard D Wright. "Antibiotic resistance: it's bad, but why isn't it worse?" In: *BMC biology* 15.1 (2017), pp. 1–8.

[14] J. Davies. "Origins and evolution of antibiotic resistance." In: *Microbiology and molecular biology reviews* 74.3 (2010), pp. 417–433. ISSN: 02134101. DOI: 10.1128/mmbr.00016-10.

[15] Michael N Alekshun and Stuart B Levy. "Molecular mechanisms of antibacterial multidrug resistance." In: *Cell* 128.6 (2007), pp. 1037–1050.

[16] Wanda C Reygaert. "An overview of the antimicrobial resistance mechanisms of bacteria." In: *AIMS microbiology* 4.3 (2018), p. 482.

[17] Alison H. Holmes, Luke S.P. Moore, Arnfinn Sundsfjord, Martin Steinbakk, Sadie Regmi, Abhilasha Karkey, Philippe J. Guerin, and Laura J.V. Piddock. "Understanding the mechanisms and drivers of antimicrobial resistance." In: *The Lancet* 387.10014 (2016), pp. 176–187. ISSN: 1474547X. DOI: 10.1016/S0140-6736(15)00473-0.

[18] Stefan Ebmeyer, Erik Kristiansson, and DG Larsson. "A framework for identifying the recent origins of mobile antibiotic resistance genes." In: *Communications biology* 4.1 (2021), pp. 1–10.

[19] PM Bennett. "Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria." In: *British journal of pharmacology* 153.S1 (2008), S347–S357.

[20] Alvaro San Millan. "Evolution of plasmid-mediated antibiotic resistance in the clinical context." In: *Trends in microbiology* 26.12 (2018), pp. 978–985.

[21] Sally R Partridge, Stephen M Kwong, Neville Firth, and Slade O Jensen. "Mobile genetic elements associated with antimicrobial resistance." In: *Clinical microbiology reviews* 31.4 (2018), e00088–17.

[22] Timothy M Ghaly and Michael R Gillings. “New perspectives on mobile genetic elements: a paradigm shift for managing the antibiotic resistance crisis.” In: *Philosophical Transactions of the Royal Society B* 377.1842 (2022), p. 20200462.

[23] Angela HAM Van Hoek, Dik Mevius, Beatriz Guerra, Peter Mullany, Adam Paul Roberts, and Henk JM Aarts. “Acquired antibiotic resistance genes: an overview.” In: *Frontiers in microbiology* 2 (2011), p. 203.

[24] Markus HK Johansson, Valeria Bortolaia, Supathep Tansirichaiya, Frank M Aarestrup, Adam P Roberts, and Thomas N Petersen. “Detection of mobile genetic elements associated with antibiotic resistance in Salmonella enterica using a newly developed web tool: MobileElementFinder.” In: *Journal of Antimicrobial Chemotherapy* 76.1 (2021), pp. 101–109.

[25] Johan Bengtsson-Palme, Erik Kristiansson, and DG Joakim Larsson. “Environmental factors influencing the development and spread of antibiotic resistance.” In: *FEMS microbiology reviews* 42.1 (2018), fux053.

[26] Rene S Hendriksen, Patrick Munk, Patrick Njage, Bram Van Bunnik, Luke McNally, Oksana Lukjancenko, Timo Röder, David Nieuwenhuijse, Susanne Karlsmose Pedersen, Jette Kjeldgaard, et al. “Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage.” In: *Nature communications* 10.1 (2019), pp. 1–12.

[27] Mei Zhuang, Yigal Achmon, Yuping Cao, Xiaomin Liang, Liang Chen, Hui Wang, Bupe A Siame, and Ka Yin Leung. “Distribution of antibiotic resistance genes in the environment.” In: *Environmental Pollution* 285 (2021), p. 117402.

[28] Richard Wise, Tony Hart, Otto Cars, Marc Streulens, Reinen Helmuth, Pentti Huovinen, and Marc Sprenger. *Antimicrobial resistance*. 1998.

[29] Ellen K Silbergeld, Jay Graham, Lance B Price, et al. “Industrial food animal production, antimicrobial resistance, and human health.” In: *Annual review of public health* 29.1 (2008), pp. 151–169.

