



PhD-FSTM-2022-035
The Faculty of Sciences, Technology and Medicine

DISSERTATION

Defence held on 30/03/2022 in Esch-sur-Alzette

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN BIOLOGIE

by

Laura DE NIES

Born on 27 May 1994 in Zwijndrecht, (the Netherlands)

MICROBIOME RESERVOIRS OF ANTIMICROBIAL
RESISTANCE

Dissertation defence committee

Dr Paul Wilmes, dissertation supervisor
Professor, Université du Luxembourg

Dr Anupam Sengupta, Chairman
Assistant professor, Université du Luxembourg

Dr Patrick May, Vice Chairman
Université du Luxembourg

Dr Willem van Schaik
Professor, University of Birmingham, United Kingdom

Dr Gabriele Berg,
Professor, Technical University of Graz, Austria

Declaration

I hereby declare that this dissertation has been written only by the undersigned and without any assistance from third parties. Furthermore, I confirm that no sources have been used in the preparation of this thesis other than those indicated herein.

Laura de Nies,
Esch-sur-Alzette, Luxembourg
April 30, 2022

Acknowledgements

I would like to thank my supervisor Prof. Paul Wilmes, for giving me the opportunity to join his group, and for his continued support of my Ph.D. studies and the related research. Thank you for your support, encouragement, and guidance during these years.

I further would like to thank Dr. Patrick May and Dr. Judith Hübschen for their valuable feedback and ideas during the CET meetings.

Furthermore, I want to thank Prof. Anupam Sengupta for joining my dissertation committee and agreeing to be chairperson for the defense. My gratitude goes to Prof. Gabriele Berg and Prof. Willem van Schaik for agreeing to be on my dissertation committee. Thank you for taking the time to attend my defense.

I want to express my gratitude to all the collaborators involved in various parts of this project: Dr. Elisabeth Lettelier, Mina Tsenkova, Sara Lopes, Dr. Anna Heintz-Buschart and Prof. Tom Battin. I would further like to thank all the members of the Microbiomes in One Health (MICROH) Doctoral Training Unit (DTU), which has supported this PhD. With special thanks to fellow Ph.D. students and members of the self-proclaimed “AMR Hustle” Rebecca Czolk and Maureen Feucherolles.

My thanks go to Dr. Cedric Lazny, Dr. Rashi Halder, Dr. Benoit Kunath, Dr. Linda Wampach, Catherine Sedrani, Dr. Charlotte de Rudder, Dr. Camille Martin Gallausiaux, Audrey Frachet Bour, Dr. Susana Martinez, and Janine Habier, for their help and support during my Ph.D. To the whole Systems Ecology group, both former and current members – thank you for a great and unforgettable time! A very special thanks to my office mates, aka co-conspirators, aka partners-in-crime, Dr. Susheel Bhanu Busi and Dr. Valentina Galata for their help, encouragement, and support.

I’m grateful to my fellow Ph.D. students of the University of Luxembourg, and to my friends both in and out of Luxembourg, for all the fun we have had over the last few years.

Last but not least, I would like to thank my family for their constant encouragement and for always believing in me.

Abstract

Antimicrobial resistance (AMR) presents a global threat to public health due to the inability to comprehensively treat bacterial infections. Emerging resistant bacteria residing within human, animal and environmental reservoirs may spread from one to the other, at both local and global levels. Consequently, AMR has the potential to rapidly become pandemic whereby it is no longer constrained by either geographical or human-animal borders. Therefore, to enhance our understanding on the dissemination of AMR we systematically resolved different reservoirs of antimicrobial resistance, leveraging animal, environmental and human samples, to provide a One Health perspective.

To identify antimicrobial resistance genes (ARGs) and compare their identity and prevalence across different microbial reservoirs, we developed the PathoFact pipeline which also contextualizes ARG localization on mobile genetic elements (MGEs). This methodology was applied to several metagenomic datasets covering microbiomes of infants, laboratory mice, a wastewater treatment plant (WWTP) and biofilms from glacier-fed streams (GFS). Investigating the infant gut resistome we found that the abundance of ARGs against (semi)-synthetic agents were increased in infants born via caesarian section compared to those born via vaginal delivery. Additionally, we identified mobile genetic elements (MGEs) encoding ARGs such as glycopeptide, diaminopyrimidine and multidrug resistance at an early age. MGEs are often pivotal in the accumulation and dissemination of AMR within a microbial population. Therefore, we assessed the effect of selective pressure on the evolution and consecutive dissemination of AMR within the commensal gut microbiome, utilizing a mouse model. While plasmids and phages were found to contribute to the spread of AMR, we found that integrons represented the primary factors mediating AMR in the antibiotic-treated mice.

In addition to the above-described studies, we investigated the environmental resistome, comprising both the urban environment, i.e., the WWTP, and a natural environment, GFS biofilms. Utilizing a multi-omics approach we investigated the WWTP resistome over a 1.5 years timeseries and found that a core group of fifteen AMR categories were always present. Additionally, we found a significant difference in AMR categories encoded on phages versus plasmids indicating that the MGEs contributed differentially to the dissemination of AMR. On the other hand, the GFS biofilms represent pristine environments with limited anthropogenic influences. Therein, we found that eukaryotes, as well as prokaryotes, may serve as AMR reservoirs owing to their potential for encoding ARGs. In addition to our identification of

biosynthetic gene clusters encoding antibacterial secondary metabolites, our findings highlight the constant intra- and inter-domain competition and the underlying mechanisms influencing microbial survival in GFS epilithic biofilms.

In general, we observed that the overall AMR abundances were highest in human and animal microbial reservoirs whilst environmental reservoirs demonstrated a higher diversity of ARG subtypes. Additionally, we identified human-associated, MGE-derived ARGs in all three components of the One Health triad, indicating possible transmission routes for AMR dissemination. In summary, this work provides a comprehensive assessment of the prevalence of antimicrobial resistance and its dissemination mechanisms in human, animal, and environmental mechanisms.

Scientific output

Major parts of this thesis are based upon work that has either been published or is in preparation for submission with the candidate as first author. In addition, the candidate has co-authored several publications of which minor parts are incorporated in the thesis. The full list of scientific outputs is listed below, and the original manuscripts are provided in **Appendix A**.

First author papers in peer-review journals

- **Laura de Nies**, Sara Lopes, Susheel Bhanu Busi, Valentina Galata, Anna Heintz-Buschart, Cedric Christian Laczny, Patrick May and Paul Wilmes (2021). PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. *Microbiome* 9 (1), 1-14, **published [Appendix A.2]**
- Susheel Bhanu Busi*, **Laura de Nies***, Janine Habier, Linda Wampach, Joelle V Fritz, Anna Heintz-Buschart, Patrick May, Rashi Halder, Carine de Beaufort and Paul Wilmes (2021). Persistence of birth mode-dependent effects on gut microbiome composition, immune system stimulation and antimicrobial resistance during the first year of life. *ISME Communications* 1 (1), 1-12, **published [Appendix A.3]**
- **Laura de Nies**, Susheel Bhanu Busi and Paul Wilmes (2021). Reservoirs of antimicrobial resistance in the context of One Health. *Current Microbiology*, **in review [Appendix A.1]**
- **Laura de Nies***, Susheel Bhanu Busi*, Mina Tsenkova, Elisabeth Lettelier and Paul Wilmes (2021). Evolution of the gut resistome following a selective antibiotic sweep. *Nature Communications*, **accepted [Chapter 4]**
- **Laura de Nies**, Susheel Bhanu Busi, Benoit Josef Kunath, Patrick May and Paul Wilmes (2021). Mobilome-driven segregation of the resistome in biological wastewater treatment. *eLife*, **in review [Appendix A.4]**
- Susheel Bhanu Busi*, **Laura de Nies***, Paraskevi Pramateftaki, Massimo Bourquin, Leïla Ezzat, Tyler J. Kohler, Stilianos Fodelianakis, Grégoire Michoud, Hannes Peter, Michail Styllas, Matteo Tolosano, Vincent De Staercke, Martina Schön, Valentina Galata, Tom Battin and Paul Wilmes. (2021) Glacier-fed stream biofilms harbor diverse resistomes and biosynthetic gene clusters. *Microbiome*, **in review [Appendix A.5]**

Co-author papers in peer-review journals

- Susana Martinez Arbas, Susheel Bhanu Busi, Pedro Queiros, **Laura de Nies**, Malte Herold, Patrick May, Paul Wilmes, Emilie EL Muller and Shaman Narayanasamy (2021). Challenges, Strategies, and Perspectives for Reference-Independent Longitudinal Multi-Omic Microbiome Studies. *Frontiers in Genetics* 12, 858, **published [Appendix A.7]**
- Valentina Galata*, Susheel Bhanu Busi*, Benoit Josef Kunath, **Laura de Nies**, Magdalena Calusinska, Rashi Halder, Patrick May, Paul Wilmes and Cedric Christian Laczny (2021). Functional meta-omics provide critical insights into long-and short read assemblies. *Briefings in bioinformatics* 22 (6) bbab330, **published [Appendix A.6]**
- Susheel Bhanu Busi, Massimo Bourquin, Stilianos Fodelianakis, Gregoire Michoud, Tyler J Kohler, Hannes Peter, Paraskevi Pramateftaki, Michail Styllas, Matteo Tolosano, Vincent De Staercke, Martina Schon, **Laura de Nies**, Ramona Marasco, Daniele Daffonchio, Leila Ezzat, Paul Wilmes and Tom J Battin. (2021) Genomic and metabolic adaptations of biofilms to ecological windows of opportunities in glacier-fed streams. *Nature Communications*, **accepted [Appendix A.8]**

Manuscripts in preparation

- **Laura de Nies**, Susheel Bhanu Busi, Benoit Josef Kunath, Oskar Hickl, Patrick May and Paul Wilmes. (2022) Leveraging metagenomics for a global One Health understanding of antimicrobial resistance.

Oral presentations in scientific conferences, symposia and workshops

- PathoFact: Predicting Virulence Factors and Antimicrobial Resistance in Human Case-Control Studies (2021). *8th International Human Microbiome Consortium Congress*. Barcelona, Spain (Virtual)
- Generation-scale evolution of the gut resistome under selective antibiotic pressure (2021). *4th Luxembourg Microbiology Day*. Belval, Luxembourg
- A metagenomic perspective of antimicrobial resistance in a One Health context (2021). *Life Sciences PhD Days*. Belval, Luxembourg (virtual)

Poster presentations in scientific conferences, symposia and workshops

- Microbial reservoirs of antimicrobial resistance (2019). *Metagenomics Bioinformatics EMBL course*. Hinxton, UK.
- Microbial reservoirs of Antimicrobial Resistance: A One Health Perspective (2019). *3rd Luxembourg Microbiology Day*. Luxembourg, Luxembourg.
- The effect of birth mode on gut microbiome composition, immune system stimulation and antimicrobial resistance during the first year of life (2019). *Life Science PhD Days*. Belval, Luxembourg.
- PathoFact: A pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data (2020). *ISME virtual summit*, virtual

Awards and recognitions

- Best poster: Microbial reservoirs of antimicrobial resistance (2019). *Metagenomics Bioinformatics EMBL course*. Hinxton, UK.
- Best talk (PhD category): Generation-scale evolution of the gut resistome under selective antibiotic pressure (2021). *4th Luxembourg Microbiology Day*. Belval, Luxembourg

Table of Contents

ACKNOWLEDGEMENTS	i
ABSTRACT	ii
SCIENTIFIC OUTPUT	iv
TABLE OF CONTENTS	vii
LIST OF FIGURES	xi
LIST OF TABLES	xiii
 CHAPTER 1. INTRODUCTION	 1
1.1 Mechanisms of antimicrobial resistance	2
1.2 Dissemination of antimicrobial resistance through horizontal gene transfer.....	4
1.3 Methods for detecting antimicrobial resistance.....	5
1.4 Microbial reservoirs of antimicrobial resistance	7
1.4.1 Human.....	7
1.4.2 Livestock, poultry and other animals.....	8
1.4.3 Environment.....	10
1.5 Metagenomic approaches in assessing antimicrobial resistance: a One Health perspective	13
 CHAPTER 2. PATHOFACT: A PIPELINE FOR THE PREDICTION OF VIRULENCE FACTORS AND ANTIMICROBIAL RESISTANCE GENES IN METAGENOMIC DATA.	 17
2.1 Background.....	18
2.2 Methods	20
2.2.1 PathoFact architecture	20
2.2.2 Workflow for the prediction of virulence factors	21
2.2.3 Workflow for the prediction of toxin genes	23
2.2.4 Workflow for the prediction of antimicrobial resistance genes	24
2.2.5 MGEs: plasmids and phages	24
2.2.6 Evaluation of the PathoFact pipeline.....	25
2.2.7 Data analysis and data availability of publicly available datasets	26
2.2.8 Data analysis and data availability of a simulated dataset.....	26
2.3 Results and Discussion	27
2.3.1 Benchmarking	27

2.3.2 Validation of the PathoFact pipeline.....	27
2.3.3 Performance evaluation using a simulated dataset	28
2.3.4 Performance of PathoFact on metagenomic datasets	29
2.4 Conclusion and outlook	35

CHAPTER 3. PERSISTENCE OF BIRTH MODE-DEPENDENT EFFECTS ON GUT MICROBIOME COMPOSITION AND ANTIMICROBIAL RESISTANCE DURING THE FIRST YEAR OF LIFE. 37

3.1 Background.....	38
3.2 Methods	40
3.2.1 Ethics statement.....	40
3.2.2 Sample collection	40
3.2.3 Faecal processing and nucleic acid extraction.....	41
3.2.4 DNA sequencing	41
3.2.5 Data processing for metagenomics, including genome reconstruction.....	41
3.2.6 Metagenomic taxonomic classification, virome and functional analyses	42
3.2.7 Identification of antimicrobial resistance genes and association with mobile genetic elements	42
3.2.8 LPS isolation and <i>in vitro</i> immunostimulation for cytokine profiling.....	44
3.2.9 Data analysis.....	44
3.3 Results.....	45
3.3.1 Birth mode-dependent gut microbiota differences during the first year.....	45
3.3.2 Assessment of differences in metagenomic functional potential at one year of age	47
3.3.3 Pro-inflammatory immune responses elevated in CSD after one year of life.....	49
3.3.4 Antimicrobial resistance modulated by birth mode.....	51
3.3.5 Taxa associated with antimicrobial resistance	53
3.3.6 Role of mobile genetic elements in antimicrobial resistance.....	54
3.3.7 Distribution of AMR categories encoded by mobile genetic elements	56
3.3.8 Phage-mediated horizontal gene transfer (HGT)	56
3.4 Discussion	57

CHAPTER 4. EVOLUTION OF THE GUT RESISTOME FOLLOWING A SELECTIVE ANTIBIOTIC SWEEP 62

4.1 Background.....	63
---------------------	----

4.2 Methods	64
4.2.1 Power calculation and sample size estimate	64
4.2.2 Mice model and antibiotic exposure	64
4.2.3 Faecal processing and nucleic acid extraction	65
4.2.4 DNA sequencing	65
4.2.5 Data processing for metagenomics, including genome reconstruction	66
4.2.5 Identification of antimicrobial resistance genes and association with mobile genetic elements	66
4.2.6 Linking ARGs with integrons	67
4.2.7 Data analysis	67
4.3 Results	68
4.3.1 Selection of specific taxa due to antibiotic-mediated depletion of the gut microbiome	68
4.3.2 Resistome after antibiotic treatment	69
4.3.3 Antibiotic-induced changes in taxonomic composition	71
4.3.4 MGEs linked to AMR dissemination	71
4.3.5 Integrons mediate AMR in antibiotic-treated mice	74
4.4 Discussion	76

CHAPTER 5. MOBILOME-MEDIATED SEGREGATION OF ANTIMICROBIAL RESISTANCE IN A WASTEWATER TREATMENT PLANT

5.1 Introduction	81
5.2 Methods	83
5.2.1 Sampling and biomolecular extraction	83
5.2.2 Sequencing and data processing for metagenomics and metatranscriptomics ..	83
5.2.3 Identification of antimicrobial resistance genes and association with mobile genetic elements	83
5.2.4 Metaproteomic sequencing and data analyses	84
5.2.5 Multi-omic integration	85
5.2.6 MGE partition assessment	85
5.2.7 Data analysis	85
5.3 Results	86
5.3.1 Longitudinal assessment of the resistome within a WWTP	86
5.3.2 Microbial populations and co-occurrence patterns of AMR	88

5.3.3 Monitoring pathogenic microorganisms within BWWTPs.....	91
5.3.4 Differential transmission of antimicrobial resistance via mobile genetic elements	92
5.3.5 Taxonomic affiliations of MGE-derived resistance genes	95
5.3.6 Metaproteomic validation of AMR abundance and expression	97
5.4 Discussion	98

CHAPTER 6. DIVERSITY OF THE RESISTOME AND BIOSYNTHETIC GENE CLUSTERS IN GLACIER-FED STREAM BIOFILMS 103

6.1 Introduction	104
6.2 Methods	106
6.2.1 Sampling and biomolecular extraction	106
6.2.2 Sequencing and data processing for metagenomics	106
6.2.3 Identification of antimicrobial resistance genes, antibiotics biosynthesis pathways and BGCs.....	106
6.2.4 Data analysis.....	107
6.3 Results.....	107
6.3.1 Antimicrobial resistance in a pristine environment.....	107
6.3.2 Antibiotic biosynthesis pathways and biosynthetic gene clusters	110
6.4 Discussion	112
6.5 Conclusions	115

CHAPTER 7. GENERAL CONCLUSION AND FUTURE PERSPECTIVES..... 116

7.1 General overview.....	117
7.2 AMR within and across biomes	118
7.3 Dissemination of MGE-derived AMR	122
7.4 The risk of antimicrobial resistance	123
7.5 Further perspectives on understanding One Health reservoirs	125

BIBLIOGRAPHY 127

APPENDIX A. MANUSCRIPTS

APPENDIX B. SUPPLEMENTARY FIGURES

APPENDIX C. SUPPLEMENTARY TABLES

List of Figures

Figure 1.1	AMR dissemination in One Health	13
Figure 2.1	Framework of the PathoFact pipeline	21
Figure 2.2	Classification framework for the prediction of virulence factors	22
Figure 2.3	Performance evaluation of PathoFact on a high-complexity simulated dataset	29
Figure 2.4	Virulence factors in three case-control metagenomic datasets.	31
Figure 2.5	Bacterial toxins in three case-control metagenomic datasets.....	32
Figure 2.6	Antimicrobial resistance in three case-control metagenomic datasets	33
Figure 2.7	Identification of MGEs within three case-control metagenomic datasets.....	35
Figure 3.1	Workflow representation of DNA and LPS isolation from faecal samples for metagenomic, immune and functional AMR analyses	45
Figure 3.2	Gut microbiome profiles throughout the first year of life.	47
Figure 3.3	Functional differences at 1 year of age	48
Figure 3.4	Immunostimulatory potential at 1 year of age	50
Figure 3.5	Antimicrobial resistance gene abundances over time.....	52
Figure 3.6	Taxa associated with antimicrobial resistance.....	53
Figure 3.7	Mobile genetic elements associated with antimicrobial resistance.	55
Figure 3.8	Horizontal gene transfer (HGT) events.	57
Figure 3.9	Summary figure.....	58
Figure 4.1	Representative illustration of the experimental design.	65
Figure 4.2	Metagenome-assembled genome profiles.....	69
Figure 4.3	Resistome in antibiotic-treated mice	70
Figure 4.4	AMR-associated taxonomy.	72
Figure 4.5	Abundance levels of AMR categories associated with MGEs	73
Figure 4.6	Comparison of Akkermansia muciniphila genomes.	74
Figure 4.7	AMR-mediated via integrons in mice administered with antibiotics.	75
Figure 5.1	Longitudinal metagenomic assessment of AMR.....	86
Figure 5.2	Longitudinal metatranscriptomic assessment of AMR	87
Figure 5.3	Microbial population-linked AMR	89
Figure 5.4	Assessment of AMR associated with clinical pathogens.	91
Figure 4.5	MGE-derived AMR within the BWWTP resistome	93
Figure 5.6	Taxonomic affiliations of MGE-derived resistance genes	95

Figure 5.7	Integrative multi-omic assessment of AMR.....	97
Figure 5.8	Separation of MGE-derived AMR within the BWWTP	99
Figure 6.1	Epilithic biofilms in GFSs harbors a diverse resistome.....	107
Figure 6.2	Taxonomic affiliation of AMR in GFSs epilithic biofilms.....	108
Figure 6.3	Biosynthetic gene clusters indicate the resistome potential.....	110
Figure 6.4	Association of BGCs with AMR.....	111
Figure 7.1	Comparison of the human, animal, and environmental resistome.....	118
Figure 7.2	ARG diversity in different microbial reservoirs.....	120
Figure 7.3	Association of MGEs with AMR in different microbial reservoirs.....	122
Figure 7.4	Risk of AMR	123

List of Tables

Table 2.1:	PathoFact runtimes with different threads/computational resources.....	27
Table 2.2:	Validation of the PathoFact pipeline.....	28
Table 5.1:	WHO priority list of antibiotic resistant bacteria.....	90

Chapter 1. Introduction

Parts of this chapter are based on the following publication submitted for peer-review:

Laura de Nies, Susheel Bhanu Busi, Paul Wilmes (2021). Reservoirs of antimicrobial resistance in the context of One Health

Current Microbiology in review [**Appendix A.1**]

Throughout history, bacterial infections have been a major cause of human disease and mortality. The discovery, subsequent development, and medical use of antibiotics brought an end to this pre-antibiotic era by providing effective treatment against bacterial infections. However, the use of antibiotics has gone hand-in-hand with the emergence and spread of antimicrobial resistance (AMR). Although antibiotic resistance in itself is a prehistoric phenomenon [1], the over- and mis-use of antibiotics has led to a global and immense increase in AMR over the past decades. As a result, many bacteria have now acquired resistance against multiple antibiotics which has led to the emergence of multi-resistant microbes, i.e., “superbugs” [2]. This phenomenon, for instance, has led to an overgrowth of pathobionts, encoding antimicrobial resistance genes (ARGs), causing alterations to the microbiome both in chronic diseases as well as in infections [3,4]. Consequently, this threatens human health through the spread of multidrug-resistant bacteria with an estimated number of deaths which may exceed ten million annually by 2050 [5,6].

1.1 Mechanisms of antimicrobial resistance

On the one hand, antimicrobial agents for fighting bacterial infections can be characterized depending on the mechanisms of their activity, i.e. agents that i) inhibit cell wall synthesis, ii) depolarize the cell membrane, iii) inhibit protein synthesis, iv) inhibit nucleic acid synthesis, or v) inhibit metabolic pathways [7]. On the other hand, various counteractive mechanisms have evolved to confer resistance. These can be characterized into categories such as those I) limiting the uptake of and exposure to antibiotics, II) modifying antibiotic targets through for example mutations, III) directly inactivating antibiotic molecules, or IV) ensuring their immediate export through active efflux pumps [7]. In this context, limitations to the uptake of antibiotics are mostly classified as intrinsic resistance. Acquired resistance in turn mostly utilizes the modification of antibiotic targets, while the inactivation or efflux of antibiotics are both intrinsic and acquired resistance mechanisms [7].

Bacteria have a natural ability which limits the uptake of antimicrobial agents. Specifically, in Gram negative bacteria the structure and functions of the LPS provide an immediate barrier to antibiotics, thereby conferring an innate resistance [8]. Gram positive bacteria, on the other hand, lack LPS and resort to mechanisms such as enzymatic degradation of antibiotics or reducing the affinity and susceptibility of antibiotic target sites [9]. Additionally, other mechanisms to limit uptake of antibiotics may involve a decrease in the number of porin

channels or mutations in the corresponding genes as well as the formation of protective biofilms [10,11].

Specific examples of antimicrobial target modifications include alterations in the structure of antibiotic binding proteins or mutations therein to prevent antibiotics from binding to those proteins[10,11]. Additionally, modifications of DNA gyrase and topoisomerase IV interfere with the antibiotics targeting the nucleic acid synthesis machinery [12], while further mutations in enzymes generate resistance to antibiotics inhibiting metabolic pathways. Alternatively, upregulation in the expression of these enzymes confers resistance through competitive inhibition [13]. Besides these mechanisms, inactivation of the antimicrobial drugs itself can occur, conferring resistance either through actual degradation or through the transfer of a chemical group to the drug. Lastly, bacteria possess various types of efflux pumps, such as the ABC, MATE, SMR, MFS and RND transporter families, which enable resistance via efflux of antimicrobial drugs [8].

With respect to these mechanisms, bacterial resistance can be classified as either natural or acquired resistance [14]. Natural resistance can be further subdivided into either 'intrinsic', which is constantly expressed in a bacterial species, or 'induced', in which resistance genes are only expressed upon exposure to antibiotics [12]. Acquired resistance can be defined as the acquisition of resistance-conferring genetic material through horizontal gene transfer (HGT), e.g., conjugation or transduction and alternatively via mutations in the chromosomal DNA after antibiotic exposure [15]. In most cases, ARGs are associated with conjugation events which are the likely mechanisms for the dissemination of AMR compared to transduction [16]. Interestingly, the rate of transfer of ARGs via the individual mechanisms is a complex process involving several factors, not limited to the mode, species of interest, bacterial environment (*in vitro* or *in vivo*), and also the antibiotics [17]. Despite previous reports suggesting low rates of ARG transfer via conjugation [18], Leclerc *et al.* [17] reported that an estimated gene transfer rate cannot be generalized across all species and antibiotics due to the several variable factors as highlighted above.

1.2 Dissemination of antimicrobial resistance through horizontal gene transfer

Horizontal gene transfer (HGT) is key to the evolution and adaptation of bacteria, allowing for the rapid gain of beneficial traits including ARGs. [19]. Employing HGT, bacteria can acquire ARGs through either conjugation or transduction via mobile genetic elements (MGEs). In conjugation, plasmids carrying one or more resistance genes are transferred between microbes, while in transduction, bacteriophages encoding ARGs infect bacteria thereby transferring resistance [20]. The collective MGEs within a given microbiome in this context are defined as the 'mobilome'.

With respect to HGT or ARGs, plasmids represent an optimal vehicle. Plasmids are composed of either circular or linear DNA distinct from bacterial chromosomal DNA, capable of autonomous replication [21]. Besides encoding for resistance to most, if not all, major classes of antibiotics, multiple genes conferring resistance to different antibiotic categories can be found on the same plasmid. This is especially evident in the case of multidrug-resistant *Klebsiella pneumoniae*, against which antibiotic combination therapies are ineffective [22][23]. Furthermore, plasmids encoding ARGs are not only found within pathogenic bacteria but can also be detected in commensals [24]. Generally, the predisposition of a HGT event has been deduced to depend on ecological and phylogenetic factors [25]. However, as described by Porse *et al.* and others [25,26], the resistance mechanisms, during HGT events, in addition to the phylogenetic relatedness of the donor and recipient species act as crucial determinants of gene functionality and fitness cost. The functional compatibility of an ARG in a new host is dependent on the interaction with the host physiology and metabolism. Consequently, resistance mechanisms, i.e., drug-modifying enzymes, with limited cellular interactions are more likely to be functionally compatible and integrate easily into a novel host physiology [25]. These observations suggest that depending on the ARG and the plasmid, they can be shared between both closely and distantly related taxonomic clades, thereby contributing to wide-spread and rapid propagation of AMR [25,27]. Alongside plasmids, integrons, often overlooked, can also play a significant role in AMR dissemination and prevalence [28]. Integrons, widely distributed and carried by plasmids, can acquire, exchange and express genes embedded within gene cassettes [29] further promoting their spread within and between microbial communities [30]. Generally, two distinct groups of integrons have been described, namely chromosomal and mobile integrons. Chromosomal integrons are encoded by many bacterial

species and are also referred to as “super-integrans” due to their large size and ability to carry up to 2000 cassettes [31]. Mobile integrans, on the other hand, are located on MGEs such as plasmids or phages and have been associated with AMR and the dissemination of resistance among bacterial populations [28]. Collectively, integrans are efficient tools for bacterial adaptation and play a significant role in the spread of AMR in conjunction with plasmids.

Besides plasmids, (bacterio-)phages contribute to the horizontal gene transfer of ARGs via transduction. Transducing phages mediating AMR can be either virulent or temperate [32]. Upon infection, temperate phages integrate their DNA into the host chromosome in which the prophage subsequently becomes dormant. When induced by stress factors such as DNA damage [33] and/or environmental cues [34], the phage will be excised from the chromosome, inducing phage particle formation and lysis of the host cell [35]. In contrast, lytic phages immediately induce the formation of phage particles resulting in lysis of the host cell [36]. Additionally, transduction can be further separated into two types: generalized or specialized. In generalized transduction, the genetic material is transferred to another bacterial cell where it is further integrated through homologous recombination. Specialized transduction, on the other hand, results in the packaging of bacteria DNA into phages at a higher frequency compared to generalized transduction. This lateral transfer of ARGs through phage-mediated transduction could be an important contributing factor in the global spread of AMR [37].

1.3 Methods for detecting antimicrobial resistance

Traditionally, culture-based methods, such as antimicrobial susceptibility testing (AST), have been, and still are, used in clinical settings to investigate AMR and resistant bacteria [38]. For phenotypic testing, bacterial isolates are cultured from samples using either non-selective or selective growth media. Subsequently, the susceptibility of the isolates to antibiotics is tested to identify AMR. These solid media techniques use Kirby-Bauer disc diffusion or gradient diffusion strips to measure the zone of inhibition, and thereby provide a proxy for the level of resistance [38,39]. Although these methods provide crucial information regarding AMR, they are only suitable for bacteria which are readily culturable using standard cultivation methods. However, microbial communities such as those inhabiting the human gut are composed of significant proportions of, at present, difficult to culture or outright unculturable taxa. In this context, the ability to sequence DNA from samples (clinical and environmental) using high-throughput sequencing methodologies have improved our ability to investigate and identify AMR. Sequencing-based metagenomics, which involves the study of the total genetic material

(e.g., DNA or RNA) recoverable directly from samples, allows for the genomic analysis of all organisms within a microbial ecosystem without previous identification [40]. This enables the investigation of the resistome, i.e., AMR, including the mechanisms and spread of ARGs, without the immediate need to isolate microorganisms.

Different bioinformatic workflows have been developed to investigate the presence of AMR and MGEs within metagenomes. These include both read- and *de novo* assembly-based methods which have been extensively discussed by Boolchandani *et al.* [38]. While read-based methods allow for identification of low-abundance ARGs [38], *de novo* assembly strategies enable the genetic contextualization of AMR surveillance, such as their presence on MGEs [41]. Some of the AMR prediction tools including DeepARG [42], RGI [43], Resfinder [44], ARG-ANNOT [45], and NCBI-AMRFinder [46] can be used for ARG identification, albeit through use of their associated databases. For example, while DeepARG, RGI and NCBI-AMRFinder use the recently updated CARD database [43], other tools provide custom versions leading to discrepancies in identified ARGs. Nonetheless, none of the above tools provide information with respect to contextualization of ARGs on MGEs which represent critical elements for AMR transmission. Alternatively, many tools have been developed for the independent prediction of MGEs alone, such as PlasFlow [47], MOB-suite [48] and gplas [49] for plasmid identification. For the prediction of phages in general, the following tools exist: DeepVirFinder [50], VirSorter [51], MARVEL [52] and PPR-Meta [53]. Each of these tools are specialized to allow identification of a single or limited set of MGEs. While some tools like DeepVirFinder are based on machine-learning methodologies, others are restricted to databases populated with previously identified MGEs. The former allows for discovery of putatively novel MGEs, while the latter methods allow for precision and confidence in the identified MGEs. In a One Health context, bridging together human, animal and environmental health [54], it is crucial to study both the prevalence and spread of AMR simultaneously. Such methods to systematically assess AMR within and between biomes have long remained elusive [55]. Therefore, to precisely address this gap in methodologies, a pipeline is needed which genomically contextualizes ARGs, including their localization on MGEs. Tools such as MOCAT2 [56] and HUMAnN3 [57] also enable ARG identification, however, do not provide any information with respect to MGE contextualization. In a One Health setting, by combining effective study designs with computational analyses methods, it is thereby now possible to trace the origins and dissemination of AMR from one reservoir to another using metagenomic sequencing coupled to *de novo* reconstruction of genomic fragments.

1.4 Microbial reservoirs of antimicrobial resistance

Natural microbial communities, or microbiomes, represent multi-species assemblages which interact in a contiguous environment [58]. Current evidence suggests that the structure of human and animal microbiomes are shaped by several factors, including exposure to microorganisms through contacts with exogenous sources (e.g. parents, animals, environment), specific host-microbe interactions linked in particular to host immune responses, and the outcome of competitive, cooperative and/or predatory (phage) interactions [59]. Although in recent years an increase in AMR has primarily been pinned on the use and misuse of antibiotics in humans and in animals, there is strong evidence suggesting that AMR dissemination is fueled by other factors with the environment being an important conduit [60]. However, AMR in itself is an ancient phenomenon [1] that has largely evolved in response to natural antibiotics produced by microbes themselves to provide a competitive advantage. As a result of these microbe-microbe interactions bacteria have developed resistance strategies against these natural products to mitigate competition [61]. Additionally, to avoid suicide, antibiotic-producing microbes themselves often contain at least one gene conferring resistance against the potentially harmful secondary metabolites that the microbe produces [62,63]. Leveraging these naturally available compounds produced by bacteria, anthropogenic efforts have led to antibiotic production which are either natural products of microorganisms, semi-synthetically produced from natural products, and/or chemically synthesized based naturally available products [64]. Therefore, the use of antibiotics, both natural and (semi-)synthetic, has created unparalleled conditions for the spread of AMR through various reservoirs.

1.4.1 Human

It has long been recognized that the microbiome affects human health through its influence on gut maturation, immune responses, digestion of food, and pathogen resistance [65]. A majority of the microorganisms constituting the human microbiome are commensals contributing to both essential functions and physiological development. However, commensal and bacteria from the immediate and built environments can also be key distributors of AMR to the microbial community with the potential to spread to pathogenic bacteria [66,67]. Recent evidence suggests that ARGs in environmental bacteria can be taken up by human-associated and pathogenic bacteria [68], thereby posing a considerable threat to human health. Schmidt *et al.* demonstrated that the gut microbiota strains found in patients across five countries indicated an endogenous transmission, whereby strains found in the oral cavity were transmitted to the

gut [69]. Interestingly, the oral cavity has been reported to be a microbial reservoir contributing to the resistome [70] and it is plausible that this in turn is linked to the environment itself [71,72] including sanitary conditions [73]. While sanitary conditions such as open defecation, access to clean water have been discussed extensively [73], ARGs were recently discovered to be transmitted via air [74,75] in conjunction with a report by Gilbert *et al.* where ARGs were found in airborne bacteria found in a hospital setting [76].

During the recent decades, research has predominantly focused on AMR prevalence within clinically relevant bacteria. For example, extended spectrum beta-lactamases (ESBL)-producing and carbapenem resistant *K. pneumoniae* isolates have been characterized as early as 2001 by Yigit *et al.* [77]. Similarly, several studies have reported on the mechanisms of ESBL- [78–84] and plasmid-mediated AmpC-producing *Escherichia coli* [85,86] rendering the bacteria resistant to third-generation cephalosporins. From a surveillance perspective, Sepp *et al.* screened 10,780 clinical strains using whole genome sequencing to investigate the prevalence of ESBL-, AmpC-, and Carbapenemase-producing *E. coli* across northern and eastern Europe [87]. Despite a low prevalence of ESBL-, AmpC-, Carbapenemase-producing *E. coli* strains, they identified inter-country differences in the distribution and prevalence of resistance genes [87]. Other studies have included research on carbapenem-resistant *Acinetobacter baumannii* [88–90] and *Pseudomonas aeruginosa* [30,91,92], vancomycin-resistant *Enterococcus faecium* [93,94], methicillin-resistant *Staphylococcus aureus* [95–97], penicillin-resistant *Streptococcus pneumoniae* [98,99] as well as fluoroquinolone resistant *Salmonella* [100,101] and *Shigella* species [102,103]. More recently, less known human pathogens such as *Corynebacterium diphtheriae* isolates have been reported to carry penicillin, macrolide and multidrug resistance [104].

1.4.2 Livestock, poultry and other animals

In livestock and poultry, especially in food production, antibiotics are used as metaphylactics and prophylactics, for disease control and treatment, as well as for growth augmentation. On the one hand, metaphylactics involve the treatment of all animals belonging to the same flock or pen where a clinically sick animal is identified. This is a mitigation strategy which allows for treatment prior to observable clinical signs of disease, for example, by water-based medication [105] [106], simultaneously shortening the overall treatment period. Holman *et al.* [107] investigated the effect of metaphylactic antibiotic usage of the common veterinary antibiotics (oxytetracycline and tulathromycin) on the bovine fecal and nasopharyngeal microbiomes. In

addition to shifts in the microbial composition after the first five days of treatment, they found an increase in the relative abundance of several antibiotic resistance genes in both microbiomes at either day 12 or 34 after treatment [107]. Prophylactics, on the other hand, are used to either eradicate a specific pathogen or treat healthy animals as a preventive measure especially during periods of disease susceptibility, e.g. early weaning of piglets [106]. Despite the utility of such treatments including low-dose antibiotics, Agga *et al.* demonstrated that prophylactic treatment limited shipping fever in weanling pigs [108], they may over protracted periods of use result in a selective pressure yielding resistant bacteria. Consequently, in many countries the use of antibiotics as prophylactic or for pathogen eradication in livestock is prohibited [106].

Apart from their use in infectious disease management, antibiotics are also used as growth promoters, whereby industrialized animal production includes antibiotics as feed supplements [109]. The low concentrations of antibiotics, similar to the levels used in prophylactics, additionally raises the possibility of emergent resistant bacteria due to longer-term selective pressure. In this context, in a five-year longitudinal study, Aarestrup *et al.* investigated the use of growth promotion in pigs and broilers. They found a concomitant increase in AMR in *Enterococcus* spp. isolated from the animals. Moreover, the mitigation of AMR was associated with the banning of antibiotics as growth promoters over the years [110], strongly suggesting the need for measures to reduce the emergence of resistant bacteria.

Even though the emergence of resistant pathogens is a critical consideration, of more immediate concern is the spread of ARGs from the animal microbiome to human microbiota through the acquisition of ARG complements. Such spread can occur via multiple routes, one of which is the direct transmission through food products, i.e., meat and eggs, especially through confined animal feeding operations (CAFOs). Multiple studies have reported food animals as a source of AMR. Examples include multidrug-resistant *Salmonella* from poultry [111], cephalosporin resistant *E. coli* from veal calves [112] and carbapenem resistant *E. coli* from pigs [113,114] to name a few. In a study by Morrison and Rubin a number of carbapenem resistant bacteria including *Pseudomonas*, *Stenotrophomonas* and *Myroides* species were identified in a variety of seafood products [115]. This phenomenon reiterates the argument that non-pathogenic bacteria, regularly excluded from surveillance programs, may indeed serve as a reservoir for AMR along the food supply chain [115,116]. Furthermore, resistant bacteria may also be spread from animals to humans through direct contact such as in the agricultural sector

[59]. For example, in a study by Rinsky *et al.* livestock-associated multidrug-resistant *S. aureus* was identified in workers at an industrial livestock operation but was not detected in workers at an antibiotic-free livestock operation [117]. These reports collectively underline the need for a more comprehensive analysis and monitoring of livestock reservoirs of AMR.

Interestingly, CAFOs have also been reported to be AMR reservoirs and a source of resistant organisms in migratory birds [118]. Similarly, other studies following the migratory patterns of birds found multi-drug resistant bacteria (*Enterococcus* spp., *Salmonella* spp. and *Vibrio* spp.) in bird fecal material [119]. Other findings simultaneously highlight the role of migratory birds travelling to Bangladesh in disseminating extended-spectrum β -lactamase (ESBL)-producing *E. coli* [120]. Given the propensity for these birds to come in contact with humans in populated countries like Bangladesh, it is likely that these ARGs may have anthropogenic influences.

On the contrary, the role of human-influenced environments in disseminating AMR is largely unexplored. A comprehensive study by Plasa-Rodriguez *et al.* found AMR associated with several bacterial species in wild boar, roe deer, wild ducks and geese [121]. Atterby *et al.* [122] previously reported the possibility of human-mediated environmental pollution as a source of AMR in wild gulls. Simultaneously, other reports indicate that clinically relevant AMR bacteria have been found in synanthropic birds partially mediated via human-influenced habitats such as landfills or areas with intensive agriculture [123]. Such anthropogenic influences have spread even to the polar regions, where antibiotic-resistant *E. coli* were found in penguin feces, while ESBL-type resistant genes were observed in bacteria such as *E. coli* and *K. pneumoniae* isolated from both seawater and Arctic birds [124].

1.4.3 Environment

The environment is a critical factor for the prediction of emergent and resistant pathogens by understanding the presence, origins and mechanisms of dissemination of AMR. Polluted environments (e.g., with heavy-metals, biocides) further contribute to the evolution and spread of AMR through co-selection. For instance, through cross-resistance, a single genetic mutation may mediate resistance to both metals and antibiotics, or through co-segregation where both metal- and antibiotic resistance genes are localized on the same MGE [125,126]. The risk of a specific environment being contaminated with AMR is often based on the interaction between the different environments. Built environments in particular, e.g., hospitals and extended care facilities, where bacteria are exposed to high and repeated doses of antibiotics, are hotspots

of AMR. Hospitals in specific are of high interest to study both the evolution and dissemination of AMR through the prevalence of hospital-acquired infections of resistant bacteria. Resistant pathogens may enter the hospital environment via infected patients or acquire resistance through in-hospital evolution. In both cases resistant pathogens may spread epidemically between patients or the ARG itself can be transmitted through HGT into other genetic backgrounds [127]. Furthermore, sewage from both the hospital and the general population are ultimately transported to wastewater treatment plants (WWTP).

Urban WWTPs therefore provide a vast reservoir of antimicrobial resistance [128] and are considered to be AMR hotspots with respect to resistant bacteria and ARGs [129]. Moreover, the extensive dissemination of ARGs between various bacterial species through HGT may facilitate the transfer of ARGs to pathogenic bacteria. For example, Alexander *et al.* identified facultative pathogenic bacteria such as *E. coli*, *K. pneumoniae*, *P. aeruginosa*, and Enterococci with 12 clinically relevant ARGs within 23 different WWTPs [130]. Additionally, since WWTPs generally do not have the necessary measures to remove either ARGs or resistant bacteria, these are then released into the receiving water bodies, promoting their dissemination into and through the aquatic environment [130]. This is in line with a study by Osinska *et al.* where a significant increase in ARGs (e.g., *bla*_{TEM}, *tetA*, *sulI1*) was identified downstream of the WWTP when compared to the upstream river water [131]. A similar study by Bueno *et al.* reported a significant increase in 17 ARGs contributing to aminoglycoside-, beta-lactam-, diaminopyrimidine-, fluoroquinolone-, sulfonamide-, tetracycline- and multidrug-resistance, in the receiving water of three different WWTPs [132], thereby highlighting the overall role and impact of the built-environment in AMR dissemination.