[30] Frank Møller Aarestrup, Anne Mette Seyfarth, Hanne-Dorthe Emborg, Karl Pedersen, René S Hendriksen, and Flemming Bager. “Effect of abolishment of the use of antimicrobial agents for growth promotion on occurrence of antimicrobial resistance in fecal enterococci from food animals in Denmark.” In: *Antimicrobial Agents and chemotherapy* 45.7 (2001), pp. 2054–2059.

[31] Bonnie M Marshall and Stuart B Levy. “Food animals and antimicrobials: impacts on human health.” In: *Clinical microbiology reviews* 24.4 (2011), pp. 718–733.

[32] Vibe Dalhoff Andersen, Frank Møller Aarestrup, Patrick Munk, Marie Stengaard Jensen, LV de Knegt, Valeria Bortolaia, BE Knudsen, O Lukjancenko, AC Birkegård, and H Vigre. “Predicting effects of changed antimicrobial usage on the abundance of antimicrobial resistance genes in finisher’gut microbiomes.” In: *Preventive veterinary medicine* 174 (2020), p. 104853.

[33] Alfonso J. Alanis. "Resistance to antibiotics: Are we in the post-antibiotic era?" In: *Archives of Medical Research* 36.6 (2005), pp. 697–705. ISSN: 01884409. DOI: 10.1016/j.arcmed.2005.06.009.

[34] World Health Organization et al. "Critically important antimicrobials for human medicine." In: (2019).

[35] Peter C Collignon, John M Conly, Antoine Andremont, Scott A McEwen, Awa Aidara-Kane, Bogotá Meeting on Integrated Surveillance of Antimicrobial Resistance (WHO-AGISAR) World Health Organization Advisory Group, Yvonne Agerso, Antoine Andremont, Peter Collignon, John Conly, et al. "World Health Organization ranking of antimicrobials according to their importance in human medicine: a critical step for developing risk management strategies to control antimicrobial resistance from food animal production." In: *Clinical Infectious Diseases* 63.8 (2016), pp. 1087–1093.

[36] Mary Barber, Mary Rozwadowska-Dowzenko, et al. "Infection by Penicillin-Resistant Staphyloeoeci." In: *Lancet* (1948), pp. 641–4.

[37] Stuart B Levy and Bonnie Marshall. "Antibacterial resistance worldwide: causes, challenges and responses." In: *Nature medicine* 10.12 (2004), S122–S129.

[38] Mohammad H Gharaibeh and Shoroq Q Shatnawi. "An overview of colistin resistance, mobilized colistin resistance genes dissemination, global responses, and the alternatives to colistin: a review." In: *Veterinary World* 12.11 (2019), p. 1735.

[39] Nadheema Hammood Hussein, Israa AL-Kadmy, Butheina Mohammed Taha, and Jumaah Dakel Hussein. "Mobilized colistin resistance (mcr) genes from 1 to 10: a comprehensive review." In: *Molecular Biology Reports* 48.3 (2021), pp. 2897–2907.

[40] Tanise V Dalmolin, Daiana de Lima-Morales, and Afonso L Barth. "Plasmid-mediated colistin resistance: what do we know?" In: *Journal of Infectiology and Epidemiology* 1.2 (2018).

[41] Yi-Yun Liu, Yang Wang, Timothy R Walsh, Ling-Xian Yi, Rong Zhang, James Spencer, Yohei Doi, Guobao Tian, Baolei Dong, Xianhui Huang, et al. "Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study." In: *The Lancet infectious diseases* 16.2 (2016), pp. 161–168.

[42] Robert L Skov and Dominique L Monnet. "Plasmid-mediated colistin resistance (mcr-1 gene): three months later, the story unfolds." In: *Eurosurveillance* 21.9 (2016), p. 30155.

[43] Qixia Luo, Yuan Wang, and Yonghong Xiao. "Prevalence and transmission of mobilized colistin resistance (mcr) gene in bacteria common to animals and humans." In: *Biosafety and Health* 2.02 (2020), pp. 71–78.