The contamination of natural environments with antibiotics originating from built environments as well as agricultural sources, results in selective pressure promoting both the evolution and the spread of ARGs. Additionally, many antibiotics are naturally produced by fungal and bacterial strains and consequently have been used by microorganisms as a competitive mechanism [133,134]. Due to their high complexity and multi-faceted microbe-microbe interactions, soil microbiomes are considered a hotbed for the evolution and development of AMR [135]. Multiple bacteria identified in soils encode genes that either degrade or inactivate antibiotics. For instance, Dantas *et al.* isolated hundreds of soil-dwelling bacteria capable of utilizing antibiotics as a carbon source and found up to 17 antibiotics, including those of synthetic origin, supporting the growth of clonal soil bacteria [136]. Furthermore, bacteria

isolated from forest, urban and agricultural soils have been found to have highly varied resistomes, even in some cases harboring novel mutation sites conferring resistance [137]. Linked to agricultural soils, the plant rhizosphere is of further interest due to the transmission of ARGs from soil to plants via the rhizosphere microbial community. Wolters *et al.* investigated the effect of various organic soil fertilizers such as manure and found increased relative abundances of sulfonamide and tetracycline resistance in the maize rhizosphere [138]. Similarly, Song *et al.* investigated the abundance of 35 antibiotic resistance genes in the rhizosphere of 10 plant species and identified a positive association between ARGs and MGEs [139].

Similar to soils, aquatic environments also represent known reservoirs of AMR. Aquatic habitats harbor resistant microbes such as carbapenem-resistant *Acinetobacter* spp. in rivers [140], carbapenem-resistant *Pseudomonas* in coastal waters [141], and carbapenem-resistant Enterobacteriales in seawater [142]. Environments are further affected greatly when in proximity to anthropogenic activity such as pharmaceutical industries. Consequently, these environments are abundant with ARGs and multidrug-resistant bacteria which have been associated with a high impact on human health [143]. For instance, Flach *et al.* found that antibiotic-polluted lakes harbored considerably higher proportions of ciprofloxacin- and sulfamethoxazole-resistant bacteria as well as several novel multi-resistance plasmids compared to non-polluted lakes [144]. Additionally, Kristiansson *et al.* identified a similar phenomenon in river sediments exposed to antibiotic pharmaceutical wastewater and reported high levels of ciprofloxacin-resistance as well as corresponding mobile quinolone resistance genes [145].

AMR, on the other hand, does not exclusively exist in human-impacted environments. As several studies have revealed, vast reservoirs of AMR are also found in environments pre-dating the antibiotic era [134]. These include glacier lakes, remote lakes [90] and oceans [103, 104]. Polar regions in particular, as one of the least human-impacted environments to date, are of interest for the study of AMR. Arctic soil isolates have previously revealed the presence of multidrug efflux pumps [146], while in a study by Dancer *et al.*, bacterial isolates from arctic glacial ice and water were found to carry resistance to antibiotics such as cefazolin, cefamandole and ampicillin [147]. The melting of glaciers and icecaps due to climate change, therefore, may give important insights into, potentially, prehistoric mechanisms of AMR. On the

other hand, this may also lead to the remobilization of ARGs, which we have not seen since before the dawn of human evolution.

1.5 Metagenomic approaches in assessing antimicrobial resistance: a One Health perspective

Resistant bacteria residing within human, animal and environmental reservoirs may spread from one to the other, at both local and global levels (**Figure 1.1**). This phenomenon has the potential to rapidly trigger a pandemic where AMR is no longer constrained by either geographic or human-animal borders [148]. It is therefore necessary to understand the dissemination of antibiotic resistance by characterizing the resistome within various environments and to unravel how they act as a reservoir for bacterial pathogens in the context of overall pandemic preparedness. A One Health perspective integrating research on AMR as well as resistant microbes, circulating in humans, animals and the environment is therefore crucial to enhance our understanding of the complex epidemiology of antimicrobial resistance [148].

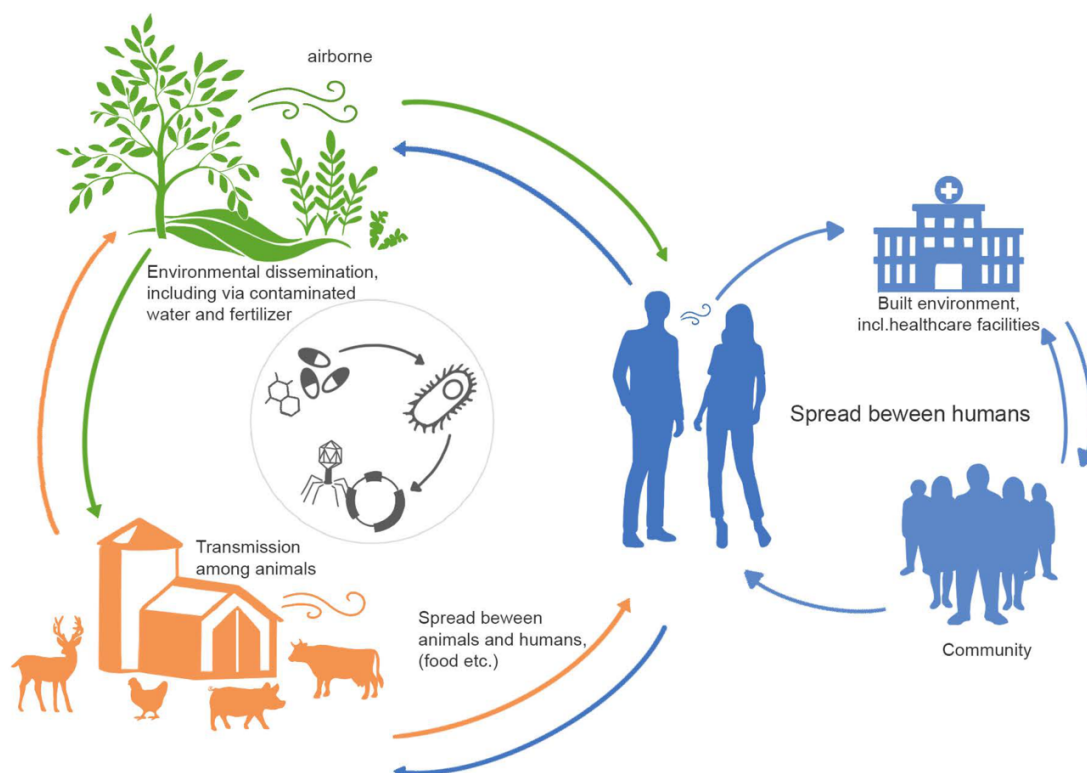


Figure 1.1: AMR dissemination in One Health. MGE-mediated (i.e., phage, plasmid and integrons) dissemination of AMR across different biomes)

In recent years, many studies, of which some have been discussed in the previous sections, have used different techniques to sample the resistomes of soils, wastewater, as well as human and animal microbiota [16]. Recent metagenomic studies by both Gibson *et al.* [114] and Munck *et al.* [115] suggest that ARGs predominantly cluster by microbial reservoir, implying that the resistomes in soils and WWTPs differ significantly from those found in human pathogens. Gibson *et al.* found that resistance against β -lactams and tetracyclines differed mostly between ecosystems [149] while Munck *et al.* highlighted that only a few genes within the WWTP core resistome were found in other environments [150]. Nonetheless, part of these resistomes may still be shared and the importance of continued exploration of the resistome in such environments should be stressed [16]. While shared resistome elements between various microbial reservoirs are of interest to understand the dissemination of AMR, resistome differences between ecosystems are equally, if not more important. They represent a pool of potential novel resistance mechanisms and thereby a likely threat to public health.

As described, several published studies have investigated AMR in humans, animals or the wider environment. However, many of these focus specifically on the ESKAPEE pathogens (*Enterococcus faecium*, *S. aureus*, *K. pneumoniae*, *A. baumannii*, *P. aeruginosa*, *Enterobacter* spp., and *E. coli*), which have been classified by the World Health Organization for their high to critical drug-resistance. They are also of particular interest due to their increased resistance to last-resort drugs [151]. There presently exists no lack in reports of the resistance mechanisms encoded by the ESKAPEE pathogens in different microbial reservoirs. Methicillin-resistant *S. aureus* (MRSA) for instance, has been reported by van den Broek *et al.* [152] and Lewis *et al.* [153] to be both human- and animal-associated with a high risk for zoonotic transmission. Similarly, Ruiz-Roldan *et al.* reported the presence of resistant *P. aeruginosa* in animals in addition to humans. On the other hand the drug resistant strains of ESKAPEE pathogens belonging to the Enterobacteriales order (i.e. *E. coli* and *P. aeruginosa*) have been extensively described in all microbial reservoirs [77–79,142,154,155]. Recent research has been extended to focus on other pathogens posing a threat to human health such as resistant *Campylobacter jejuni* where infections have been reported in both humans, animals and the environment [46,156,157][46,156]. Similarly, other reports include multidrug-resistant *Salmonella* which have been identified in human [158,159], animal [158,160] and environmental reservoirs [161].

While the above studies are focused on specific pathogens or resistance categories, research utilizing sequence-based metagenomics provides a comprehensive perspective on all ARGs within different microbial reservoirs. For instance, Forslund *et al.* provide extensive insights into the human gut resistomes of 832 individuals spanning 10 geographical areas. They reported significant differences in gut resistance potential between countries resulting from differences in antibiotic usage as well as direct links to medical and food production activities [162]. Other metagenomic studies have focused on the development of the resistome early on in life with several studies reporting a diversity of ARGs within the infant gut [163,164]. During the first days of life the bacteria colonizing the infant gut originates primarily from the mother's birth canal, the living environment and handling by other individuals. Birth mode affects colonization since vaginally born infants are colonized firstly by fecal and vaginal bacteria from the mother, while infants born via cesarean section are initially exposed to bacteria originating from both the hospital environment and healthcare workers [65,165]. Therefore, infants born by cesarean section may also have a higher chance of acquiring hospital-mediated AMR and thereby resistant bacteria [163]. Other metagenomic studies have focused on the animal resistome, especially food production animals, such as dairy cattle, revealing an increase in AMR linked to heavy metal-contaminated environments [166]. Furthermore, a study by Skarzynska *et al.* leveraged metagenomic data to study AMR in the gut of both wild (boars, foxes and rodents) and domestic (chicken, turkey and pig) animals. Importantly, they identified increased AMR abundance in farm animals compared to wildlife [167]. Furthermore, the lowest AMR abundance in this study was observed in wild rodents due to their limited exposure to antimicrobials. In this context, further evidence was found linking ARGs conferring resistance to important antimicrobials such as quinolones and cephalosporins to wild foxes [167]. Alongside human and animal studies, metagenomic studies on the environmental resistome focus on characterizing AMR either in WWTPs or the natural environment or built environment (e.g. healthcare facilities). However, few are specifically tailored towards understanding the role of the environmental ecosystems as microbial reservoirs of AMR, especially in a One Health setting.

The few metagenomic studies that are focused on multiple microbial reservoirs still largely target only one side of the One Health triad, e.g. human-animal [168–171], animal-environment [172–175] or environment-human [149,150,176–179]. Nonetheless, some studies have pursued a complete One Health AMR approach [180,181]. Li *et al.* investigated wide-spectrum profiles of ARGs and their co-occurrence patterns in 50 samples from 10 microbial reservoirs,

spanning human, environment and animal habitats. They found that samples could be clustered into four groups according to AMR abundance, with samples derived from livestock and wastewater demonstrating the highest abundance followed by humans, and with the lowest abundance found in sediments, soil, river and drinking water, in that particular order. A widespread occurrence of vancomycin resistance genes was identified in all environments except from river sediments and drinking-water [181]. Another study by Pal *et al.* investigated AMR, MGEs and bacterial taxonomic compositions of 864 human, 145 animal and 369 environmental metagenomes. Both human and animal microbial communities demonstrated a limited taxonomic diversity, a low abundance/diversity of biocide and metal resistance genes and MGEs, yet a high abundance in ARGs. Additionally, a number of ARGs corresponding to aminoglycoside, macrolide, beta-lactam and tetracycline resistance was found to be widespread and present in almost all of the investigated environments [180]. Collectively, these studies report the cross-domain similarities and likely transmission of AMR in a One Health setting, potentially highlighting the need for more in-depth characterization of AMR transmission mechanisms.

Chapter 2. PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data.

A major part of this chapter was adapted and modified from the following first-author peer-reviewed publication:

Laura de Nies, Sara Lopes, Susheel Bhanu Busi, Valentina Galata, Anna Heintz-Buschart, Cedric Christian Laczny, Patrick May and Paul Wilmes (2021). PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. *Microbiome* 9 (1), 1-14 [**Appendix A.2**]

2.1 Background

Most of the microorganisms constituting the human microbiome are commensals [182]. They contribute essential functions to the human host and contribute to its physiological development. In contrast, pathogenic microorganisms including bacteria, viruses, fungi, and protozoa cause disease by invading, colonizing and damaging the host. Virulence factors, including bacterial toxins amongst others, contribute to this pathogenicity by enhancing not only the infectivity of pathogenic bacteria but also by exacerbating antimicrobial resistance which in turn restricts treatment options [182].

Virulence factors enable pathogenic microorganisms to colonize host niches, ultimately resulting in tissue damage as well as local and systemic inflammation. These factors are important for pathogens to establish an infection and span a wide range, thus contributing both directly and indirectly to disease processes [183]. These virulence traits include cell-surface structures, secretion machineries, siderophores, regulators, etc. [184,185]. Of all virulence factors employed by pathogens, bacterial toxins often have a crucial function in the pathogenesis of infectious diseases [186]. Different types of bacterial toxins have evolved over time to counteract human defenses. These bacterial toxins can be coarsely categorized into two groups: the cell-associated endotoxins and the extracellular diffusible exotoxins. Exotoxins are typically polypeptides and proteins that act to stimulate a variety of host responses either through direct action with cell receptors or via enzymatic modulation [186,187].

Partly through the utilization of these virulence factors, and toxins in particular, pathogenic microorganisms have been a major cause of infectious diseases including in the context of viral co-infections [182]. The development and medical use of antibiotics has limited the development and spread of these pathogens by providing an effective treatment for bacterial infections. However, the over- and mis-use of antibiotics has resulted in a global increase in antimicrobial resistance (AMR) which now threatens human health through the emergence and spread of multidrug resistant bacteria [182,188] (**section 1**). Furthermore, the acquisition of antimicrobial resistance genes (ARGs) is not restricted to a single strain or species of bacteria. While commensal bacteria provide a source of ARGs, antimicrobial resistance can be transferred to pathogenic species through horizontal gene transfer, e.g., conjugation or transduction (**section 1.2**) [20,189,190]. As a result, many pathogenic bacteria have now acquired resistance against the main classes of antibiotics which has led to a dramatic rise in untreatable infections, resulting in the emergence of so-called “superbugs” [191].

Consequently, AMR is an urgent and growing threat to public health with an estimated number of deaths exceeding ten million annually by 2050 [5,6].

Pathogenic microorganisms have modified and adapted their virulence to host defense systems over millions of years. Similarly, AMR is thought to have evolved over extensive periods of time in bacteria, indicating that it is an ancient phenomenon [1]. However, with an increase in selective pressure through the use of antibiotics an excessive increase in the spread and evolution of AMR has been observed in the last fifty years. Yet, despite differences in evolutionary paths, virulence factors and AMR share common characteristics. Most importantly, virulence factors and AMR are necessary for pathogenic bacteria to adapt to, and survive in, competitive microbial environments [188]. Additionally, both virulence and resistance mechanisms are frequently transferred between bacteria by horizontal gene transfer [190]. Furthermore, both processes make use of similar systems (i.e., cell wall alterations, efflux pumps, two-component systems and porins) that activate or repress the expression of various genes [192–194]. Therefore, although AMR in itself is not a virulence factor, in environments with selective antibiotic pressure, opportunistic pathogens are able to colonize through acquisition or presence of AMR [182].

Considering the burden of bacterial infections in which virulence factors and ARGs play crucial roles, it is important to be able to identify these in microbial communities in situ. The advent of high-throughput DNA sequencing provides a powerful means to profile the full complement of DNA derived from genomic extracts obtained from a wide range of environments [42]. As such metagenomic sequencing represents a pertinent technique for in situ studies as it provides less biased view of the genomic complements of individual microbial populations compared to amplicon-based methods [195,196]. However, currently there is a lack of automated pipelines to simultaneously identify these different factors in metagenomic datasets. Various tools exist for the prediction of ARGs themselves, such as DeepARG [42], RGI [43], ResFinder [197] and ARGsOAP [198], with a very few prediction tools for virulence factors existing, such as MP3 [199] and VirulentPred [200]. Most of the latter tools are based on outdated databases of virulence factors which have since been expanded greatly. Moreover, there is a lack of recent bioinformatics tools for the prediction of bacterial toxin genes in particular. Furthermore, although various AMR prediction tools exist, these primarily focus on the prediction of genes without considering their location, i.e., these tools do not differentiate between localization on mobile genetic elements (MGEs) or on bacterial genomes. Since MGEs are the main

mechanism by which ARGs are transmitted, it is crucial to identify the relationship between ARGs and MGEs. Outside of these prediction tools, it is common practice to use standard homology search algorithms against specific databases. However, such practices require several intermediate steps which may vary from lab to lab. Additionally, using these methods is restrictive in the sense that only a single database can be searched at a time.

Here, we present PathoFact, a pipeline for the simultaneous prediction of virulence factors, bacterial toxins in particular, and ARGs. Our tool furthermore contextualizes these with respect to their localization on MGEs. Moreover, PathoFact aggregates the information obtained via different prediction tools and databases into a single output, allowing both novices and experts in bioinformatics alike to parse information as needed. PathoFact thus provides a unified perspective on pathogenic mechanisms. We provide evaluation results on our tool's sensitivity, specificity and accuracy, and demonstrate PathoFact's versatility using both a simulated metagenomic dataset and public case-control metagenomic datasets for Parkinson's disease, psoriasis, and *Clostridioides difficile* infection. Using the simulated metagenomic dataset, we further perform a comparison of PathoFact with other metagenomic characterization workflows namely MOCAT2 [56] and HUMANN3 [201].

2.2 Methods

2.2.1 PathoFact architecture

PathoFact is a command-line tool for UNIX-based systems that integrates three distinct workflows for the prediction of (i) virulence factors, (ii) bacterial toxins, and (iii) antimicrobial resistance genes from metagenomic data (**Figure 2.1**). Each workflow can be applied individually or in combination with the other workflows. Our tool is written in Python (version 3.6) and uses the Snakemake (version 5.5.4) workflow management software [202]. This implementation offers several advantages, including workflow assembly, parallelism, and the ability to resume processing following an interruption. Each step of the pipeline is implemented as a rule in the Snakemake framework specifying the input needed and the output files generated. We use conda (version 4.7) environments wherever possible thus reducing the need for explicit installation of software dependencies. Moreover, the use of conda environments makes it possible to incorporate prediction tools dependent on older Python versions incompatible with version 5.5 of Snakemake. As such, Python, Snakemake and (mini)conda (version 4.7) [203] installations are required. PathoFact is open-source and freely available at <https://pathofact.lcsb.uni.lu>.

The input to the PathoFact pipeline consists of an assembly fasta file containing nucleotide sequences of the contigs. PathoFact subsequently predicts the ORFs using Prodigal (version 2.6.3) [204] for the prediction of virulence factors, toxins and antimicrobial resistance genes. The MGEs are predicted from the initial assembly file and a mapping file is generated by PathoFact which aggregates all the results. PathoFact aggregates the information obtained from the different sub modules into both module-specific reports as well as a complete final report. The reports describe all virulence factors, bacterial toxins and antimicrobial resistance genes identified from the input as well as their assigned confidence level (virulence factors/ bacterial toxins), their resistance mechanisms (AMR) and their corresponding localization on MGEs.

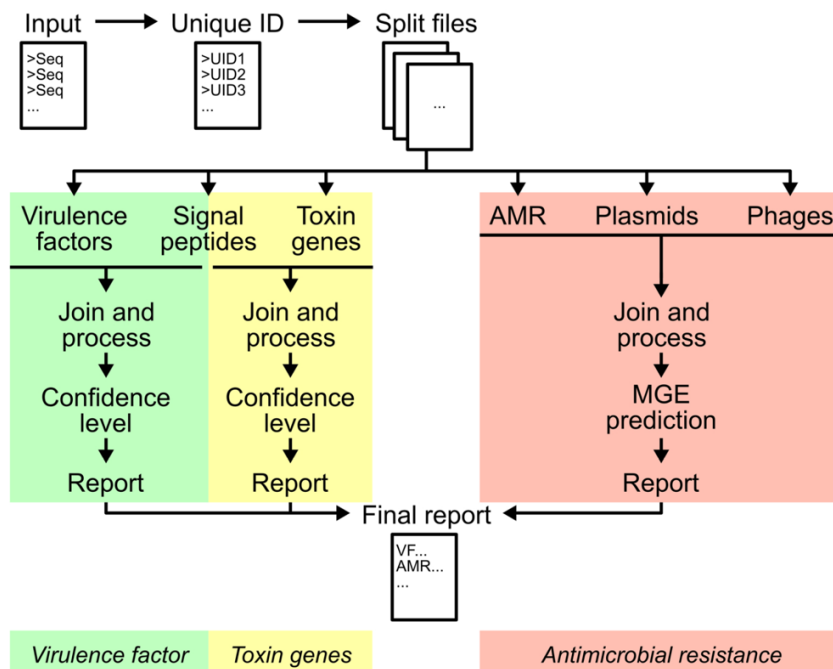


Figure 2.1: Framework of the PathoFact pipeline. The pipeline consists of three different modules related to (i) virulence factors, incl. (ii) bacterial toxins, and (iii) antimicrobial resistance genes. All modules can either be run independently or jointly.

2.2.2 Workflow for the prediction of virulence factors

For the prediction of virulence factors, we created a prediction tool consisting of two parts; (i) a database consisting of virulence factor HMM profiles (HMMER3 v3.2.1) [205], and (ii) a random forest model. Hits against the virulence factor HMM database are then combined with the classification of the random forest model to result in the final prediction (**Figure 2.2**). The development of the tool was inspired by the MP3 software tool for the prediction of virulence

factors which has not received an update since 2014 and was thus outdated [199]. In addition, PathoFact combines these annotations with the prediction of signal peptides by SignalP (v5.0) [206] to distinguish between secreted and non-secreted virulence factors.

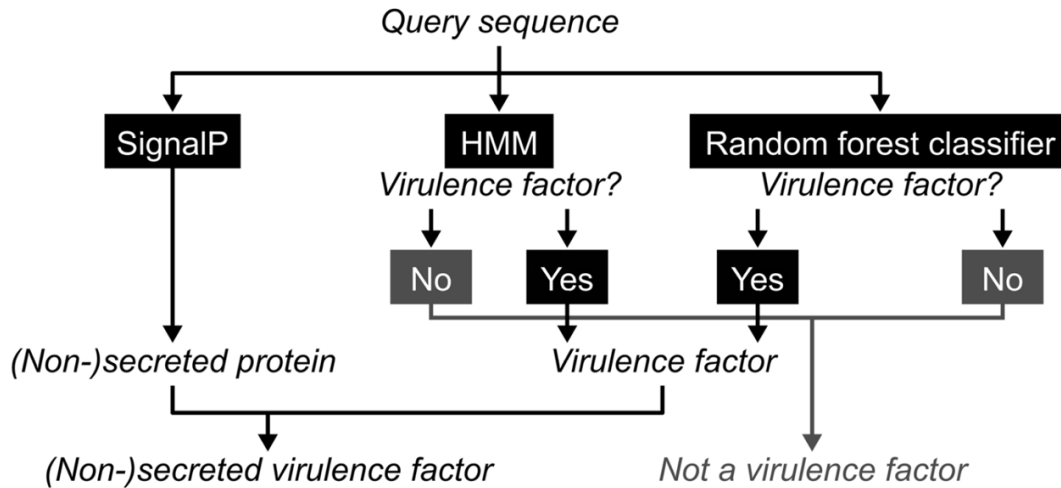


Figure 2.2: Classification framework for the prediction of virulence factors. The prediction of virulence factors depends on two different aspects: (i) a HMM domain database, (ii) a random forest classifier. Sequences predicted positive from both are classified as virulence factors. The incorporation of SignalP in the framework allows integration of information regarding the likely secretion of the virulence factors.

2.2.2.1 Dataset for the prediction of virulence factors

A dataset, consisting of both a positive and negative subset, was constructed for the training of the virulence factor prediction tool. The positive subset consisted of known virulence factor sequences retrieved from the Virulence Factors Database (8945 sequences) (VFDB) [184]. All sequences were obtained from the VFDB core dataset containing (translated) gene sequences associated with experimentally verified virulence factors. The negative subset of the training set consisted of protein sequences that were retrieved from the Database of Essential Genes (DEG) (7995 sequences) [207] and which were known not to be virulence factors. For both subsets, all sequences were clustered with CD-HIT [208] and sequences with a 90% sequence identity were collapsed to prevent redundancy within the subsets. This 90% cutoff is routinely used to reduce redundancy in similar protein datasets, improving efficiency without foregoing specificity given the large metagenomic database sizes [209,210]. The resulting training set was used for (i) the implementation of the HMM profiles, and (ii) the training of the random forest model.

2.2.2.2 Construction of the virulence factor HMM Database

For the construction of the virulence HMM database, HMM profiles were annotated for the training set using HMMER3 (version 3.2.1) against multiple pre-compiled and in-house annotation databases [211]: PFAM-A [212], TIGR [213], KEGG [214], MetaCyc [215] and Swissprot [216]. The best hit in each HMM set was assigned to each gene in the training set if the HMM score was higher than the binary logarithm of the number of target genes, in accordance with the recommendations in the HMMer manual. HMM profiles were subsequently retrieved, and the databases were concatenated to form the virulence HMM database. Binary compressed data files were constructed with the hmmpress (HMMER3 v3.2.1) [205]. For the prediction of virulence factors by the virulence HMM database, identified HMM profiles are separated by those matching the positive or negative subset of the training set, as well as HMM profiles ambiguous for both positive and negative subset.

2.2.2.3 Machine learning model for the prediction of virulence factors

In addition to the virulence HMM database, we created a random forest model [217]. A random forest model operates from decision trees and outputs classification of the individual trees while correcting for overfitting of the training set. While overfitting, in which models perform highly on the training set but poorly on the test set, is a common problem in machine learning, a random forest model corrects for overfitting by continuously creating trees on random subsets. This does not mean that random forest classifiers are not capable of overfitting. However, they are less sensitive to variance and effects of overfitting are therefore rarely observed [218]. For training of the random forest model, the following five features of the sequences were selected and implemented: amino acid composition (AAC), dipeptide composition (DPC), composition (CTDC), transition (CTDT) and distribution (CTDD) [219]. A feature matrix was built with rows corresponding to the sequence composition of the features. The random forest model was implemented using pandas (v 0.25.0) [220], numpy (v 1.17.0) [221] and scikit-learn (v0.21.3) [222] and consisted of 1600 trees with a maximum depth of 340.

2.2.3 Workflow for the prediction of toxin genes

For the prediction of toxin genes, a workflow consisting of a toxin HMM database combined with SignalP version 5.0 [206] was developed. The toxin HMM database consists of bacterial toxin domains to identify toxin-related domains in the query sequences. Using the hmmsearch function of the HMMER3 (v3.2.1) program [205], the input query sequences are searched

against the collection of profiles present in the toxin HMM database. In addition, analyses are combined with SignalP [206] to differentiate between secreted and non-secreted toxins.

2.2.3.1. Construction of the toxin HMM database

For the toxin HMM database an HMM model based on a training set of known toxins was developed and implemented. The training set was compiled from the Toxin and Toxin Target Database (T3DB) [223] and the training set derived from the DBETH prediction tool [186]. Protein sequences from within the training set with a similarity greater than 90% were clustered and collapsed with CD-HIT-2D to reduce redundancy [208]. The corresponding toxin HMM profiles were identified from the same five HMM databases as used for the virulence factors (see above). The datasets were extended with HMM profiles already annotated as bacterial toxin domains in the PFAM, TIGR, KEGG, MetaCyc and Swissprot databases. Finally, in order to have a short description of all HMM profiles present in the toxin HMM database, a toxin library was created. These lists (i) all HMM profiles, (ii) their names, (iii) their alternative names, and (iv) the original database from which the HMM profile was derived.

2.2.4 Workflow for the prediction of antimicrobial resistance genes

For the prediction of ARGs, the workflow is separated into two parts: (i) the prediction of ARGs, and (ii) the prediction of MGEs. For the prediction of ARGs, the tools DeepARG (v1.0.1) [42] and RGI (v5.1.0) [43] are used. DeepARG uses a deep learning approach that improves classification accuracy while at the same time reducing false negatives. It offers a powerful approach for metagenomic profiling of ARGs as it expands on the available databases for ARGs by combining the widely used CARD [224], ARDB [225], and UNIPROT [226] databases. Additionally, RGI [43], is included which is able to identify mutation-driven AMR within genes, allowing for a strain-resolved profiling of ARGs.

2.2.5 MGEs: plasmids and phages

The prediction of MGEs is split into two parts focusing on the prediction of (i) plasmids, and (ii) phages. For the prediction of plasmids, PlasFlow (v1.1) [47] is used, while for the prediction of phages VirSorter (v1.0.6) [51] and DeepVirFinder (v1.0) [50] were incorporated. All three tools were selected because of their performance compared to other, similar tools [47,50,51]. The predictions of these different tools are merged with the prediction of ARGs to provide localization information of the resistance genes to either MGEs or genomes. Considering the different predictions of MGEs, the final classification includes plasmid, phage, genome,

unclassified, and ambiguous when localization predictions contradict each other, for example predicted to be both phage and plasmid.

2.2.6 Evaluation of the PathoFact pipeline

To evaluate the performance of PathoFact, validations were conducted for the prediction of toxins, for virulence factors, and for ARGs. The prediction quality was evaluated by sensitivity, specificity and accuracy criteria as defined below.

$$Sensitivity = \frac{tp}{tp + fn} \quad Specificity = \frac{tn}{tn + fp} \quad Accuracy = \frac{tp + tn}{tp + fn + tn + fp}$$

Where tp represents true positives (i.e., virulence factors (incl. bacterial toxins) or ARG is predicted correctly), tn (i.e. a gene is correctly predicted not to be a virulence factor, toxin genes or ARG), fp false positive (i.e., a gene incorrectly identified as a virulence factor, toxin genes or ARG), and fn false negatives (i.e., a virulence factor, toxin genes or ARG is incorrectly identified as non-pathogenic). We evaluated the sequence similarities between the training and validation (test set) datasets after removing the sequences from the validation set with 90% identity to the training set sequences using sourmash [227] (**Appendix B.1: Supplementary figure 2.1**).

2.2.6.1 Validation of virulence factors

A validation dataset was constructed to assess the performance of the prediction of virulence factors. Analogous to the training set, the validation set consisted of a positive subset of 2639 sequences (VFDB database) and a negative subset of 2628 (DEG database) sequences. Importantly, the sequences in the validation dataset were removed from the training set to avoid overfitting. The test set for virulence predictions was used to run both the standalone MP3 (v1.0) tool and our newly generated tool for prediction of virulence factors. For MP3 the standard advised parameters were used: set on metagenomic protein fragments, a minimum length of 90 bases and a threshold value of 0.2 for the svm module [199].

2.2.6.2 Validation of toxin genes

For the validation of toxin genes, a validation dataset containing both positive and negative subsets was constructed. The positive subset was constructed from sequences in the EMBL-EBI database annotated as bacterial toxins. The results were limited to protein sequences described in the UniProtDB. Further filtering of the protein sequences removed sequences with uncertain predictions (i.e., hypothetical, probable). To limit redundancy within the dataset,

sequences were clustered in terms of similarity by using a 90% sequence identity cut-off. Furthermore, to limit redundancy between the validation and the training set, sequences with a similarity of greater than 90% were discarded. The remaining 202 positive sequences were combined with 202 random selected sequences from the negative dataset, consisting of housekeeping genes representing the validation dataset.

2.2.6.3 Validation of AMR prediction

For the prediction of ARGs, both the DeepARG and RGI prediction tools were used. DeepARG has proven to be more accurate than most AMR prediction tools with a great reduction in false negatives [42], while RGI is capable of annotating SNPs contributing to AMR. For further validation, before inclusion in the pipeline, the prediction tools were tested using the NCBI's resistance gene database (5265 sequences) [46]. This positive subset was combined with a negative subset (consisting of sequences retrieved from the Database of Essential Genes) of equal size. For DeepARG default settings were applied, while parameters for *model* were set to LS and type was set to *prot*. Similar to DeepARG, default settings of RGI were applied while input-type was set to *protein*.

2.2.7 Data analysis and data availability of publicly available datasets

Metagenomic sequences for the publicly case-control metagenomic datasets were obtained from the European Bioinformatics Institute-Sequence Read Archive database, with accession numbers PRJNA297269 (Milani *et al.* [228]), PRJNA281366 (Tett *et al.* [229]) and ERP019674 (Bedarf *et al.* [230]). Information on the analyzed samples per study can be found in **Appendix C.1: Supplementary table 2.1**. Metagenomic reads were processed and assembled using IMP (v2) [231]. The resulting fasta files containing the assembled contigs and genes were used as input for PathoFact. For analyses of the predictions, FeatureCounts (v1.6.4) [232] was used to extract the number of reads per functional category. Thereafter, the relative abundance of the toxin genes was calculated using the Rnum_Gi method described by Hu *et al.* [233]. Additionally, the DESeq2 (v1.24) [234] package was used to analyze the differential abundance of virulence factors, toxins and ARGs.

2.2.8 Data analysis and data availability of a simulated dataset

To evaluate the performance of PathoFact compared to other metagenome characterization workflows, a high-complexity stimulated dataset consisting of 5 time series samples with 596 genomes and 478 circular elements was obtained from CAMI [64]. As with the case-control

metagenomic datasets, reads were processed and assembled using IMP (v2), after which the dataset was run through PathoFact. In addition, both MOCAT2 and HUMAnN3 were run on the stimulated metagenomic dataset using default settings of both workflows. Further data analysis was performed as described for the case-control datasets.

2.3 Results and Discussion

2.3.1 Benchmarking

The PathoFact pipeline has an in-built multi-threading option to improve computational efficiency. In fact, certain tools, e.g., DeepVirFinder, are memory intensive and may require additional resources. Table 2.1 corresponds to the runtime of a metagenomic dataset (363 933 metagenomic sequences) with differing numbers of threads. A minimum usage of 8 threads, in this case corresponding to 28 GB/thread, is advised for running the pipeline. Additionally, for the installation of PathoFact an initial storage of 6.3 GB is required.

Table 2.1: PathoFact runtimes with different threads/computational resources. Evaluated running times of PathoFact with different threads (8,16) and corresponding computation resources.

Threads	Memory	Running time
8	224 GB	25h 19m
16	448 GB	15h 58m

2.3.2 Validation of the PathoFact pipeline

For the prediction of virulence factors the prediction tool consists of two parts: a virulence factor HMM database and a random forest classifier. The random forest classifier's out-of-bag (OOB) error value reported an accuracy of 0.822. To improve performance for virulence prediction, the random forest model was combined with the HMM database which resulted in an overall sensitivity of 0.886, specificity of 0.957 and an accuracy of 0.921 (Table 2.2). Additionally, we compared our tool to the MP3 tool for the prediction of virulence factors (**Appendix C.1: Supplementary table 2.2**). PathoFact scored overall higher than MP3 which scored 0.125, 0.992, 0.558, respectively. In addition to the prediction of virulence factors, for the prediction of bacterial toxins an overall sensitivity of 0.777, specificity of 0.989 and accuracy of 0.832 were obtained.

Finally, for the prediction of ARGs the sensitivity, specificity and accuracy of both DeepARG and RGI was determined at 0.720, 0.996, 0.858 and 0.920, 0.997, 0.958, respectively. A combined approach merging the use of both tools resulted in the highest scores with an overall sensitivity of 0.963, specificity of 0.994 and accuracy of 0.979 for the prediction of ARGs.

Table 2.2: Validation of the PathoFact pipeline. Evaluated performance of PathoFact regarding the prediction of virulence factors, bacterial toxins, and antimicrobial resistance genes.

	Toxin prediction	Virulence factor prediction	AMR prediction
Sensitivity	0.777	0.886	0.963
Specificity	0.989	0.957	0.994
Accuracy	0.832	0.921	0.979

2.3.3 Performance evaluation using a simulated dataset

To further evaluate the performance of PathoFact and compare it to other existing tools, the PathoFact pipeline was run on a simulated metagenome comprised of high-quality annotated genomes, i.e., the CAMI High Complexity Toy Test Dataset. Both MOCAT2 [56] and HUMAnN3 [201] were run on the original reads of the simulated CAMI datasets, while the same read datasets were processed and assembled with IMP followed by execution of PathoFact. Subsequently, annotations resulting from the different workflows were compared to evaluate the performance of PathoFact (**Figure 2.3a**). PathoFact demonstrated increased numbers of predictions compared to both MOCAT2 and HUMAnN3 regarding virulence and toxin predictions (< 0.05 , ANOVA) while performing similarly regarding AMR prediction compared to MOCAT2. Furthermore, and importantly, no additional curation or data-wrangling is needed for PathoFact compared to the other workflows tested above.

Additionally, we aimed to further characterize the performance of the metagenomic workflows against annotations of the CAMI High Complexity Toy Test Dataset. To achieve this, we annotated the underlying genomic data using the NCBI database of resistance genes [46], as well as a BLAST search of the original 450 genomes against known virulence factors and toxin genes [184,186]. The resulting annotations were compared to the prediction reports of PathoFact, MOCAT2 and HUMAnN3. PathoFact identifies a similar number of virulence factors and toxin genes in the annotated genomes compared to the original annotations, while MOCAT2 and HUMAnN3 identified a significantly lower number (**Figure 2.3b**). Regarding

antimicrobial resistance, PathoFact was able to identify many more gene variants compared to MOCAT2 and HUMAnN3 (**Figure 2.3c**).

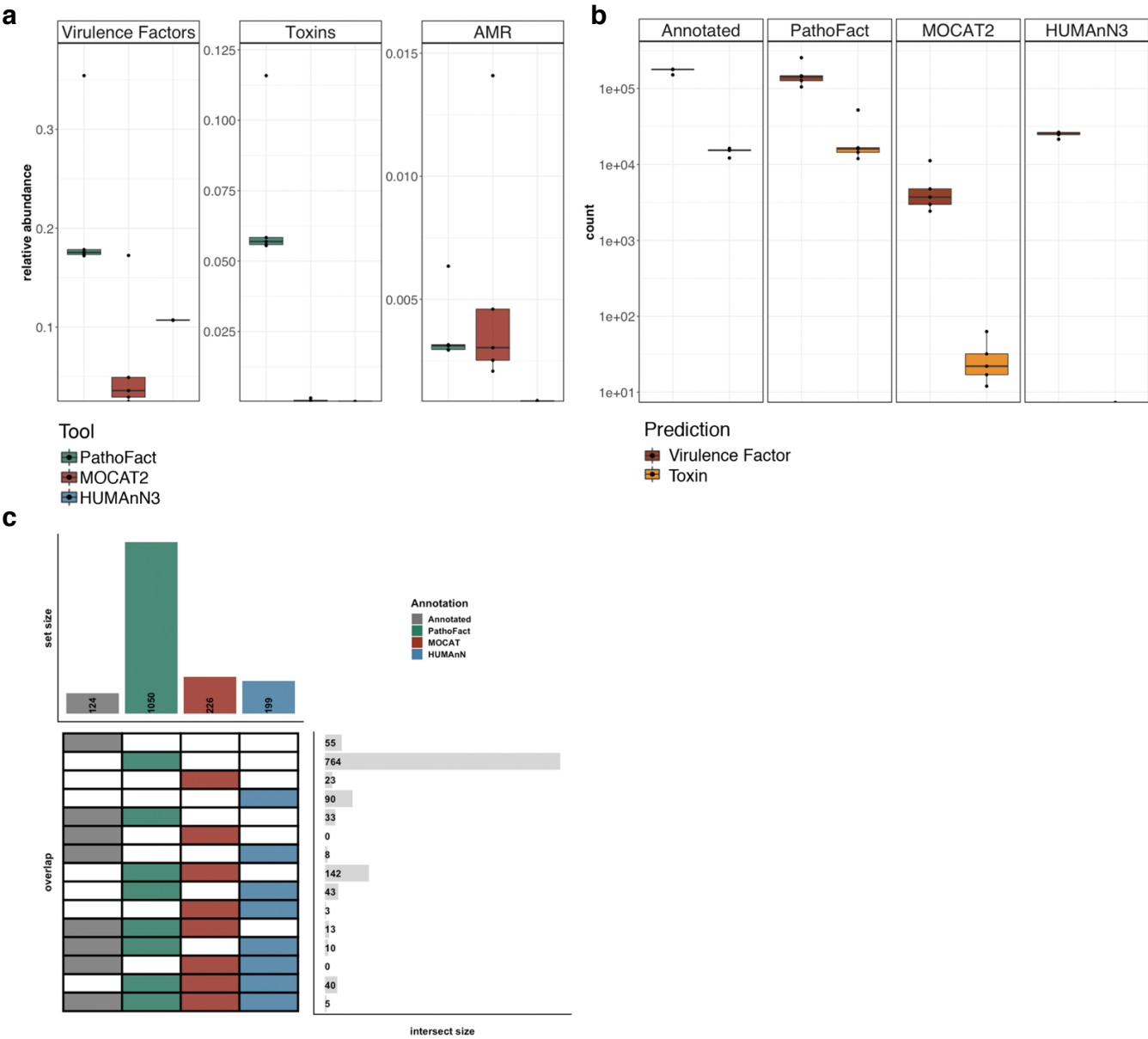


Figure 2.3: Performance evaluation of PathoFact on a high-complexity simulated dataset. **a.** The relative abundances (%) of virulence factors, including bacterial toxins, as well as antimicrobial resistance genes, as predicted by PathoFact, MOCAT2, and HUMAnN3, *two-way ANNOVA, *P* value < 0.05. **b.** Total number of virulence factors and toxin genes identified in the annotated genome and as predicted by PathoFact, MOCAT2, HUMAnN3. **c.** Number of unique ARGs as annotated by the NCBI resistance database and as predicted by PathoFact, MOCAT2, and HUMAnN3.

Virulence factors and toxins may contribute to dysbiosis of the microbiome and favor a pro-inflammatory environment [65]. In addition, particular pathogenic bacteria may adapt to, and survive in, the presence of antimicrobials through acquisition or expression of AMR. Thereby, virulence factors, toxins and AMR may all contribute to the pathogenic potential of the microbiome, which in turn may have an effect on the onset and development of disease and infection. The performance of PathoFact was demonstrated using three publicly available case-control metagenomic datasets which were chosen considering the following criteria: representing an actual infection or a chronic disease in which either pathogenic potential or toxins are believed to play a role. The Milani *et al.* [228] study represents actual infections with *Clostridioides difficile* (CDI) in the human gut microbiome of five patients along with five healthy controls. Furthermore, skin metagenomes of five psoriasis patients along with five healthy controls from Tett *et al.* [229] were chosen to represent a chronic disease in which a pathogenic potential is believed to have a function. Additionally, from Bedarf *et al.* [230] the metagenomes of fecal microbiomes derived from 10 early stage Parkinson's disease (PD) patients, as well as 10 age-matched controls, was obtained to represent a chronic disease in which bacterial toxins are believed to be involved [230].

2.3.4.1 Prediction of virulence factors and bacterial toxins

The predictions from PathoFact resulted in the identification of virulence factors in all three case-control metagenomic datasets. Furthermore, predicted virulence factors were characterized as secreted and non-secreted through the incorporation of SignalP in the pipeline. No statistically significant (P -value < 0.05 , Wilcoxon rank sum test) different relative abundance of the different virulence factors was found in any of the three studies when comparing diseased state and control (**Figure 2.4**).