[44] Chengcheng Wang, Yu Feng, Lina Liu, Li Wei, Mei Kang, and Zhiyong Zong. "Identification of novel mobile colistin resistance gene mcr-10." In: *Emerging microbes & infections* 9.1 (2020), pp. 508–516.

[45] Shalini Kunhikannan, Colleen J Thomas, Ashley E Franks, Sumana Mahadevaiah, Sumana Kumar, and Steve Petrovski. "Environmental hotspots for antibiotic resistance genes." In: *Microbiologyopen* 10.3 (2021), e1197.

[46] Saadia Andleeb, Muhsin Jamal, Sayed Bukhari, Sumbal Sardar, and Mahnoor Majid. "Trends in Antimicrobial Use in Food Animals, Aquaculture, and Hospital Waste." In: *Antibiotics and Antimicrobial Resistance Genes*. Springer, 2020, pp. 95–138.

[47] Timothy P Robinson, DP Bu, Juan Carrique-Mas, Eric M Fèvre, Marius Gilbert, Delia Grace, Simon I Hay, Jatesada Jiwakanon, Manish Kakkar, S Kariuki, et al. "Antibiotic resistance is the quintessential One Health issue." In: *Transactions of the Royal Society of Tropical Medicine and Hygiene* 110.7 (2016), pp. 377–380.

[48] Ea Zankari, Henrik Hasman, Salvatore Cosentino, Martin Vestergaard, Simon Rasmussen, Ole Lund, Frank M Aarestrup, and Mette Voldby Larsen. "Identification of acquired antimicrobial resistance genes." In: *Journal of antimicrobial chemotherapy* 67.11 (2012), pp. 2640–2644.

[49] M Kresken and B Wiedemann. "Development of resistance to nalidixic acid and the fluoroquinolones after the introduction of norfloxacin and ofloxacin." In: *Antimicrobial Agents and Chemotherapy* 32.8 (1988), pp. 1285–1288.

[50] Daniel Fernández-Villa, Maria Rosa Aguilar, and Luis Rojo. "Folic acid antagonists: antimicrobial and immunomodulating mechanisms and applications." In: *International journal of molecular sciences* 20.20 (2019), p. 4996.

[51] ONKA Kitamoto, N Kasai, K Fukaya, and A Kawashima. "Drug sensitivity of the Shigella strains isolated in 1955." In: *J. Jpn. Assoc. Infect. Dis* 30 (1956), pp. 403–404.

[52] D Dámaso, M Moreno-López, and J Martinez-Beltran. "Evolution of sensitivity to fosfomycin in bacteria isolated in 1973, 1974 and 1975 in the Servicio de Microbiologia y Epidemiologia of the 'Clinica Puerta de Hierro', Madrid." In: *Chemotherapy* 23.Suppl. 1 (1977), pp. 104–111.

[53] Katherine S Long, Jacob Poehlsgaard, Corinna Kehrenberg, Stefan Schwarz, and Birte Vester. "The Cfr rRNA methyltransferase confers resistance to phenicols, lincosamides, oxazolidinones, pleuromutilins, and streptogramin A antibiotics." In: *Antimicrobial agents and chemotherapy* 50.7 (2006), pp. 2500–2505.

[54] I Chopra. "Mechanisms of resistance to fusidic acid in Staphylococcus aureus." In: *Microbiology* 96.2 (1976), pp. 229–238.

[55] Frank Møller Aarestrup, Flemming Bager, NE Jensen, M Madsen, A Meyling, and Henrik Caspar Wegener. "Resistance to antimicrobial agents used for animal therapy in pathogenic-, zoonotic-and indicator bacteria isolated from different food animals in Denmark: a baseline study for the Danish Integrated Antimicrobial Resistance Monitoring Programme (DANMAP)." In: *Apmis* 106.7-12 (1998), pp. 745–770.

[56] Majda Attauabi, Birgitte Borck Høg, Berit Müller-Pebody, Ana Sofia Ribeiro Duarte, Helle Bisgaard Korsgaard, Jeppe Boel, Tine Dalby, Anette M Hammerum, Frank Hansen, Henrik Hasman, et al. "DANMAP 2020: Use of antimicrobial agents and occurrence of antimicrobial resistance in bacteria from food animals, food and humans in Denmark." In: (2021).