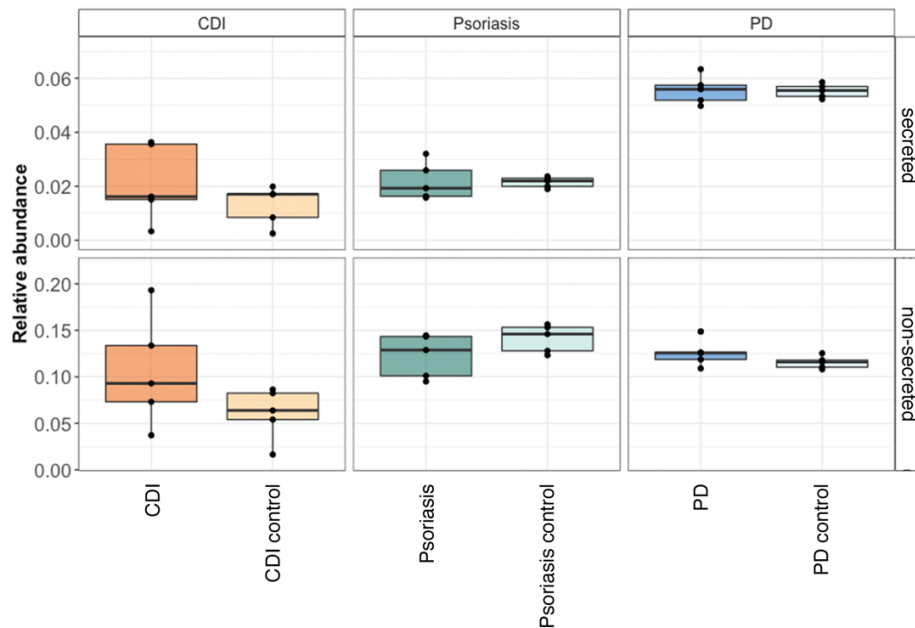


Figure 2.4: Virulence factors in three case-control metagenomic datasets. The relative abundance (%) of both secreted and non-secreted virulence factors as well as non-pathogenic sequences in three metagenomic datasets (*Clostridioides difficile* infection, Psoriasis, Parkinson's disease)

In addition to the general prediction of virulence factors using PathoFact we identified bacterial toxins, as well as their corresponding HMM domain by which they were identified. Furthermore, both secreted and non-secreted toxins were identified in both diseased and control groups in all datasets (**Figure 2.5a**) and we identified several differentially abundant bacterial toxins (**Appendix C.1: Supplementary table 2.3-2.5**). Within the CDI dataset three distinct toxin domains, PF13953, PF13954 and PF06609, were identified to be differentially abundant in CDI over control (**Figure 2.5b**). Interestingly, none of these toxin domains have yet been reported to be linked to CDI and therefore are of interest for further research. Four distinct toxin domains (K12340, PF13935, PF14449 and K11052) were found to be significantly abundant in psoriasis over control (**Figure 2.5c**). Of these toxin domains, only K12340 was previously linked to psoriasis [235]. Finally, regarding the PD study we found several differentially abundant bacterial toxins when comparing PD and control samples (**Figure 2.5d**). Of these bacterial toxins, one containing the PF09599 domains was more abundant in PD and is among others found in invasins in *Salmonella typhimurium* which has been hypothesized to be involved in Parkinson's disease [236]. Interestingly, in all three datasets additional 'unknown'

toxin domains were identified to be linked to the diseases, therefore representing interesting candidates for further research.

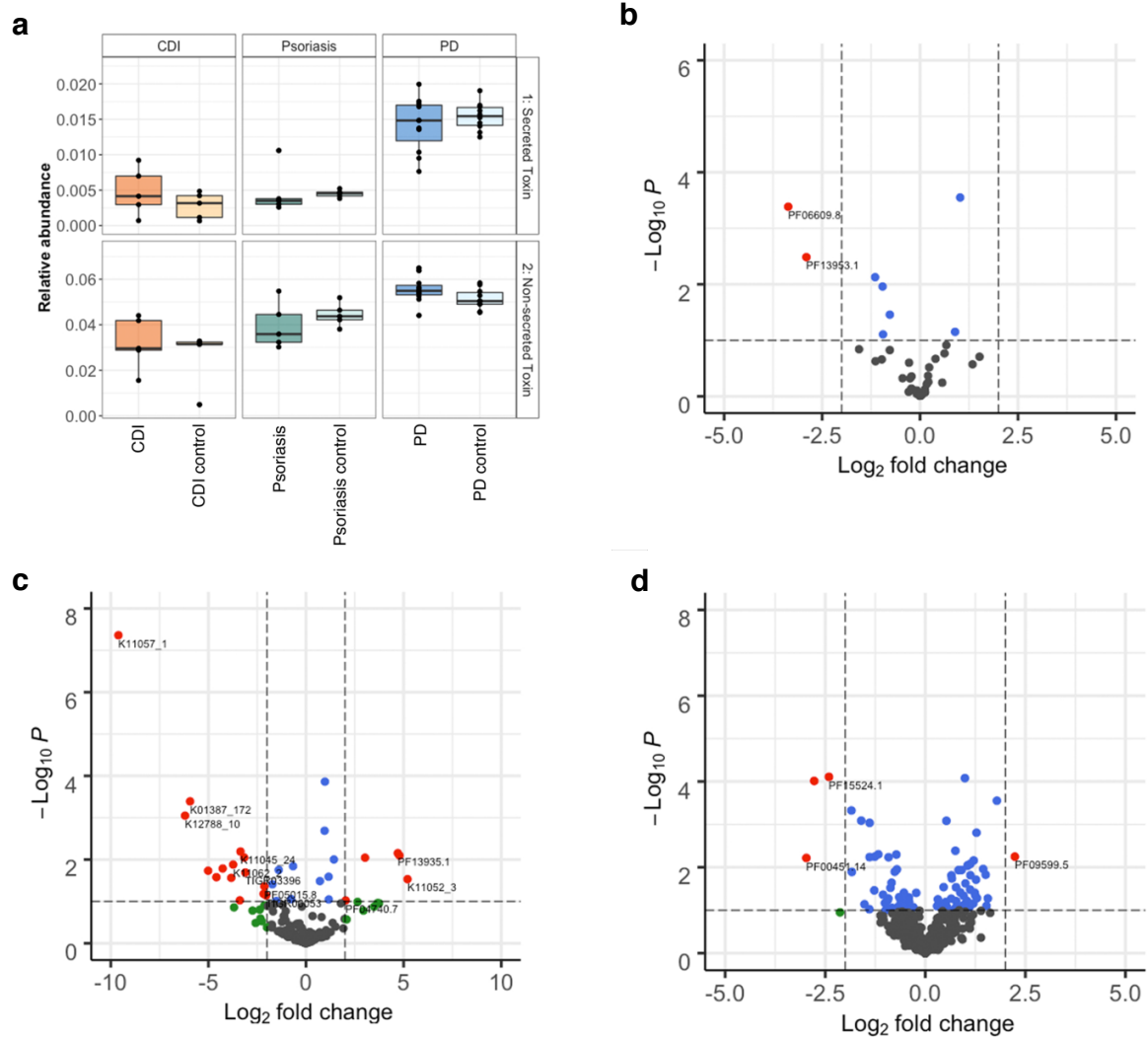


Figure 2.5: Bacterial toxins in three case-control metagenomic datasets. **a.** The relative abundance (%) of both secreted and non-secreted bacterial toxins in diseased versus control subjects. **b.** Volcano plot depicting differentially abundant bacterial toxins in *Clostridioides difficile* infections versus control. **c.** Volcano plot depicting differentially abundant bacterial toxins in Psoriasis versus control. **d.** Volcano plot depicting differentially abundant bacterial toxins in Parkinson's disease versus control.

2.3.4.2 Prediction of antimicrobial resistance

Using the PathoFact pipeline we predicted the presence of antimicrobial resistance genes in all three case-control metagenomic datasets. Within the CDI datasets 23 ARG categories were identified (**Appendix B.1: Supplementary figure 2.2a**) of which six, i.e. diaminopyrimidine, elfamycin, fluoroquinolone, nucleoside, peptide and multidrug, were significantly higher

abundant in individuals with CDI over control (**Figure 2.6a**). Antimicrobial resistance has previously been found to be associated with CDI infections [237]. In the metagenomic data of the skin microbiome 22 categories of ARGs were identified (**Appendix B.1: Supplementary figure 2.2b**). Interestingly, none of these resistance categories were found to be significantly different, neither with the diseased nor the control group. Within the PD study 33 ARG categories were identified (**Appendix B.1: Supplementary figure 2.2c**) with glycopeptide resistance significantly abundant in PD over controls, while tetracycline resistance was found to be enriched in the control group (**Figure 2.6b**). The link between antimicrobial resistance and Parkinson's disease has been mostly unexplored thus far. However, a recently published study by Mertsalmi *et al.* [238] suggests a role for antibiotics in PD through the influence on the gut microbiome.

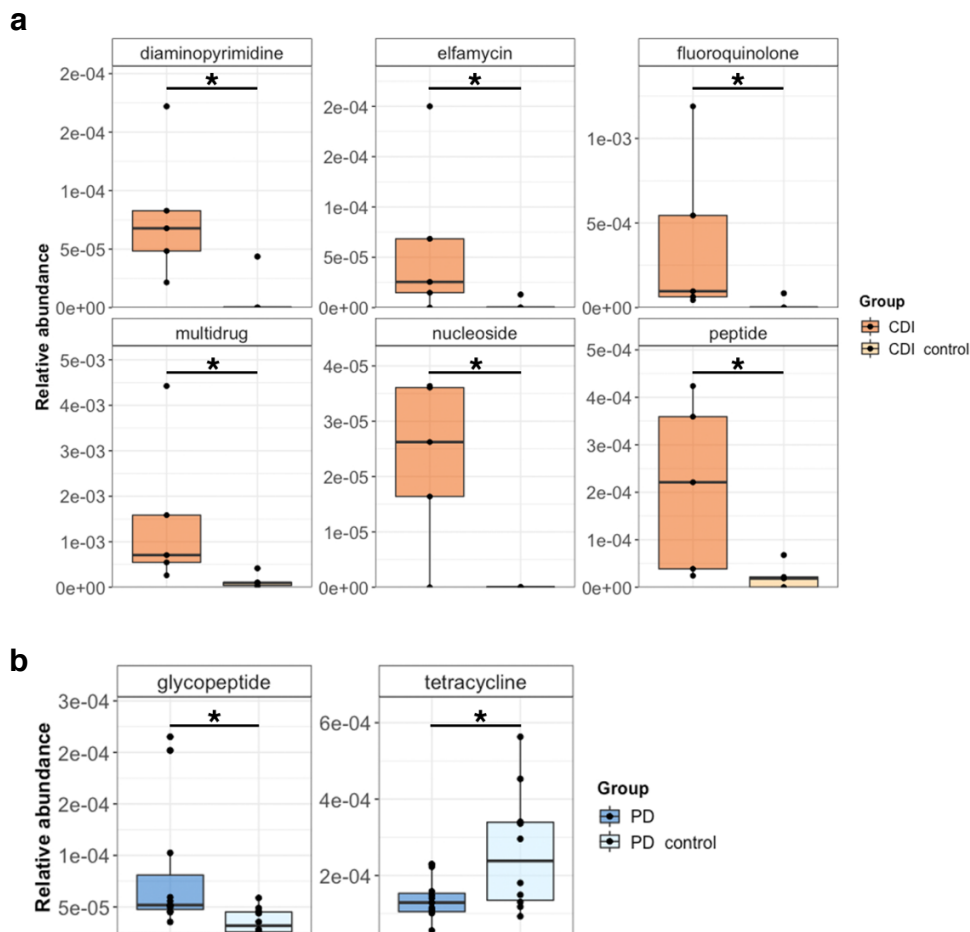


Figure 2.6: Antimicrobial resistance in three case-control metagenomic datasets. The relative abundance (%) of antimicrobial resistance categories with statistically significantly differential abundance in **a** *Clostridioides difficile* infection versus control, **b** Parkinson's disease versus control. * P value < 0.05 .

Although we propose the primary usage of PathoFact for metagenomic analyses, as seen with these three case-control metagenomic datasets, it can also be applied to single genome assemblies. Using the *Klebsiella pneumoniae* subsp. *pneumoniae* HS11286 reference genome, we identified 86 resistance genes of which 6 contained SNPs contributing to resistance (**Appendix C.1: Supplementary table 2.6**).

2.3.4.3 Prediction of mobile genetic elements linked to virulence factors

Using the predictions generated by PathoFact, we resolved the genomic contexts and identified MGEs in all three case-control metagenomic datasets (**Figure 2.7a**) (**Appendix B.1: Supplementary figure 2.3**). Within all three datasets the presence of both phage- and plasmid-derived sequences was detected, although no significant difference was observed between diseased and control. We found that in all datasets the majority of MGEs were found to be both linked to virulence factors as well as AMR (~50%), closely followed by MGEs linked solely to virulence factors, including bacterial toxins, with AMR contributing to the remaining MGEs (**Figure 2.7b**). Furthermore, a number of MGEs were found to be both linked to virulence factors as well as AMR.

Of the ARGs linked to MGEs, the prevalence of the different resistance categories were identified using our tool. Within the CDI dataset, the majority of the MGEs were linked to phenicol and beta- resistance in both diseased and control groups (**Appendix B.1: Supplementary figure 2.4a**). Additionally, plasmids linked to diaminopyrimidine and sulfonamide resistance were identified within the disease group while found to be absent in the control. Within the skin metagenomes, the majority of the predicted resistance genes linked to MGEs included beta-lactam, tetracycline and multidrug resistance in both diseased and control groups (**Appendix B.1: Supplementary figure 2.4b**). However, MGEs linked to beta-lactam resistance were found to be enriched in the diseased group. Finally, of the resistance genes within the PD study, both peptide and tetracycline resistances were found to be linked to phage and plasmids. Peptide resistance was abundant in controls whereas tetracycline was identified primarily in diseased (**Appendix B.1: Supplementary figure 2.4c**).

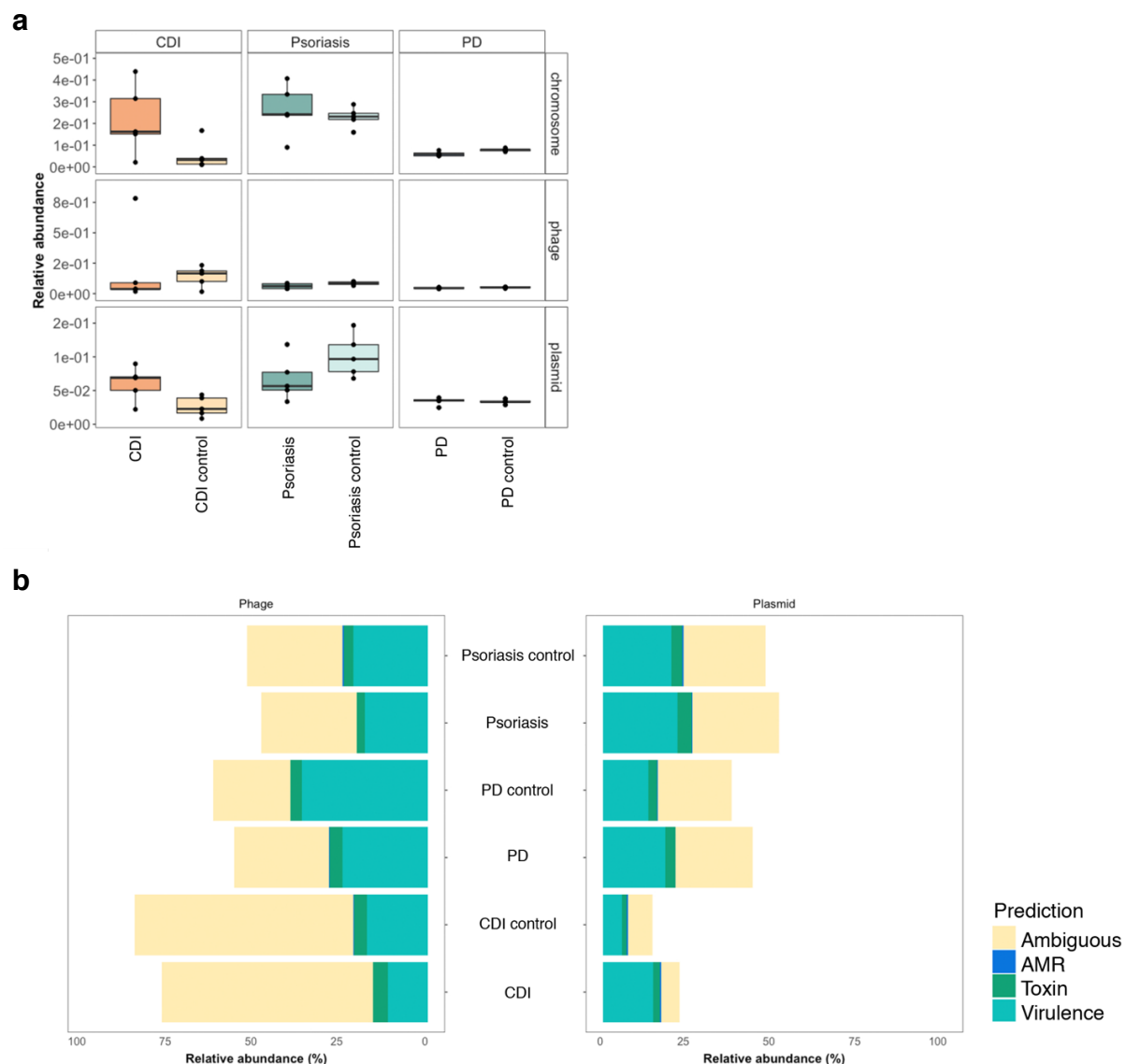


Figure 2.7: Identification of MGEs within three case-control metagenomic datasets. Relative abundance of MGEs within three metagenomic datasets (*Clostridioides difficile* infection, psoriasis (skin), and PD). **a.** The overall relative abundance of phage and plasmid within the *Clostridioides difficile* infection, psoriasis, and Parkinson's disease datasets. **b.** The distribution of virulence factors, incl. toxins, and AMR between phage and plasmid in all datasets.

2.4 Conclusion and outlook

The identification of virulence factors, toxins and antimicrobial resistance genes are of immediate importance for understanding the pathogenic state of microbiomes. Using our newly developed tool, PathoFact, we were able to identify virulence factors and bacterial toxins within

three publicly available case-control metagenomic datasets. Furthermore, we were able to identify differentially abundant bacterial toxins when comparing diseased and control groups in all datasets. Additionally, antimicrobial resistance genes were identified in two of the datasets with a significant difference of certain resistance categories between diseased and control individuals. The inclusion of MGEs is of particular importance in understanding the possible transmission of MGE-born virulence factors. With PathoFact we identified MGEs in all three datasets and were able to link these simultaneously to the corresponding virulence factors, toxins, and antimicrobial resistance genes.

Until now, no single tool has existed which has combined these distinct aspects. Although several prediction tools exist for AMR, of which DeepARG and RGI have been chosen for their accuracy and ability to identify mutations' contribution to resistance (RGI), to be included in our pipeline. Limited or no tools were available on the other hand for the prediction of toxins and virulence factors. PathoFact utilizes the wealth of currently available software (e.g., AMR and MGE predictions) as well as newly generated tools (e.g. virulence factors and toxins). Furthermore, PathoFact can conveniently integrate updates and newly developed prediction tools. In conclusion, our tool combines the strength of AMR predictions linked to MGE predictions and integrates this with the prediction of toxins and virulence factors. PathoFact is a versatile and reproducible pipeline by its ability to run either the complete workflow or each module on its own, giving the investigator flexibility in their analysis.

Chapter 3. Persistence of birth mode-dependent effects on gut microbiome composition and antimicrobial resistance during the first year of life.

A major part of this chapter was adapted and modified from the following first-author peer-reviewed publication:

Susheel Bhanu Busi*, **Laura de Nies***, Janine Habier, Linda Wampach, Joelle V Fritz, Anna Heintz-Buschart, Patrick May, Rashi Halder, Carine de Beaufort and Paul Wilmes (2021). Persistence of birth mode-dependent effects on gut microbiome composition, immune system stimulation and antimicrobial resistance during the first year of life. *ISME Communications* 1 (1), 1-12 [**Appendix A.3**]

* Co-first author

3.1 Background

The rate of caesarean section delivery is constantly increasing worldwide, which is partly driven by increases in overall income and access to health facilities [239]. According to a 2015 report, 29.7 million births occurred via CSD in that year accounting for ~18% of the births in 169 countries [239]. At 25% in Europe, this number is higher than the global average [240]. The short-term risks of CSD include delayed or altered development of the immune system [241], reduced gut microbiome diversity⁴, limited transmission of bacterial strains from mother to neonate [242,243], and microbiome-borne functional deficiencies [165,244–246]. Although few studies associate CSD with metabolic disorders [247,248] and allergies [249,250], the long-term effects of birth mode are not well understood. Shao *et al.* reported that CSD may predispose individuals to colonization by opportunistic pathogens including those carrying antimicrobial resistance (AMR) genes [251]. On the one hand, several reports including our previously published study [165] addressed questions concerning the very early development of the neonate's gut microbiomes [250,252] and immune system priming [241] in relation to disease development [253,254]. On the other hand, only few reports [255–258] follow the effects of birth mode during the first year of life especially in relation to immune system priming, development and evolution of AMR, and the contribution of mobile genetic elements to the persistence of ARGs.

Factors including environmental exposure [250], breastfeeding and diet [255,259,260], and genetics [261] play crucial roles in the development of an infant. Aside from this, it is now generally accepted that birth mode, i.e. vaginal delivery (VD) or CSD, has a pronounced impact on early microbiome structure [165,241,247,262,263]. While the majority of these studies focus on the overall microbiome structure, analyses of the functional contribution of the microbiome have attracted attention due to its sensitivity to perturbation [264]. For example, we previously reported that the microbiome in VD-born babies was enriched in bacterial genes encoding for lipopolysaccharide (LPS) biosynthesis, cationic antimicrobial peptide resistance as well as two-component systems [165]. Interestingly, higher levels of LPS biosynthesis genes were associated with increased immune responses in VD neonates, whereas CSD neonates had reduced levels of TNF- α and IL-18 immediately after birth. Noteworthy in this context is previous work by Vatanen *et al.* which showed that differing LPS immunogenicity contributes to autoimmunity thereby affecting the long-term health outcomes of infants exposed to different antigens [265]. Furthermore, others have hypothesized and reported [241,266] a similar phenomenon, whereby the gut microbiome contributes to the development of the immune

system during a “critical window” of development [241,266–272]. In a neonatal cohort at risk for the development of asthma, bacterial metabolites were shown to specifically impede immune tolerance [273]. However, some of the reports described above do not elaborate on the continuous effect of early immune system priming in the context of the birth mode and especially over the course of the first year of life, including whether these effects normalize over time.

Aside from the well-studied factors and consequences of development described above, the role of commensal microbiota in the emergence and spread of AMR is not well understood. Recent studies have reported that antibiotic exposure in infancy affects microbial diversity and enriches for ARGs. Interestingly, Ravi *et al.* have suggested that the infant gut microbiome acts as a reservoir for multidrug resistance that persists throughout infancy up to two years of age [274]. They reported that integrons (*int1* gene) in the gut could potentially be responsible for this phenomenon. Nevertheless, the effect of birth mode, CSD or VD, on the transmission and occurrence of AMR remains unresolved.

Here, we address the aforementioned gaps in knowledge concerning the effect of birth mode on the persistence of the gut microbiota over the first year of life including their inherent functions, immunogenic properties and their role in conferring AMR. Our results highlight birth mode-dependent differences in gut microbiome structure and their association with immune function. We found that the gut microbiota becomes similar between CSD and VD babies at one year of age, with the exception of an immunostimulatory commensal, *Faecalibacterium prausnitzii*, which was enriched in the VD group. Additionally, we identified an increased abundance in ARGs directed against synthetic and semi-synthetic antibiotics in CSD as early as five days *postpartum*. Strikingly, we found that mobile genetic elements (MGEs) including plasmids and bacteriophages are key contributors to the establishment and persistence of AMR, irrespective of birth mode. Collectively, our findings suggest that birth mode-dependent effects persist through the first year of life including the delayed immunostimulation of CSD infants likely affecting tolerance mechanisms as well as the apparent role of bacteriophages in conferring AMR.

3.2 Methods

3.2.1 Ethics statement

All aspects concerning the recruitment and collection of mother-neonate pairs including handling, processing and storing of samples as well as data were approved by the Luxembourg Comité national d'éthique de recherche, under reference number 201110/06 and by the Luxembourg National Commission for Data Protection under reference number A005335/R000058. Prior to specimen collection, following a detailed consultation; written and informed consent was obtained from all mothers enrolled in the study.

3.2.2 Sample collection

Based on our previous study [165], the present study design aimed at testing the hypothesis that birth mode elicits longer-term functional microbiome changes which may impact neonatal health and development (with particular foci on antimicrobial resistance and lipopolysaccharide biosynthesis) and we performed the corresponding power analyses using data from our previous study. Founded on the increase in fold-change [caesarean section delivery (CSD) versus vaginal delivery (VD)] in antimicrobial resistance genes, a sample size calculation revealed a minimum number of four individual mother-infant pairs per group to achieve a power of 80% with a significance threshold of 5%. For the LPS-mediated functional cytokine measurements, we estimated a minimum sample size per group of six pairs based on a fold-change of 1.40 x in TNF- α , *i.e.*, a 40% difference of means between the samples (**Appendix B.2: Supplementary figure 3.1**). As previously published [165,275], we found that the functional microbiome differences provide clearer delineations when comparing groups than the typically reported taxonomic profiles. Based on our hypothesis, we focused on functional endpoints, in particular on the emergence and acquisition of ARGs and the LPS-mediated immune stimulation. As per the results of the power analyses highlighting a minimum requirement of 6 mother-infant pairs per group, we further inflated the per-group sample size by 50 % leading to a minimum of 9 mother-infant pairs per group. In the present study, babies delivered via caesarean (CSD, n=11) and vaginal (VD, n=9) deliveries were sampled during the first days of life and were followed-up at 1 month, 6 months and at one year of age. Samples were collected during follow-up visits into sterile plastic vials and immediately flash-frozen in liquid nitrogen. Faecal samples were stored until further processing at -80 °C.

3.2.3 Faecal processing and nucleic acid extraction

Genomic DNA was isolated from 50 mg of frozen stool samples aseptically weighed into sterile vials, prior to processing with the DNeasy PowerSoil Kit (Qiagen, Luxembourg) including an additional incubation step at 65 °C and milling, as described previously [165]. All the study samples yielded sufficient DNA for metagenomic sequencing including artefact-curated metagenomic data as described previously [165] for subsequent analyses. DNA extracted from all timepoints was thereafter stored at -80 °C until further use.

3.2.4 DNA sequencing

All DNA samples were subjected to random shotgun sequencing. Briefly, 250ng of DNA was sheared using Bioruptor NGS (Diagenode, UCD300) with 30s ON and 30s OFF for 15 cycles. The sequencing libraries were prepared using TruSeq Nano DNA library preparation kit (Illumina, FC-121-4002) using the protocol provided with the kit. The libraries were prepared considering 350bp average insert size. Prepared libraries were quantified using Qubit (Invitrogen) and the quality was checked on a Bioanalyzer (Agilent). Sequencing was performed on the NextSeq500 (Illumina) instrument using 2x150 bp read length at the LCSB Sequencing Platform.

3.2.5 Data processing for metagenomics, including genome reconstruction

Paired forward and reverse sequences were processed using the metagenomic workflow of the Integrated Meta-omic Pipeline [231] (IMP). The metagenomic processing workflow includes pre-processing, assembly, genome reconstruction and functional annotation of genes based on custom databases in a reproducible manner. Briefly, the adapter sequences were trimmed in the pre-processing step including the removal of human reads. Thereafter the *de novo* assembly was performed using the MEGAHIT (version 2.0) assembler[276]. Default IMP parameters were retained for all samples. Subsequently, we used MetaBAT2 [277] and MaxBin2 [278] for binning in addition to an in-house binning methodology previously described [211]. This involved ignoring ribosomal RNA sequences in kmer profiles based on the clustering from VizBin embeddings [279], which uses density-based non-hierarchical clustering algorithms and depth of coverage for genome reconstructions. The reconstructed genomes are hereafter referred to as bins or metagenome-assembled genomes (MAGs). We obtained a non-redundant set of MAGs using DASTool [280] with a score threshold of 0.7 for downstream analyses.

3.2.6 Metagenomic taxonomic classification, virome and functional analyses

Trimmed and pre-processed read pairs were used as input to determine the microbial abundance and population genomic profiles based on the mOTUs [281] (version 2) tool. Based on the marker genes in the mOTU2 database taxonomic profiling was performed. The relative abundances of the mOTUs were estimated using a minimum alignment length of 125 base pairs (bp), where the read counts were normalized to the gene length while also accounting for base coverage of the genes. This was done using the *motus profile* option with the built-in option (-c) for relative abundance values per sample. Simultaneously, to improve specificity and minimize false positives, a cut-off of seven genes that deviated from the median was used as an additional parameter to improve both sensitivity and precision. For the reconstructed MAGs, completeness and contamination was determined using CheckM [282], while the taxonomy for each MAG was assigned using the GTDB (Genome Taxonomy Database) toolkit (gtdb-tk) [283] using the *lineage_wf* option and by using the fasta files as inputs for the MAGs.

For the analyses of functional potential from the assembled contigs, open-reading frames were predicted from the assembled contigs using a modified version of Prokka [284] that includes Prodigal [204] gene predictions for complete and incomplete open reading frames. The identified genes were annotated with a Hidden Markov Models [285] (HMM) approach, trained using an in-house database [211] including all KO [214], TIGRFAM and SWISS-PROT [216] groups and using *hmmsearch* from HMMER 3.1 [205]. Where multiple functional groups were assigned to genes, the best hits based on bit scores were selected. FeatureCounts [232] was used to extract the number of reads per functional category, using the arguments -p and -O, thus yielding counts for each functional category. After the LPS-cytokine analysis, insufficient faecal sample for one of the CSD samples (C118) remained for metagenomic sequencing. Therefore, the sample was removed from subsequent metagenomic analyses. For the virome analyses, we used an iterative annotation method to recover microbial (bacterial and archaeal) viruses [286], and subsequently taxonomically annotated using a network-based classification protocol defined by Bolduc *et al.* [287]. Samples C109 was not included in the virome analyses due to viral contigs being below detection confidence thresholds.

3.2.7 Identification of antimicrobial resistance genes and association with mobile genetic elements

We used a deep-learning approach, DeepARG [42], to predict and identify ARGs within our metagenomic data. The output from Prokka, i.e., the translated fasta sequence files for all open

reading frames, was used as input for the AMR analyses. ARGs were collapsed into categories based on the Comprehensive Antibiotic Resistance Database (CARD) [43] and identified using DeepARG. Thereafter, the relative abundance of the ARGs was calculated using the Rnum_Gi method described by Hu *et al.* [233].

Identified ARGs and their categories were consecutively linked to associated bacterial taxonomy using the metagenomic bin classification. Furthermore, ARGs were linked to predicted mobile genetic elements (MGEs; phages and plasmids) to identify probable transmission of AMR between taxa. For the identification of plasmids in the metagenomic data, PlasFlow [47] was used with a threshold for filtering set to 0.7. Simultaneously, DeepVirFinder [50] and VirSorter [51] were used to identify phage sequences within the VD and CSD groups. Predictions from both these tools were subsequently merged to obtain a comprehensive catalog of phage sequences. For the prediction of phage sequences the DeepVirFinder thresholds for filtering were set at a p-value of <0.05 and a score of 0.7, while for VirSorter the category 1 and 2 predictions were used for downstream analyses. To link both the MGEs and the taxonomy to the ARGs, we mapped the genes to assembled contigs, followed by identifying the corresponding bins (MAGs) to which the contigs belonged. By considering all different predictions of MGEs, a final classification was made based on the genomic contexts of the ARGs encoded on plasmids, phages or chromosomes, including classification of those that could not be resolved (ambiguous). Those ARGs that could not be assigned to either the MGEs or bacterial chromosomes were further referred to as unclassified genomic signatures. Certain ARGs were encoded on both the bacterial chromosome and phage genomes. In such cases, we recorded the encoded ARG as being ambiguous. The confirmation of ARGs and their associated mode of transfer was performed manually alongside the mapping of identical 1Kbp flanking regions, via the MetaCHIP analyses pipeline [288]. Briefly, groups of genes among all input MAGs with maximum average identity were considered putative HGT genes. To validate the predicted candidates, a pairwise BLASTN was used to assess each pair of flanking regions of 10 Kbp. Visual representations of the genomic regions were extracted alongside the results for visual interpretation and inspection. Coverage of the genomic regions was additionally assessed through the IGV viewer using the bam file, and manually plotted based on per base coverage statistics for the latter.

3.2.8 LPS isolation and *in vitro* immunostimulation for cytokine profiling

From the one year of age time point, 150 mg faecal samples were weighed aseptically, and lipopolysaccharide (LPS) was extracted alongside an extraction blank to serve as a negative control. We also used an in-house pure culture of *E. coli*, from which extracted LPS was used as a positive control. To maximize yields, the samples were divided into triplicates, i.e., 50 mg per vial, prior to LPS extraction using the hot phenol-water protocol as previously described [165]. After extractions the triplicates from each sample were pooled and quantified using an endotoxin-detection assay (Endolisa, #609033, Hyglos GmbH, Germany). All samples produced sufficient quantities of LPS. The purified LPS was used to stimulate monocyte-derived dendritic cells (MoDCs). Briefly, primary human monocytes were derived from blood samples from four healthy donors obtained through the Luxembourg Red Cross. The monocytes were further differentiated into MoDCs, in RPMI 1640 medium (ThermoFisher Scientific) supplemented with 10% foetal bovine serum (ThermoFisher Scientific), 20 ng ml⁻¹ of granulocyte-macrophage colony-stimulating factor (Peprotech, London, UK), 20 ng ml⁻¹ IL-4 (Peprotech) and 1% penicillin–streptomycin (Invitrogen). Subsequently, the immunostimulatory potential of the LPS fractions isolated from the one year of age faecal samples was determined. For this, MoDCs were treated with LPS extracts from VD and CSD samples. The amount of LPS from each sample that was used to stimulate the MoDCs was adjusted as described by Wampach *et al.* [165]. Briefly, the MoDCs were stimulated with 7.5 µl/well of LPS while a positive control was established using 15 EU/well LPS isolated from *E. coli*, and a negative control was set up by incubating MoDCs with 7.5 µl/well of the LPS extraction blank. For the *in vitro* stimulation, the amount of MoDCs was 1 x 10⁵ cells/well. Treatments were performed on cells from all the healthy donor-derived samples and analyzed for the presence of pro- and anti-inflammatory cytokines (TNF-α, IL-8, IL-18, IL-1b, IL-12, and IL-10) using both Human Instant and uncoated ELISA kits (ThermoFisher Scientific).

3.2.9 Data analysis

All figures for the study including visualizations derived from the taxonomic, functional, and cytokine profiling were created using version 3.6 of the R statistical software package [289]. DESeq2 [234] and Wilcoxon rank-sum tests with FDR-adjustments for multiple testing were used to assess significant differences for the AMR and taxonomic analyses whereas a paired two-way ANOVA (Analysis of Variance) within the *nlme* package was used for identifying statistically significant differences in the cytokine profiles. Volcano plots were generated using the *EnhancedVolcano* package [290]. Corrpplots were generate using the *corrgram* package

developed for R [291]. The *metacoder* [292] package was used to visualize the AMR-linked taxonomy in R.

3.3 Results

3.3.1 Birth mode-dependent gut microbiota differences during the first year

We previously described the initial seeding and colonization processes within the human gut microbiome and identified differences in microbiome structure and function as well as linked immunogenicity and immune system priming, which stratified according to birth mode [165,275]. Building on this work, we aimed to understand the long-term effects in relation to the observed differences, especially through the first year of life which represents a “critical window” of development including physiological growth and immune system maturation. To achieve this, we followed VD and CSD neonates in our cohort and collected faecal samples at crucial intervals after birth, including five days, 1 month, 6 months, and at one year of age (**Figure 3.1**).

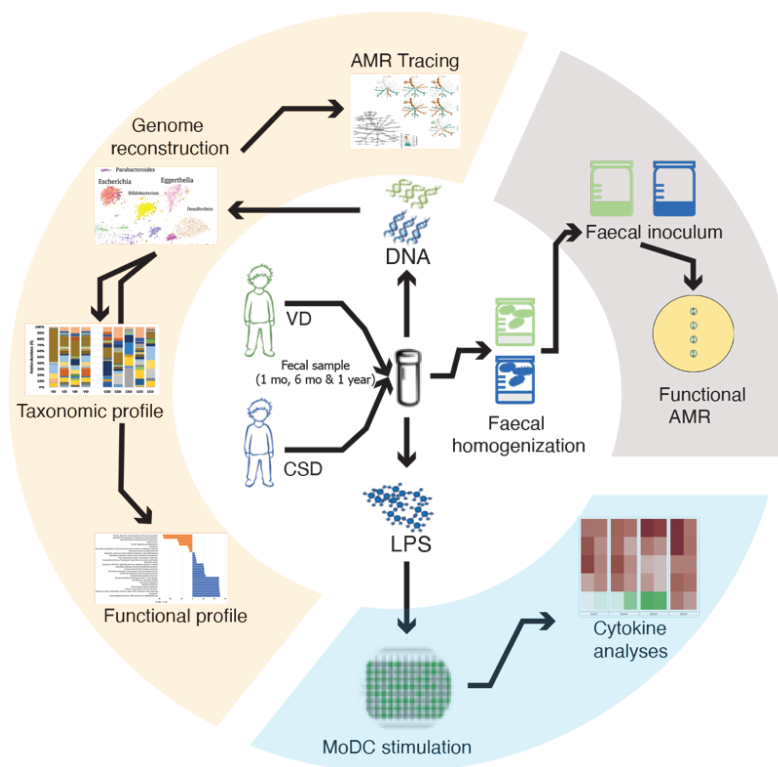


Figure 3.1: Workflow representation of DNA and LPS isolation from faecal samples for metagenomic, immune and functional AMR analyses

In one of our previous studies [165], a multivariate analysis was performed to compare the profiles of CSD (\pm SGA) to VD neonates. The results of these analyses demonstrated that delivery mode was the strongest determining factor in the microbial profile and predicted functions irrespective if the infants were born SGA (small for gestational age) or not [165]. In light of these analyses, we included the SGA samples within the CSD group. We reconstructed microbial genomes and identified differentially abundant taxa and functions between the groups using metagenomic sequencing data. Based on metagenomic operational taxonomic units (mOTUs), we calculated the Jensen-Shannon divergence index and found that the intra-group variability within CSD or VD was minimal while the inter-group variability between CSD and VD groups was significantly different (**Appendix B.2: Supplementary figure 3.1**). At the genus level, our data also recapitulated previously described [165] significantly increased levels of *Bacteroides* (FDR-adjusted $p < 0.05$, Wilcoxon rank-sum test) in the VD neonates compared to CSD at the early time points (day 5 after birth and at 1 month). *B. caccae* also showed an increasing trend in the CSD group at 6 months and after one year of age, while *B. caecimuris* was significantly increased in CSD at 5 days after birth (**Figure 3.2a**; FDR-adjusted $p < 0.05$, Wilcoxon rank-sum test) and showed an increasing trend at one year of age. However, at one year of age, the abundance of this genus in samples from CSD neonates was comparable to the levels in the VD group. In contrast, the levels of *Bifidobacterium* were increased in VD after 6 months, while *Faecalibacterium prausnitzii*, a commensal associated with healthy human microbiomes [293], was found to be significantly increased in the VD group at one year of age (**Figures 3.2a**, FDR-adjusted $p < 0.05$, Wilcoxon rank-sum test). We further found that both birth mode and the neonatal age affect the gut microbiome community structure, whereby the latter contributes highly to variation within and between the groups (**Figure 3.2b**). The taxonomic profiles at one year of age were distinct when compared to day 5, 1 month and 6 months from both groups.

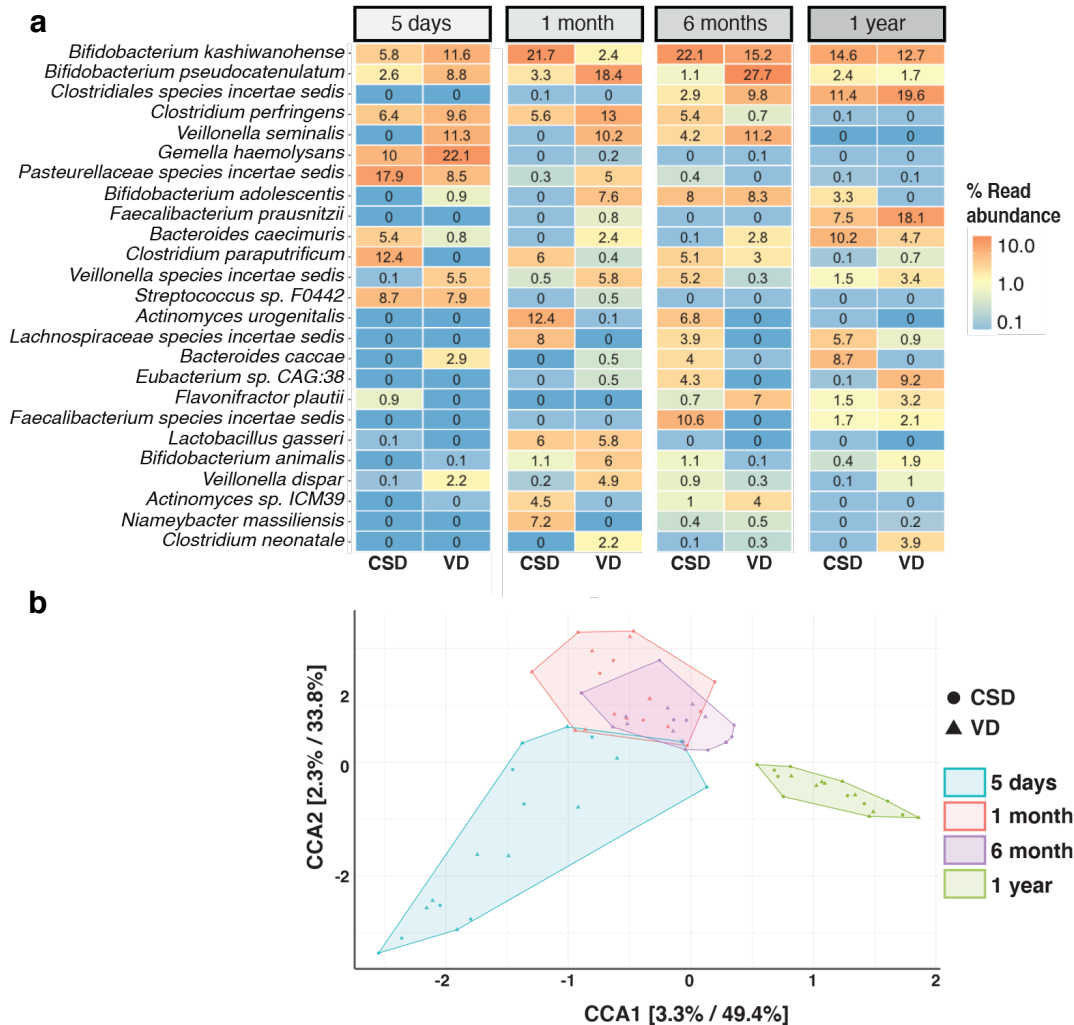


Figure 3.2 Gut microbiome profiles throughout the first year of life. a. Relative abundances of metagenomic operation taxonomic units (mOTUs) >1% abundance at day 5 after birth, 1 months, and at 1 year of age. **b.** Canonical correlation analyses (CCA) resolving the stratification of taxonomic profiles based on two covariates, i.e., birth mode and time when samples were sequenced.