[57] World Health Organization et al. "Global antimicrobial resistance surveillance system (GLASS) report: early implementation 2020." In: (2020).

[58] World Health Organization et al. "Global antimicrobial resistance and use surveillance system (GLASS) report: 2021." In: (2021).

[59] World Health Organization et al. "Health in 2015: from MDGs, millennium development goals to SDGs, sustainable development goals." In: (2015).

[60] Derek R MacFadden, Sarah F McGough, David Fisman, Mauricio Santillana, and John S Brownstein. "Antibiotic resistance increases with local temperature." In: *Nature Climate Change* 8.6 (2018), pp. 510–514.

[61] Márió Gajdács, Edit Urbán, Anette Stájer, and Zoltán Baráth. "Antimicrobial resistance in the context of the sustainable development goals: A brief review." In: *European Journal of Investigation in Health, Psychology and Education* 11.1 (2021), pp. 71–82.

[62] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools." In: *Nucleic acids research* 41.D1 (2012), pp. D590–D596.

[63] P. R. Burton, J.M. Bowden, and M.D. Tobin. "Epidemiology and Genetic Epidemiology." In: *Handbook of Statistical Genetics*. Ed. by D.J. Balding, M. Bishop, and C. Cannings. 3rd. John Wiley & Sons, L td, 2007. Chap. 32, pp. 1111–1140. ISBN: 978-0-470-05830-5.

[64] Jennifer L Gardy and Nicholas J Loman. "Towards a genomics-informed, real-time, global pathogen surveillance system." In: *Nature Reviews Genetics* 19.1 (2018), pp. 9–20.

[65] Stephen J O'Brien. "A decade of GigaScience: A perspective on conservation genetics." In: *GigaScience* 11 (2022).

[66] Philip Hugenholtz, Brett M Goebel, and Norman R Pace. "Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity." In: *Journal of bacteriology* 180.18 (1998), pp. 4765–4774.

[67] Kyle J Popovich and Evan S Snitkin. “Whole genome sequencing—implications for infection prevention and outbreak investigations.” In: *Current infectious disease reports* 19.4 (2017), pp. 1–7.

[68] Jill E Clarridge III. “Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases.” In: *Clinical microbiology reviews* 17.4 (2004), pp. 840–862.

[69] Rob Knight, Alison Vrbanac, Bryn C Taylor, Alexander Aksenov, Chris Callewaert, Justine Debelius, Antonio Gonzalez, Tomasz Kosciolek, Laura-Isobel McCall, Daniel McDonald, et al. “Best practices for analysing microbiomes.” In: *Nature Reviews Microbiology* 16.7 (2018), pp. 410–422.

[70] James D Watson and Francis HC Crick. “Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid.” In: *Nature* 171.4356 (1953), pp. 737–738.

[71] Doris T Zallen. “Despite Franklin’s work, Wilkins earned his Nobel.” In: *Nature* 425.6953 (2003), pp. 15–15.

[72] Frederick Sanger, Steven Nicklen, and Alan R Coulson. “DNA sequencing with chain-terminating inhibitors.” In: *Proceedings of the national academy of sciences* 74.12 (1977), pp. 5463–5467.

[73] James M Heather and Benjamin Chain. “The sequence of sequencers: The history of sequencing DNA.” In: *Genomics* 107.1 (2016), pp. 1–8.

[74] Taishan Hu, Nilesh Chitnis, Dimitri Monos, and Anh Dinh. “Next-generation sequencing technologies: An overview.” In: *Human Immunology* 82.11 (2021), pp. 801–811.

[75] Tom Hunkapiller, RJ Kaiser, BF Koop, and Leroy Hood. “Large-scale and automated DNA sequence determination.” In: *Science* 254.5028 (1991), pp. 59–67.

[76] Shawn E Levy and Richard M Myers. “Advancements in next-generation sequencing.” In: *Annu Rev Genomics Hum Genet* 17.1 (2016), pp. 95–115.

[77] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. “The sequence of the human genome.” In: *science* 291.5507 (2001), pp. 1304–1351.

[78] Pål Nyrén and Arne Lundin. “Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis.” In: *Analytical biochemistry* 151.2 (1985), pp. 504–509.

[79] Jay Shendure, Shankar Balasubramanian, George M Church, Walter Gilbert, Jane Rogers, Jeffery A Schloss, and Robert H Waterston. “DNA sequencing at 40: past, present and future.” In: *Nature* 550.7676 (2017), pp. 345–353.

[80] Wilhelm J Ansorge. “Next-generation DNA sequencing techniques.” In: *New biotechnology* 25.4 (2009), pp. 195–203.

[81] Lincoln D Stein. “The case for cloud computing in genome informatics.” In: *Genome biology* 11.5 (2010), pp. 1–7.

[82] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, et al. “Real-time DNA sequencing from single polymerase molecules.” In: *Science* 323.5910 (2009), pp. 133–138.

[83] Kishore R Kumar, Mark J Cowley, and Ryan L Davis. “Next-generation sequencing and emerging technologies.” In: *Seminars in thrombosis and hemostasis*. Vol. 45. 07. Thieme Medical Publishers. 2019, pp. 661–673.

[84] Miten Jain, Hugh E Olsen, Benedict Paten, and Mark Akeson. “The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community.” In: *Genome biology* 17.1 (2016), pp. 1–11.

[85] Søren M Karst, Ryan M Ziels, Rasmus H Kirkegaard, Emil A Sørensen, Daniel McDonald, Qiyun Zhu, Rob Knight, and Mads Albertsen. “High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing.” In: *Nature methods* 18.2 (2021), pp. 165–169.

[86] Karl V Voelkerding, Shale A Dames, and Jacob D Durtschi. “Next-generation sequencing: from basic research to diagnostics.” In: *Clinical chemistry* 55.4 (2009), pp. 641–658.

[87] Elaine R Mardis. “Next-generation DNA sequencing methods.” In: *Annual review of genomics and human genetics* 9.1 (2008), pp. 387–402.

[88] Masanori Arita, Ilene Karsch-Mizrachi, and Guy Cochrane. “The international nucleotide sequence database collaboration.” In: *Nucleic Acids Research* 49.D1 (2021), pp. D121–D124.

[89] Jang-il Sohn and Jin-Wu Nam. “The present and future of de novo whole-genome assembly.” In: *Briefings in bioinformatics* 19.1 (2018), pp. 23–40.

[90] Nathan D Olson, Todd J Treangen, Christopher M Hill, Victoria Cepeda-Espinoza, Jay Ghurye, Sergey Koren, and Mihai Pop. “Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes.” In: *Briefings in bioinformatics* 20.4 (2019), pp. 1140–1150.

[91] Sergey Nurk, Dmitry Meleshko, Anton Korobeynikov, and Pavel A Pevzner. “metaSPAdes: a new versatile metagenomic assembler.” In: *Genome research* 27.5 (2017), pp. 824–834.

[92] Raffaella Rizzi, Stefano Beretta, Murray Patterson, Yuri Pirola, Marco Previtali, Gianluca Della Vedova, and Paola Bonizzoni. “Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era.” In: *Quantitative Biology* 7.4 (2019), pp. 278–292.

[93] Derrick E Wood and Steven L Salzberg. “Kraken: ultrafast metagenomic sequence classification using exact alignments.” In: *Genome biology* 15.3 (2014), pp. 1–12.

[94] Justin Kuczynski, Jesse Stombaugh, William Anton Walters, Antonio González, J Gregory Caporaso, and Rob Knight. "Using QIIME to analyze 16S rRNA gene sequences from microbial communities." In: *Current protocols in microbiology* 27.1 (2012), 1E–5.

[95] Francesco Asnicar, Andrew Maltez Thomas, Francesco Beghini, Claudia Mengoni, Serena Manara, Paolo Manghi, Qiyun Zhu, Mattia Bolzan, Fabio Cumbo, Uyen May, et al. "Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0." In: *Nature communications* 11.1 (2020), pp. 1–10.

[96] Ben Langmead and Steven L Salzberg. "Fast gapped-read alignment with Bowtie 2." In: *Nature methods* 9.4 (2012), pp. 357–359.