3.3.2 Assessment of differences in metagenomic functional potential at one year of age

Taxonomic differences within the gut microbiome populations may not always manifest as differences in functional diversity due to the redundancy in the latter. To address this, we assigned KEGG [214] orthology identifiers (KOs) to each gene identified from both groups. We found 84 differentially abundant KOs between VD and CSD samples at one year of age (**Figure 3.3a**). Additionally, we linked all identified KOs (n=7,103) to their corresponding KEGG orthology pathways (**Figure 3.3b**) and performed differential pathway analyses. We found that the VD group showed an increase in the gene copy numbers of pathways involved in

carbapenem and phenazine biosynthesis (**Figure 3.3c**). We found that twenty-one unique genera were associated with carbapenem biosynthesis across both groups spanning all major phyla found within the gut.

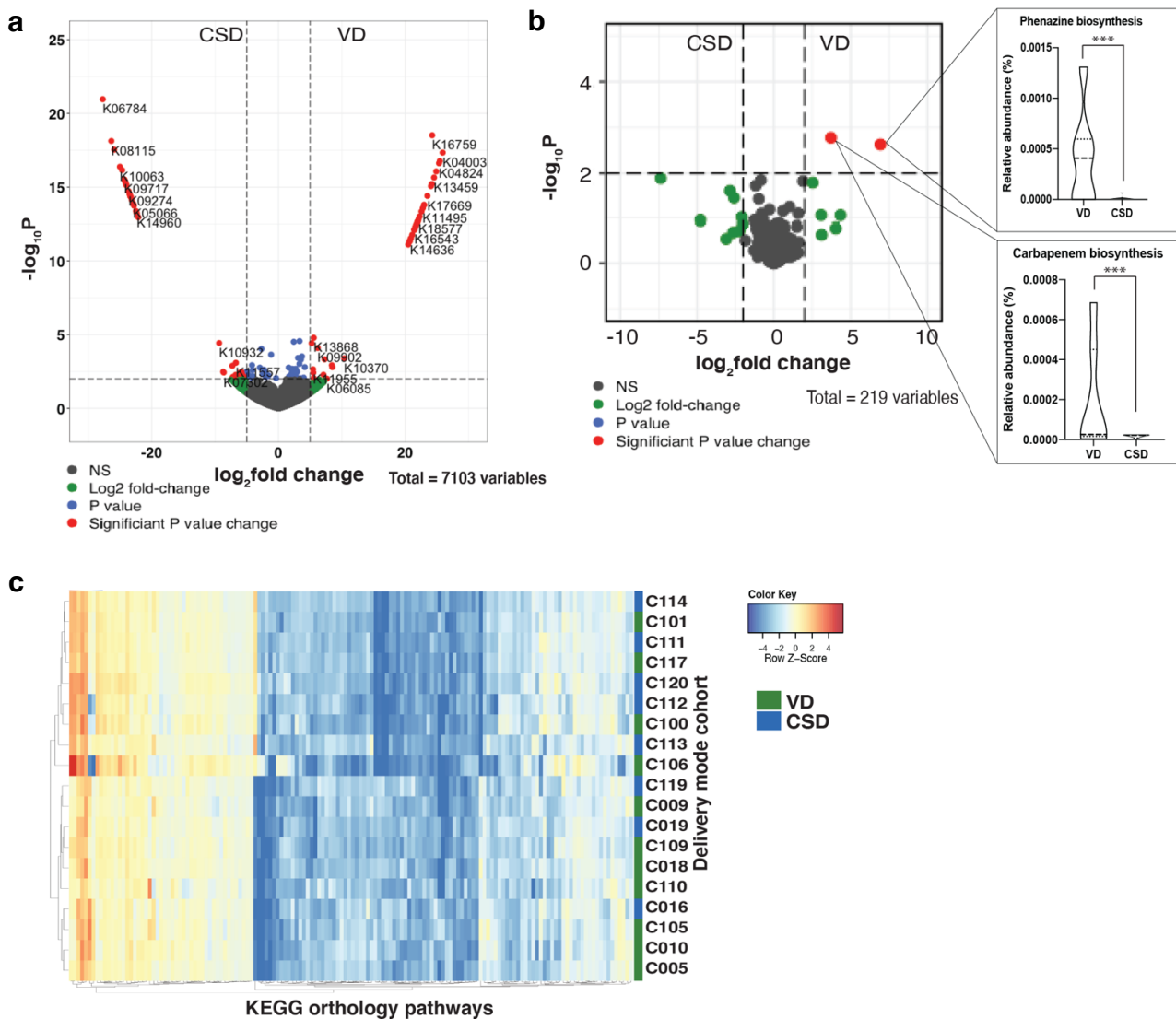


Figure 3.3 Functional differences at 1 year of age. **a.** Volcano plot depicting the statistically significantly different KEGG orthologs found in both CSD and VD groups at 1 year of age. A total of 6413 variables were tested, with $-\log_{10}(p \text{ value})$ shown on the y axis. **b.** Heatmap based on the KEGG pathways found in both CSD and VD samples at 1 year of age. Each column denotes a pathway represented by the KO genes, with the hierarchical clustering being based on Euclidean distances using Ward's clustering algorithm. **c.** Volcano plot of the 219 KEGG pathways to which the KO's were mapped, tested for significance with a fold-change cut-off of 2, and with a false-discovery rate-adjusted p value cut-off of 0.01. *** p value < 0.001

3.3.3 Pro-inflammatory immune responses elevated in CSD after one year of life

In the early stages of neonatal development, we found that the immune activation potential of LPS was significantly increased in samples from VD neonates [165], whereby the isolated LPS triggered the secretion of TNF- α and IL-18 by monocyte-derived dendritic cells (MoDCs) from four healthy adult donors. To determine if the immunostimulatory potential persisted at one year of age, we stimulated the MoDCs (obtained from four healthy adult donors) with LPS isolated from the faecal samples of the CSD and VD groups. In addition to TNF- α and IL-18, we also tested the potential of LPS to stimulate secretion of pro- and anti-inflammatory cytokines such as IL-1 β , IL-12, IL-8, and IL-10 (**Figure 3.4a**). Interestingly, at one year of age, IL-18 was below the detection limits. We did not find any significant differences between the CSD and VD groups with respect to the levels of secreted TNF- α at one year of age. However, contrary to the patterns observed at five days after birth, we found that the levels of TNF- α stimulated by LPS were significantly increased at one year of age within the CSD group ($p < 3.5 \times 10^{-5}$, Paired Two-Way ANOVA; **Figure 3.4b**). Interestingly, the increase in stimulated TNF- α levels in CSD at one year of age was similar to the level of the cytokine stimulated by LPS from the day 5 VD samples (**Figure 3.4b**). Additionally, we found that the stimulated TNF- α levels at one year of age were positively correlated with the abundance of several mOTUs, including *Bacteroides caecimuris* and *Haemophilus influenzae* (**Figure 3.4c**). Previous reports [294] suggest that Enterobacteriaceae levels correlate with inflammatory levels. However, we did not find a correlation of this taxa with LPS levels in our study (**Appendix B.2: Supplementary figure 3.2**). Our data also indicate an increase in the number of Gram negative (G-ve) bacteria at one year of age compared to day 5 after birth in the CSD group (**Appendix B.2: Supplementary figure 3.3**). The increase in Shannon diversity at one year of age compared to day 5 coupled with the increase in G-ve bacteria provides a mechanistic explanation why the LPS stimulation of donor cells from fecal samples of CSD resulted in similar levels of TNF- α (**Figure 3.4b**), as observed with fecal samples from the VD group at one year of age.

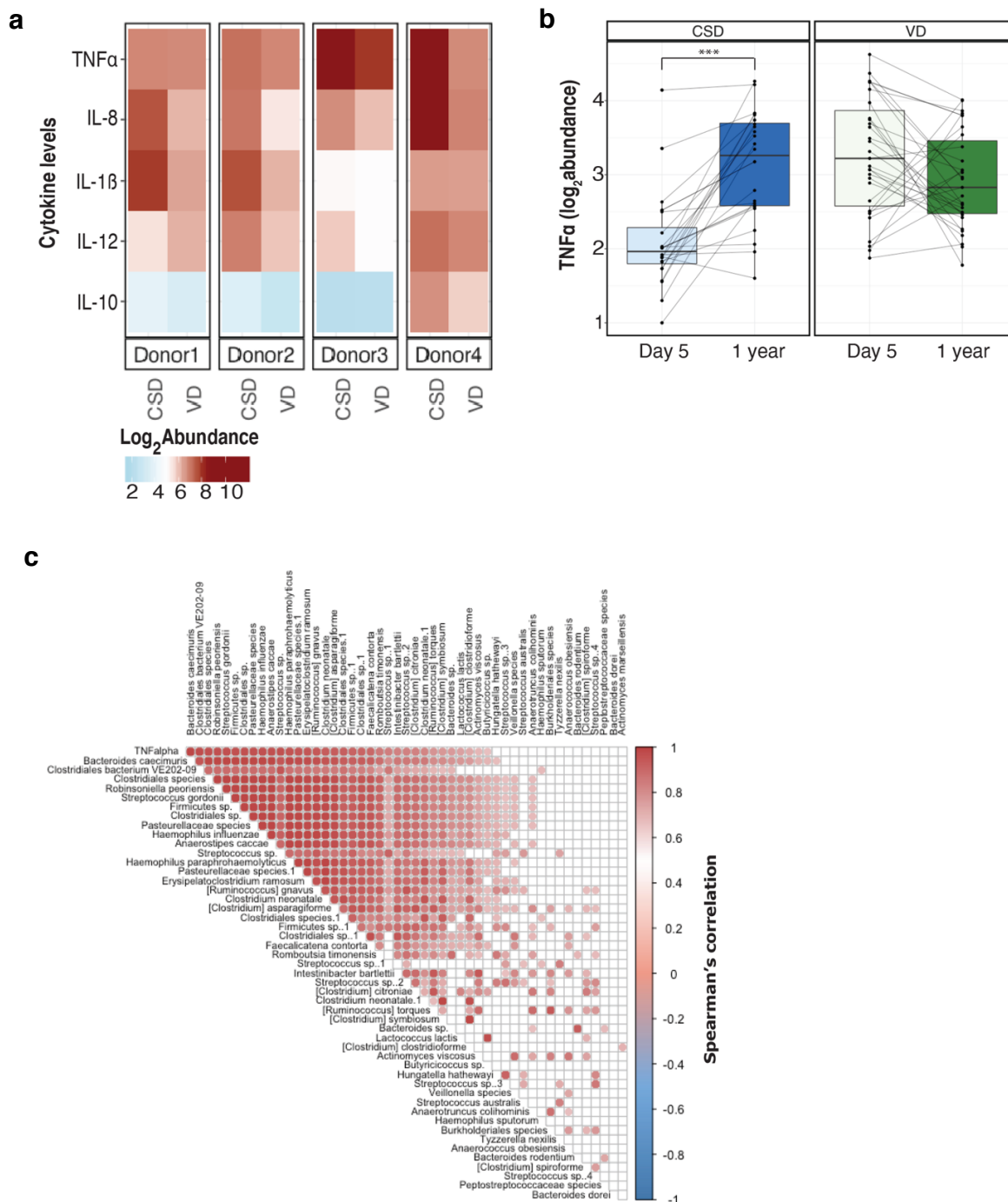


Figure 3.4 Immunostimulatory potential at 1 year of age. Heatmap depicting the abundance (log₂) of pro- and anti-inflammatory cytokines at 1 year of age. Cytokine levels were measured by stimulating MoDCs from healthy donors (Donor 1-4) with LPS isolated from faecal samples of CSD and VD neonates. **b.** Boxplots depicting the TNF-α levels in both groups (CSD and VD) at day 5 after birth and 1 year of age. Paired two-way ANOVA (analysis of variance) *** p value < 0.001. **c** Correlation of TNF-α levels (row 1) with the relative abundance of metagenomic OTUs based on canonical correlation analysis. Filled squares indicate significantly correlated taxa, whereas color indicates (red) or negative (blue) correlation.

3.3.4 Antimicrobial resistance modulated by birth mode

The analyses of the functional potential based on KEGG orthology revealed a stratification of antibiotic biosynthesis pathways based on whether an infant was born by CSD or VD (**Figure 3.2c**). To assess and validate the impact of birth mode on the presence and persistence of AMR, we used a deep-learning approach [42] to annotate antibiotic resistance genes in our metagenomic data [43]. We determined the presence and relative abundance of ARGs in samples collected from both CSD and VD at day five after birth, 1 month, 6 months and at one year of age. The samples collected from CSD neonates exhibited an increased abundance in ARGs at the earliest time point (day 5) compared to the VD group. Additionally, we found that the number of ARGs detected in CSD infants at one year of age was significantly reduced in comparison to the CSD samples at day 5 (FDR-adjusted $p < 0.0021$, Wilcoxon rank-sum test; **Figure 3.5a**, **Appendix B.2: Supplementary figure 3.4**). To corroborate our observations on the levels of ARGs, we assessed the abundance of ARGs using a random subset of samples from the resistome study by Gasparrini *et al.* [295]. We found that the overall levels of ARGs starting at 1 month through to one year of age were similar in their study to those observed in our own cohort (**Appendix B.2: Supplementary figure 3.5**). Meanwhile in our study, at one year of age, we found several genes that were differentially abundant between the CSD and VD groups (FDR-adjusted $p < 0.05$, Wilcoxon rank sum test; **Figure 3.5b**). Since various genes can confer resistance to the same antibiotic, we regrouped the genes into their respective categories such as multidrug, tetracycline resistance *etc.* We found that genes conferring glycopeptide, phenicol, pleuromutilin, bacitracin, sulfonamide and diaminopyrimidine resistance were significantly increased in CSD compared to VD at day five after birth (**Figure 3.5c**; FDR-adjusted $p < 0.05$, Wilcoxon rank sum test)). Interestingly, diaminopyrimidine, phenicol, pleuromutilin, and sulfonamide are synthetic or semi-synthetic antibiotics, likely prevalent in the hospital environment [296–298]. However, these differences did not persist over time. Additionally, the mothers in our cohort across both groups (CSD:6 and VD:1) received prophylactic treatment against group B *Streptococcus* in the form of cephalosporin. However, we did not find any distinguishing patterns within the resistance categories corresponding to this treatment regimen. Albeit a limited sample size, we also tracked the diet including feeding method (bottle- or breast-fed), antibiotic regimen and physical characteristics through the first year and did not find any significant correlations with functional pathways including AMR (**Appendix B.2: Supplementary figure 3.6**).

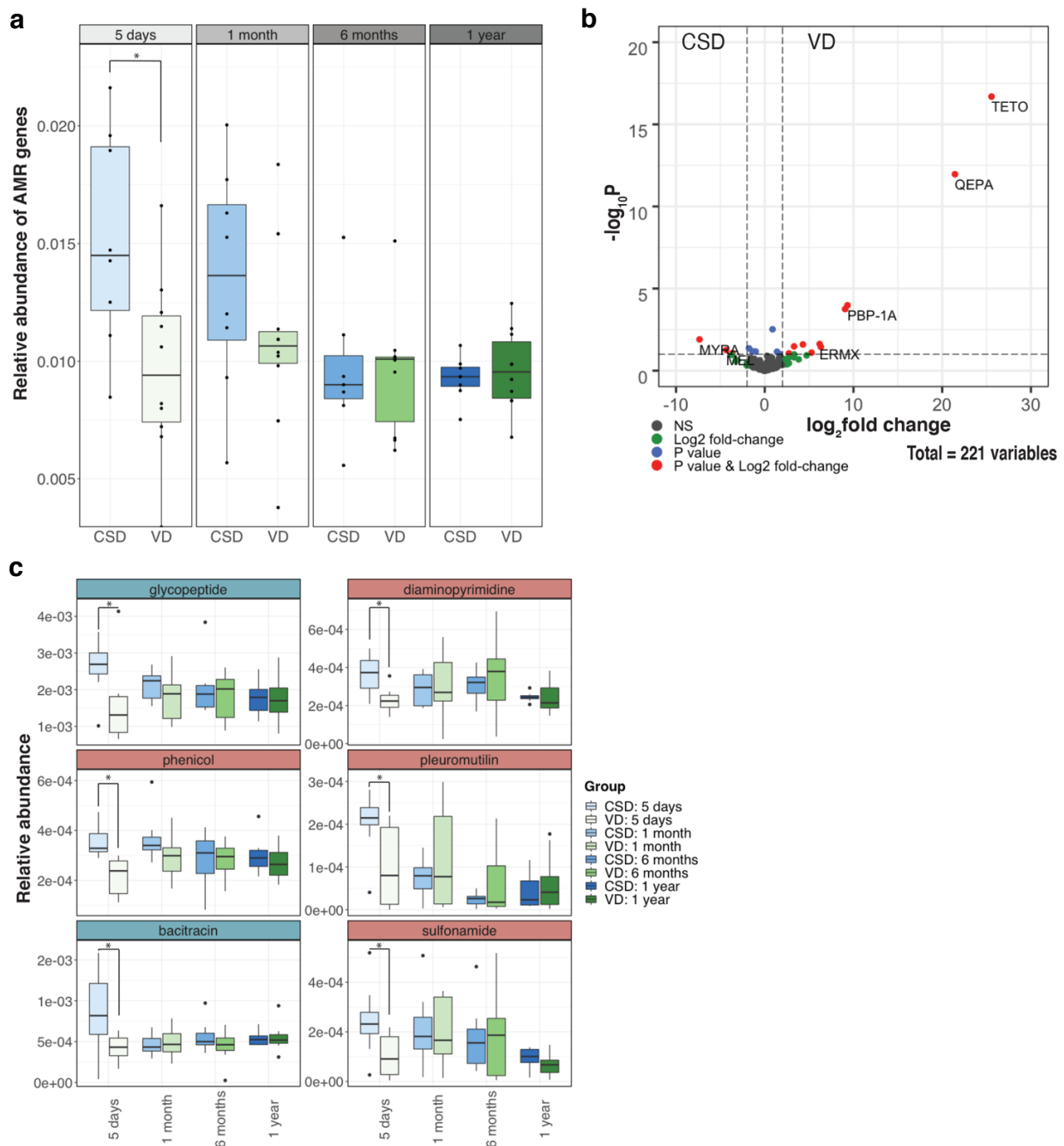


Figure 3.5: Antimicrobial resistance gene abundances over time. **a.** Boxplots of the overall ARG abundance in CSD and VD samples at different timepoints including 5 days after birth and at 1 year of age. Wilcoxon rank-sum tests were used to test for significance. $*p < 0.05$. **b.** Volcano plot depicting the significantly enriched genes in either CSD or VD samples at 1 year of age. **c.** AMR categories which are significantly different between the groups at any of the timepoints are shown. Wilcoxon rank-sum tests were used to test for significance. $*p < 0.05$

Since ARGs were found in the metagenomic data, we investigated which taxa carried these genes by reconstructing metagenome-assembled genomes (MAGs), classifying them taxonomically based on the GTDB database [283], and linking ARGs to individual MAGs. We then compared taxa contributing to AMR between birth mode at day five after birth and one year of age as well as the change over time within the individual groups (**Figure 3.6**).

Figure 2 displays the relative abundance of bacterial taxa in the rumen of CSD and VD goats. The central phylogenetic tree shows the full range of taxa, with labels for various bacterial groups such as Bacteria, Firmicutes, Actinobacteria, and Bacteroidetes. Four smaller trees show the relative abundance of taxa at Day 5 and 1 year for both CSD and VD groups. A color scale indicates Log2 ratio median proportions from -3 (brown) to 3 (green).

53

We compared samples from day five after birth versus one year of age within each birth mode group independently to differentiate between taxonomic groups contributing to AMR. Within CSD samples, we found that *Enterobacteriales* and *Staphylococcaceae* were enriched at day five after birth while major AMR contributors at one year of age were *Lachnospiraceae*, *Bacteroidaceae*, *Actinobacteria*, and *Oscillospirales*. Conversely, within VD samples early AMR resistance was mainly attributed to the abundance of *Bacteroidales*, *Lactobacillales*, *Propionibacteraceae*, and *Enterobacteriaceae* at day five after birth. Meanwhile at one year of age, VD samples were enriched in taxa including *Lachnospiraceae*, *Ruminococcaceae*, *Veillonellales*, and *Eggerthellaceae* with respect to contribution of ARGs to the resistome (**Figure 3.6**). These data suggest that ARGs are also encoded by commensals apart from pathogens which, in the context of the present study, sustain their presence throughout the first year of life.

3.3.6 Role of mobile genetic elements in antimicrobial resistance

Bacterial genomes have been fine-tuned over evolutionary timescales [299], potentially refining their defense mechanisms against various biocidal agents including chemicals. Aside from these, bacterial components such as MGEs are known to be potent factors in the spread of AMR [300] and can transfer genes across distinct taxonomic clades. An example of such MGEs are plasmids as well as viruses including bacteriophages, which actively drive the transfer of genetic material [301]. To determine the role of MGEs in conferring AMR in our neonate cohort, we analyzed the genomic context of the ARGs. The contigs were classified as chromosomal, plasmid, phage, ambiguous (those that could not be resolved), and unclassified. In this study, chromosomal sequences refer to the bacterial genome excluding plasmids, in accordance with the PlasFlow [47] methodology. These criteria were used to assess the role of MGEs at all timepoints. The majority (average of ~75%) of the ARGs were encoded on the bacterial chromosome (**Figure 3.7a**). This phenomenon was prominent in the VD samples irrespective of sampling timepoint. On the other hand, the mean relative abundance of ARGs encoded on plasmids (~5%) was marginally increased in the CSD group at both five days after birth and at one year of age. Overall, we found that phages encoded lower levels (1 - 3%) of ARGs compared to the plasmids. However, we found that the relative abundances of phages encoding AMR were significantly increased after one year of age (**Figure 3.7a**). Interestingly, we did not find any significant differences between the birth modes in relation to the virome profiles at any of the timepoints (FDR-adjusted $p > 0.05$, Two-way ANOVA, **Appendix B.2**:

Supplementary figure 3.7). However, a large proportion of the contigs were either ambiguous or unclassified but demonstrated an even distribution across all timepoints (**Appendix B.2: Supplementary figure 3.8**). When ambiguous sequences mapping to both the bacterial chromosome and phages are included in the phage abundance metrics, it results in a higher abundance of ARGs conferred by phage compared to plasmids (**Appendix B.2: Supplementary figure 3.8**).