[97] Heng Li. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM." In: *arXiv preprint arXiv:1303.3997* (2013).

[98] Philip TLC Clausen, Frank M Aarestrup, and Ole Lund. "Rapid and precise alignment of raw reads against redundant databases with KMA." In: *BMC bioinformatics* 19.1 (2018), pp. 1–8.

[99] Brian P Alcock, Amogelang R Raphenya, Tammy TY Lau, Kara K Tsang, Mégane Bouchard, Arman Edalatmand, William Huynh, Anna-Lisa V Nguyen, Annie A Cheng, Sihan Liu, et al. "CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database." In: *Nucleic acids research* 48.D1 (2020), pp. D517–D525.

[100] M Luz Calle. "Statistical analysis of metagenomics data." In: *Genomics & informatics* 17.1 (2019).

[101] Andrew D Fernandes, Jennifer Ns Reid, Jean M Macklaim, Thomas A McMurrough, David R Edgell, and Gregory B Gloor. "Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis." In: *Microbiome* 2.1 (2014), pp. 1–13.

[102] Jonathan Friedman and Eric J Alm. "Inferring correlation networks from genomic survey data." In: (2012).

[103] Zachary D Kurtz, Christian L Müller, Emily R Miraldi, Dan R Littman, Martin J Blaser, and Richard A Bonneau. "Sparse and compositionally robust inference of microbial ecological networks." In: *PLoS computational biology* 11.5 (2015), e1004226.

[104] Gregory B Gloor, Jia Rong Wu, Vera Pawlowsky-Glahn, and Juan José Egozcue. "It's all relative: analyzing microbiome data as compositions." In: *Annals of epidemiology* 26.5 (2016), pp. 322–329.

[105] Michael Worobey, Joshua I Levy, Lorena Malpica Serrano, Alexander Crits-Christoph, Jonathan E Pekar, Stephen A Goldstein, Angela L Rasmussen, Moritz UG Kraemer, Chris Newman, Marion PG Koopmans, et al. "The Huanan Seafood Wholesale Market in Wuhan was the early epicenter of the COVID-19 pandemic." In: *Science* 377.6609 (2022), pp. 951–959.

[106] Renxin Zhao, Ke Yu, Jiayu Zhang, Guijuan Zhang, Jin Huang, Liping Ma, Chunfang Deng, Xiaoyan Li, and Bing Li. "Deciphering the mobility and bacterial hosts of antibiotic resistance genes under antibiotic selection pressure by metagenomic assembly and binning approaches." In: *Water Research* 186 (2020), p. 116318.

[107] Devin B Holman, Arun Kommadath, Jeffrey P Tingley, and D Wade Abbott. "Novel insights into the pig gut microbiome using metagenome-assembled genomes." In: *Microbiology spectrum* (2022), e02380–22.

[108] Shinichi Sunagawa, Silvia G Acinas, Peer Bork, Chris Bowler, Damien Eveillard, Gabriel Gorsky, Lionel Guidi, Daniele Iudicone, Eric Karsenti, Fabien Lombard, et al. "Tara Oceans: towards global ocean ecosystems biology." In: *Nature Reviews Microbiology* 18.8 (2020), pp. 428–445.

[109] Rafael RC Cuadrat, Maria Sorokina, Bruno G Andrade, Tobias Goris, and Alberto MR Davila. "Global ocean resistome revealed: Exploring antibiotic resistance gene abundance and distribution in TARA Oceans samples." In: *GigaScience* 9.5 (2020), giaa046.

[110] Fergus WJ Collins, Calum J Walsh, Beatriz Gomez-Sala, Elena Guijarro-García, David Stokes, Klara B Jakobsdóttir, Kristján Kristjánsson, Finlay Burns, Paul D Cotter, Mary C Rea, et al. "The microbiome of deep-sea fish reveals new microbial species and a sparsity of antibiotic resistance genes." In: *Gut microbes* 13.1 (2021), p. 1921924.

[111] Charles W Knapp, Seánín M McCluskey, Brajesh K Singh, Colin D Campbell, Gordon Hudson, and David W Graham. "Antibiotic resistance gene abundances correlate with metal and geochemical conditions in archived Scottish soils." In: *PloS one* 6.11 (2011), e27300.

[112] Cheng Wang, Ruiwen Hu, PJ Strong, Wei Zhuang, Weiming Huang, Zhiwen Luo, Qingyun Yan, Zhili He, and Longfei Shu. "Prevalence of antibiotic resistance genes and bacterial pathogens along the soil–mangrove root continuum." In: *Journal of Hazardous Materials* 408 (2021), p. 124985.

[113] Frank M Aarestrup and Mark EJ Woolhouse. "Using sewage for surveillance of antimicrobial resistance." In: *Science* 367.6478 (2020), pp. 630–632.

[114] Jianhua Guo, Jie Li, Hui Chen, Philip L Bond, and Zhiguo Yuan. "Metagenomic analysis reveals wastewater treatment plants as hotspots of antibiotic resistance genes and mobile genetic elements." In: *Water research* 123 (2017), pp. 468–478.

[115] Charmaine Ng, Martin Tay, Boonfei Tan, Thai-Hoang Le, Laurence Haller, Hongjie Chen, Tse H Koh, Timothy MS Barkham, Janelle R Thompson, and Karina Y-H Gin. "Characterization of metagenomes in urban aquatic compartments reveals high prevalence of clinically relevant antibiotic resistance genes in wastewaters." In: *Frontiers in microbiology* 8 (2017), p. 2200.

[116] Shahjahon Begmatov, Alexander G Dorofeev, Vitaly V Kadnikov, Alexey V Beletsky, Nikolai V Pimenov, Nikolai V Ravin, and Andrey V Mardanov. "The structure of microbial communities of activated sludge of large-scale wastewater treatment plants in the city of Moscow." In: *Scientific reports* 12.1 (2022), pp. 1–14.

[117] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. "The FAIR Guiding Principles for scientific data management and stewardship." In: *Scientific data* 3.1 (2016), pp. 1–9.

[118] Randall J LeVeque, Ian M Mitchell, and Victoria Stodden. "Reproducible research for scientific computing: Tools and strategies for changing the culture." In: *Computing in Science & Engineering* 14.04 (2012), pp. 13–17.

[119] Johannes Köster and Sven Rahmann. "Snakemake—a scalable bioinformatics workflow engine." In: *Bioinformatics* 28.19 (2012), pp. 2520–2522.

[120] Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B Hall, Christopher H Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O Twardziok, Alexander Kanitz, et al. "Sustainable data analysis with Snakemake." In: *F1000Research* 10 (2021).

[121] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. "Nextflow enables reproducible computational workflows." In: *Nature biotechnology* 35.4 (2017), pp. 316–319.

[122] Jennifer Lu, Natalia Rincon, Derrick E Wood, Florian P Breitwieser, Christopher Pockrandt, Ben Langmead, Steven L Salzberg, and Martin Steinegger. "Metagenome analysis using the Kraken software suite." In: *Nature protocols* (2022), pp. 1–25. ISSN: 1750-2799. DOI: 10.1038/s41596-022-00738-y. URL: http://www.ncbi.nlm.nih.gov/pubmed/36171387.

[123] European Organization For Nuclear Research and OpenAIRE. *Zenodo*. en. 2013. DOI: 10.25495/7GXK-RD71. URL: https://www.zenodo.org/.

[124] Thomas C Redman. "If your data is bad, your machine learning tools are useless." In: *Harvard Business Review* 2 (2018).

[125] Richard J Abdill, Elizabeth M Adamowicz, and Ran Blekhman. "Public human microbiome data are dominated by highly developed countries." In: *PLoS biology* 20.2 (2022), e3001536.

[126] Gregory B. Gloor, Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. "Microbiome datasets are compositional: And this is not optional." In: *Frontiers in Microbiology* 8.NOV (2017), pp. 1–6. ISSN: 1664302X. DOI: 10.3389/fmicb.2017.02224.

[127] Thomas P. Quinn, Ionas Erb, Mark F. Richardson, and Tamsyn M. Crowley. "Understanding sequencing data as compositions: An outlook and review." In: *Bioinformatics* 34.16 (2018), pp. 2870–2878. ISSN: 14602059. DOI: 10.1093/bioinformatics/bty175.

[128] Karl Pearson. "Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs." In: *Proceedings of the royal society of london* 60.359-367 (1897), pp. 489–498.

[129] J Aitchison. "The Statistical Analysis of Compositional Data." In: *Journal of the Royal Statistical Society: Series B (Methodological)* 44.2 (1982), pp. 139–160.

[130] Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. *Modeling and Analysis of Compositional Data*. John Wiley and Sons, 2015, pp. 1–247. ISBN: 9781119003144. DOI: 10.1002/9781119003144.

[131] Josep-Antoni Martín-Fernández, Karel Hron, Matthias Templ, Peter Filzmoser, and Javier Palarea-Albaladejo. "Bayesian-multiplicative treatment of count zeros in compositional data sets." In: *Statistical Modelling* 15.2 (2015), pp. 134–158.

[132] K Gerald Van den Boogaart and Raimon Tolosana-Delgado. "Zeroes, Missings, and Outliers." In: *Analyzing compositional data with R*. Vol. 122. Springer, 2013. Chap. 7, pp. 209–252.

[133] Ian Holmes, Keith Harris, and Christopher Quince. "Dirichlet multinomial mixtures: generative models for microbial metagenomics." In: *PloS one* 7.2 (2012), e30126.

[134] Jonathan Thorsen, Asker Brejnrod, Martin Mortensen, Morten A Rasmussen, Jakob Stokholm, Waleed Abu Al-Soud, Søren Sørensen, Hans Bisgaard, and Johannes Waage. "Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies." In: *Microbiome* 4.1 (2016), pp. 1–14.

[135] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." In: *bioinformatics* 26.1 (2010), pp. 139–140.

[136] Joseph N Paulson, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. "Differential abundance analysis for microbial marker-gene surveys." In: *Nature methods* 10.12 (2013), pp. 1200–1202.

[137] Siddhartha Mandal, Will Van Treuren, Richard A White, Merete Eggesbø, Rob Knight, and Shyamal D Peddada. "Analysis of composition of microbiomes: a novel method for studying microbial composition." In: *Microbial ecology in health and disease* 26.1 (2015), p. 27663.

[138] Jacob T Nearing, Gavin M Douglas, Molly G Hayes, Jocelyn MacDonald, Dhwani K Desai, Nicole Allward, Casey Jones, Robyn J Wright, Akhilesh S Dhanani, André M Comeau, et al. "Microbiome differential abundance methods produce different results across 38 datasets." In: *Nature communications* 13.1 (2022), pp. 1–16.

[139] Abhishek Kaul, Siddhartha Mandal, Ori Davidov, and Shyamal D Peddada. "Analysis of microbiome data in the presence of excess zeros." In: *Frontiers in microbiology* 8 (2017), p. 2114.

[140] Mehdi Layeghifard, David M Hwang, and David S Guttman. "Disentangling interactions in the microbiome: a network perspective." In: *Trends in microbiology* 25.3 (2017), pp. 217–228.

[141] Sophie Weiss, Will Van Treuren, Catherine Lozupone, Karoline Faust, Jonathan Friedman, Ye Deng, Li Charlie Xia, Zhenjiang Zech Xu, Luke Ursell, Eric J Alm, et al. "Correlation detection strategies in microbial data sets vary widely in sensitivity and precision." In: *The ISME journal* 10.7 (2016), pp. 1669–1681.

[142] Antoni Susin, Yiwen Wang, Kim-Anh Lê Cao, and M Luz Calle. "Variable selection in microbiome compositional data analysis." In: *NAR Genomics and Bioinformatics* 2.2 (2020), lqaa029.

[143] Javier Rivera-Pinto, Juan Jose Egozcue, Vera Pawlowsky-Glahn, Raul Paredes, Marc Noguera-Julian, and M Luz Calle. "Balances: a new perspective for microbiome analysis." In: *mSystems* 3.4 (2018), e00053–18.

$\sqrt{17}$
$\Omega \int \delta e^{i\pi} = -1$
{2.7182818284}
ϑφέστυθιοπσδφγηξκλ
$\chi^2$
$\sum$