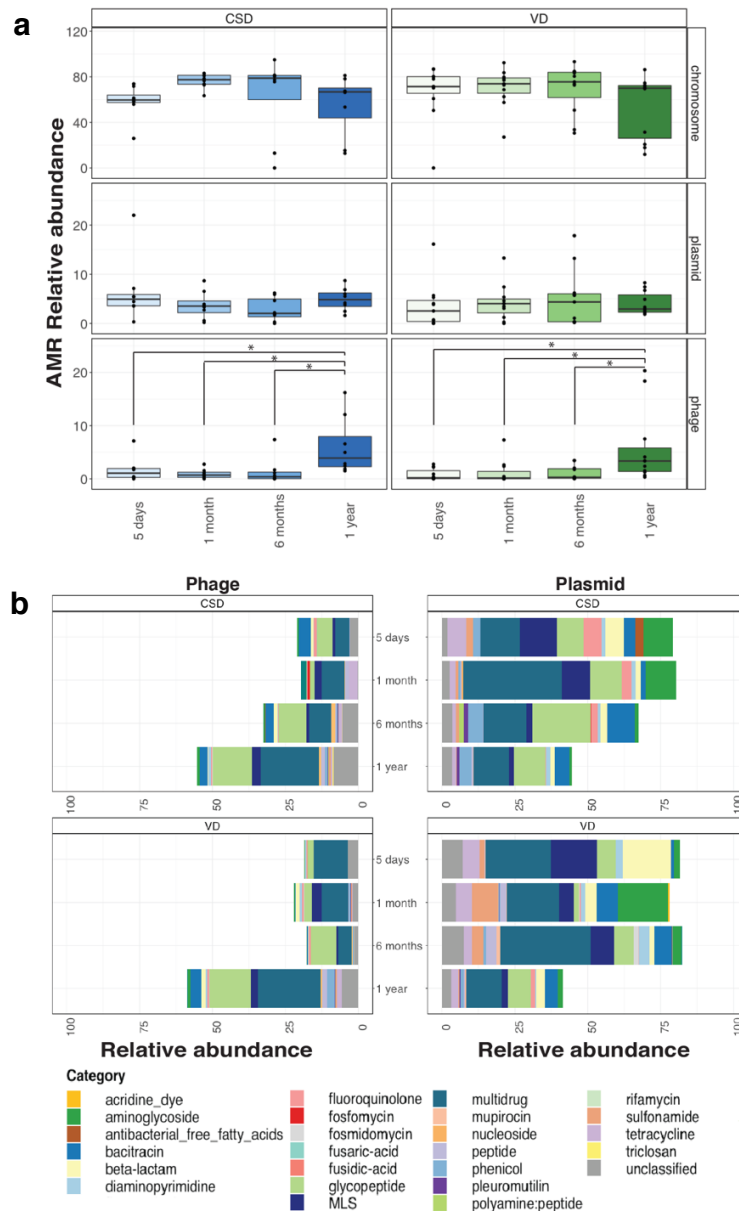


Figure 3.7: Mobile genetic elements associated with antimicrobial resistance. a. The relative abundances of ARGs found on the bacterial chromosome, plasmids, or phages at the different timepoints, ranging from day 5 after birth through to 1 year of age. Paired two-way ANOVA was used to assess significant differences. * p -value<0.05. **b.** Stacked bar plot depicting the AMR categories

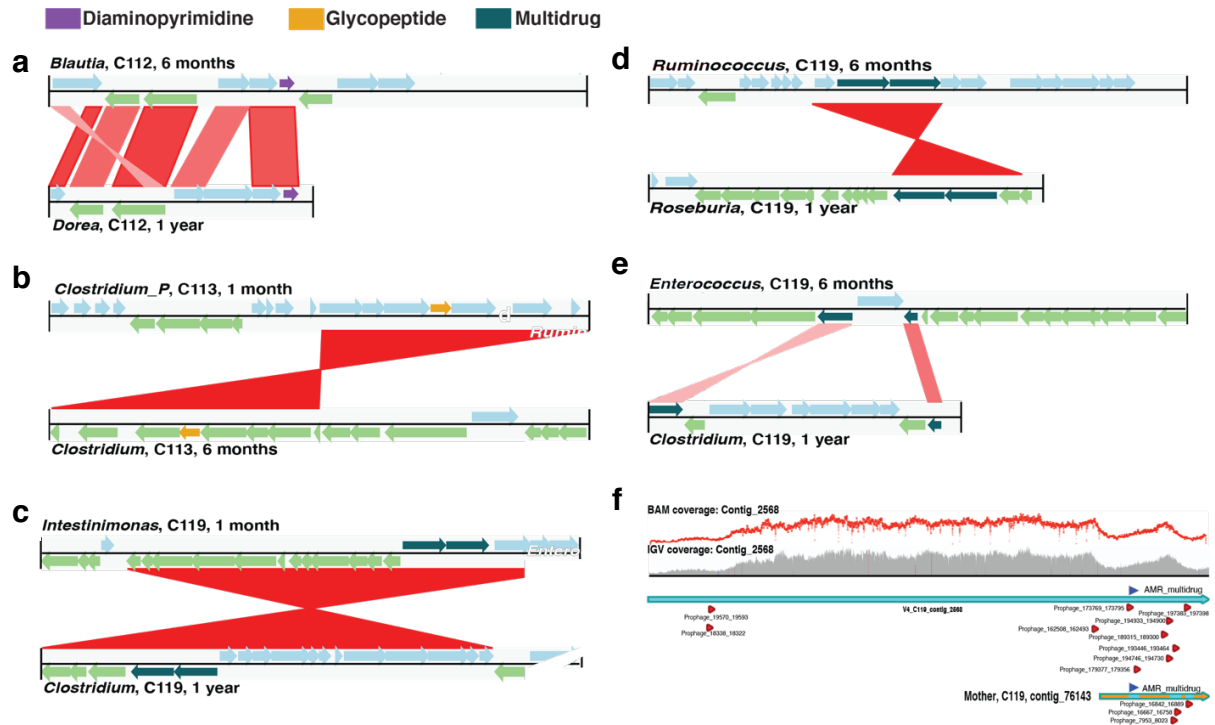
transmitted via phages and plasmid at all timepoints. Each color in the plot is associated with a category listed in the legend on the right. The plot represents mean values for all samples in each group

3.3.7 Distribution of AMR categories encoded by mobile genetic elements

Assessing AMR conferred by MGEs, we found that both plasmids and phages encoded genes conferring resistance to several classes of antibiotics (**Figure 3.7b**). Though significant differences were not apparent, we found that phage-encoded ARGs against vancomycin (glycopeptide) and numerous other antimicrobials were dominant in both birth mode groups. Additionally, plasmids conferred resistance to diaminopyrimidine and bacitracin, as well as β -lactams, phenicol, MLS, and tetracyclines (**Figure 3.7b**). Strikingly, these data suggest that MGE-mediated AMR, encoded by phages, is a potential factor in conferring AMR or serve as a reservoir for antimicrobial resistance throughout the first years of life.

3.3.8 Phage-mediated horizontal gene transfer (HGT)

To understand phage-mediated horizontal gene transfer of AMR we analyzed in detail phage contigs encoding ARGs. We identified several genes that were horizontally transferred within the CSD and VD groups (**Appendix B.2: Supplementary figure 3.9**). CSD samples (n=3; C112, C113, and C119), exhibited HGT involving ARGs including resistance to glycopeptide and multidrug (**Figure 3.8a - 3.8d**). The majority (~88%) mapped to the bacterial chromosome. However, two genes encoding multidrug resistance were encoded by both chromosome and phage (C119: contig 2568). The contig was found to be a candidate prophage based on detailed inspection and was found to encode several genomic regions with prophage signatures flanking the multidrug resistance genes (**Figure 3.8f**). Additionally, the coverage of the contig across its entire length was more variable in the genomic regions where the prophage and ARG sequences were identified. Resolution of the taxa involved indicated HGT between the *Intestinimonas butyriciproducens* (GCA 003096335) and *Clostridium bolteae* (ATCC BAA 613 GCA 000154365) strains belonging to the *Oscillospirallales* and *Lachnospirallales* orders respectively (**Appendix B.2: Supplementary figure 3.10**).



3.4 Discussion

Birth mode is postulated to represent a major factor in shaping earliest gut microbiome colonization and the linked development of neonates especially in relation to the priming of the neonates' immune system [241]. Apart from birth mode, additional aspects such as diet and medical factors have been described to have a significant effect on neonate colonization and succession [255]. Whether or not such effects persist during the first year of life remains an essential question. Here, we performed an in-depth longitudinal analysis of the gut microbiome

using high-resolution metagenomics on samples collected during the first few days *postpartum* through to the first year of age. We specifically assessed the pervasive effect of birth mode-dependent microbiome differences in relation to immune system priming and AMR. Additionally, we analyzed the contribution of mobile genetic elements and the role of horizontal gene transfer in conferring AMR (**Figure 3.9**).

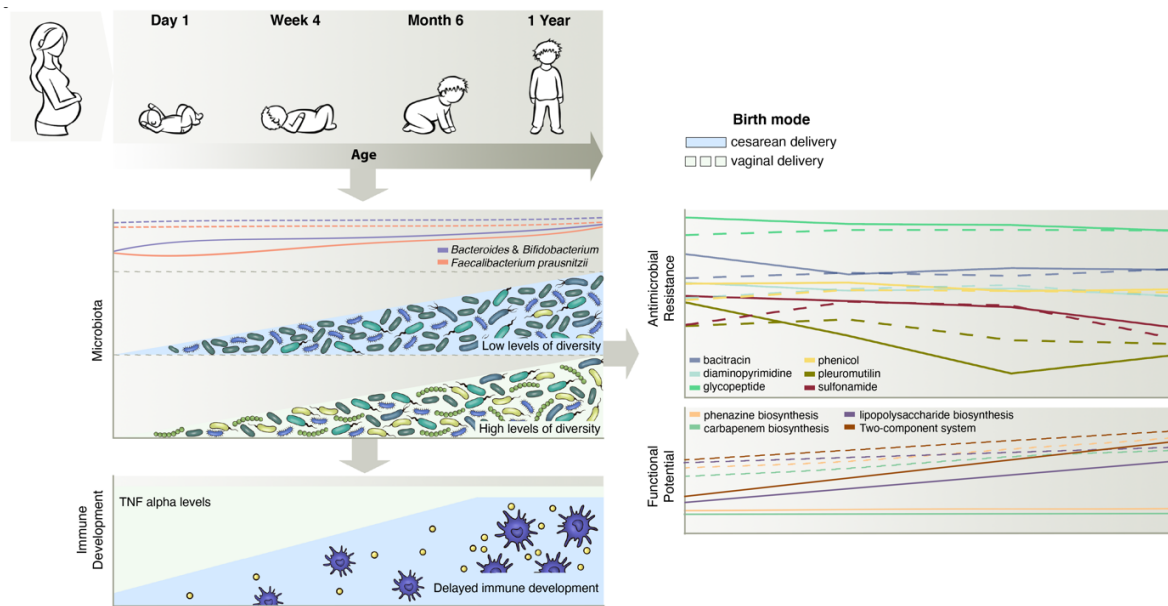


Figure 3.9: Summary figure: Depicting the longitudinal analyses of the infant gut microbiome during the first few days *postpartum* through to the first year of age. Panels representing Microbiome composition, TNF alpha levels, antimicrobial resistance levels and information the functional potential.

Our previous findings [165] along with the longer-term trends of the present study underpin the notion that persistent structural and functional differences exist in the gut microbiomes of neonates born by CSD. More specifically, our results agree with other studies which have highlighted a reduced abundance and colonization by taxa such as *Bifidobacterium* and *Bacteroides* in CSD neonates [165,243,260,302,303]. Additionally, we found that the levels of *Faecalibacterium prausnitzii* were significantly elevated in VD infants after one year of age. *F. prausnitzii* is a highly abundant commensal in the human gut including those with higher levels of diversity and richness when compared to individuals following a Western lifestyle [304]. Concurrently, this taxon has been reported to be reduced in the gut of patients with ulcerative colitis and Crohn's disease [305], which in turn may be linked to it being a keystone taxon conferring anti-inflammatory properties in humans [306,307]. Further studies are necessary to effectively understand the longer-term consequences of the differential abundance of *F. prausnitzii* in humans beyond the first year of age.

Our previously published study [165] highlighted a higher potential for LPS-mediated immune priming in VD compared to CSD at day five after birth. Conversely, we found that LPS extracts from the CSD samples taken at one year of age resulted in significantly higher TNF- α levels compared to five days after birth. Our results indicate that a reduction in earliest immune system priming through key immunogenic molecules occurs in CSD neonates. This might lead to persistent effects throughout the first year of life which, in turn, may explain the higher rates of immune system-linked diseases observed in CSD infants in later life including metabolic disorders [247,302] and allergies [248,249]. Along these lines, Jakobsson *et al.* previously showed that children born via CSD have reduced Th1 responses [308]. Furthermore, other groups have reported that early life immune system stimulation impacts immune-disorders including asthma [309], allergies [310], diabetes and IBD [311]. In this context, our findings indicate that birth mode-dependent gut microbiota alterations affect the status of the immune system throughout the first year of life, and likely beyond. This in turn may explain immunological deficits linked to numerous chronic diseases for which a higher propensity is observed in individuals born by CSD [312].

Birth mode-associated alterations of the gut microbiota may facilitate colonization by opportunistic pathogens, including those encoding antimicrobial resistance [251]. Functional analyses of our metagenomic data highlighted enrichments in carbapenem and phenazine biosynthesis genes in the VD group after the first year of life, potentially a consequence of endogenous gut bacteria-mediated resistance mechanisms against opportunistic pathogens in the gut. Both carbapenem and phenazine are known to be bacterial compounds that are used clinically in fighting Gram-positive and Gram-negative pathogens [313,314]. This data suggests that the indigenous gut microbiota plays a crucial, early role in conferring colonization resistance against pathogens. In addition, we found that CSD is associated with resistance against semi- and synthetic- antibiotics as early as five days after birth. It is well established that mothers undergoing CSD are administered antibiotics to prevent nosocomial infections, as a prophylactic policy [294,315,316]. Interestingly, we did not find resistance towards the antibiotic treatment administered to mothers in our cohort. It however remains plausible that the enrichment in ARGs especially against phenicol, pleuromutilin and diaminopyrimidine classes at day five after birth in CSD neonatal samples is linked to the hospital environment including the actual caesarean section.

In conjunction with the observed differences in AMR between CSD and VD our study also highlights the potential mode of AMR transmission via mobile genetic elements including via plasmids and/or bacteriophages [317–319]. Parnanen *et al.* reported the presence of ARGs and MGEs in infant faecal samples at 1 and 6 months of age [320]. Our findings agree with their results and expand on these by additionally providing data on the abundance of ARGs and MGEs at five days after birth. Furthermore, we identified 27 categories of ARGs and linked these to both bacterial taxonomy and MGEs. We found that both plasmids and phages encoded genes which confer resistance to several classes of antibiotics. Of all MGEs, plasmids conferred resistance to a variety of antimicrobial compounds. Furthermore, we found that glycopeptide and multidrug resistance were transferred via phages, in accordance with previous reports [321–323]. We also found that horizontal gene transfer plays a critical role in the continued transmission of AMR during the first year of life. While ~88% of HGT occurred via canonical methods involving the bacterial chromosome and plasmids, we found that prophages contributed to multidrug resistance in one CSD sample (C119). We detected prophage signature sequences flanking two multidrug resistance genes, horizontally transferred between bacteria from two distinct orders. Our findings thereby highlight the role of prophages, typically thought to mediate AMR in humans pathogens including *Staphylococcus aureus* [324], *Salmonella* and shiga-toxin producing *Escherichia coli* [32], as mediators of HGT even among commensals. Intriguingly, HGT events in the VD samples did not indicate any ARG transmission. On the other hand, considering the smaller sample size, further studies with an increased power are needed to clarify the role of phage-mediated AMR resistance especially during the first few days of life.

The persistence of differences in early life exposure is an important but challenging research question, not least because of the paucity of long-term, longitudinal studies ranging from immediately after birth until early childhood. Our findings imply that birth mode leads to persistent gut microbiota structural and functional differences. We acknowledge that the limited sample size and the lack of detailed dietary information cannot rule out other confounding factors such as the *in utero* and *postpartum* environments of the infants. However, and importantly, our data suggest that gut microbiota structural and functional effects may predispose infants delivered by CSD to delayed immune priming resulting in a deficiency in tolerance. Our results pave the way for future, rational interventions aimed at restoring key functional features of the microbiota. In this context, further studies including following the children over extended periods of time are needed to understand birth mode-mediated

manifestations of disease. Concurrently, an important research direction which arises from our study centers on the role of the gut mobilome in conferring AMR and how this affects microbiome trajectories and linked phenotypic outcomes in humans. Considering current global efforts directed at limiting the emergence of antibiotic resistance [325], appreciation of the role of phages as an additional source of resistance may be necessary for success in reducing the overall burden of AMR in the future.

Chapter 4. Evolution of the gut resistome following a selective antibiotic sweep

A major part of this chapter is based on the following publication submitted for peer-review:

Laura de Nies*, Susheel Bhanu Busi*, Mina Tsenkova, Elisabeth Lettelier and Paul Wilmes (2021). Evolution of the gut resistome following a selective antibiotic sweep. *Nature Communications* **in review** [Appendix A.4]

*Co-first author

4.1 Background

Prior to the advent of antibiotics, bacterial infections were the leading cause of disease and mortality in humans. Antibiotic usage is now commonplace in treating infections [326] as well as ensuring the safety of surgical procedures [327,328] and organ transplantation [329]. In addition, they are extensively used in animal husbandry [330] and also in animal models for studying the gut microbiome [331–336]. However, many bacterial taxa have evolved antimicrobial resistance (AMR) to several classes of antibiotics, and multidrug-resistant bacteria have now emerged, preventing comprehensive treatment of infections and resulting in a growing number of deaths [337]. Therefore, a clear understanding of the selective sweeps underlying the evolution, speed and transmission of antimicrobial resistance genes (ARGs) is of crucial importance.

In general terms, bacteria can acquire and develop AMR through two distinct genetic mechanisms. Either through the acquisition of spontaneous mutations during replication of the bacterial genome or through the accumulation and dissemination of resistance genes via mobile genetic elements (MGEs) (**section 1.2**) [20]. Although the majority of bacteria are harmless commensals which do not cause disease, they still provide a rich repertoire of resistance genes [189]. As such, through HGT, opportunistic pathogens may acquire resistance genes from other commensals. Furthermore, while resistant bacteria may remain latent, they contribute to the overall reservoir of AMR based on which resistant pathogens may emerge once selective pressure is built up [338,339]. Compounding this phenomenon, the overuse of antibiotics both in treatment of human disease and animal husbandry has fueled the build-up of AMR globally [340].

With the realization that the gut microbiome plays crucial roles in disease etiology [341], antibiotic-treated animal models remain one of the methods by which the intestinal microbiome is studied [341] [342]. The potential caveats, however, of using antibiotics for modulating the endogenous populations, including the emergence of AMR is unknown. Therefore, utilizing a mouse model, we assessed the effect of selective sweeps on the evolution of AMR within the commensal gut microbiome population over a single mammalian lifespan after a single course of antibiotic treatment. Our observations allow us to test whether specific bacteria, including commensals, are more susceptible or capable of acquiring ARGs, as well as assess the influence of HGT on shaping the gut microbiome's resistome.

4.2 Methods

4.2.1 Power calculation and sample size estimate

To determine the number of animals required per treatment and control group we performed a multifactorial power analysis based on a 2015 study by Raymond *et al.* [343]. We estimated the Jensen-Shannon Divergences (JSD) of the microbial profiles of the antibiotic-treated (cefprozil) and control groups. Based on the observed JSD the inter-group variability was significantly high, demonstrating a minimum sample size of three mice per group to account for a power of 80% and a 5% alpha error rate, indicating changes in microbial composition (**Appendix B.3: Supplementary figure 4.1**).

4.2.2 Mice model and antibiotic exposure

C57BL/6J mice were bred in-house and experiments were performed according to all applicable laws and the regulations, after receiving approval by the institution's animal experimentation ethics committee and the veterinarian service of the Luxembourg Ministry of Agriculture (Permit Number: LUPA 2019/13). To limit individual variation of the gut microbiome in experimental groups, mice of the same age were obtained from the same vendor and the same location in the vendor facility. After a 7-day quarantine and subsequent acclimation period of one week, mice were maintained in single housing conditions for each experiment. Mice were housed in Allentown NexGen Mouse 500 (194mm x 130mm x 381 mm) cages with JRS Rehofix Corncob bedding. Mice had access to reverse osmosis water with 2ppm of chlorine fed *ad libitum* along with standard A40 chow diet (SAFE, France). The animals were maintained under standard habitat conditions (humidity: 40-70%, temperature: 22°C) with a 12:12 light cycle. Two groups of mice were established (control and treatment), and each group contained 4 animals (2 males + 2 females). Antibiotics, ampicillin (1g/L), vancomycin (500mg/L), metronidazole (1g/L) and neomycin (1g/L) were chosen for their utility in several mouse models [344,345] and in line with most preoperative procedures [346]. They were administered as a cocktail within the drinking water to the treatment group starting at 8 weeks of age. Antibiotics were administered during a period of one week, after which the change was made to regular drinking water for the duration of the recovery period. Fecal samples were collected daily for a duration of 19 days (both treatment and recovery phase) starting prior to the antibiotic treatment till take down (**Figure 4.1**).

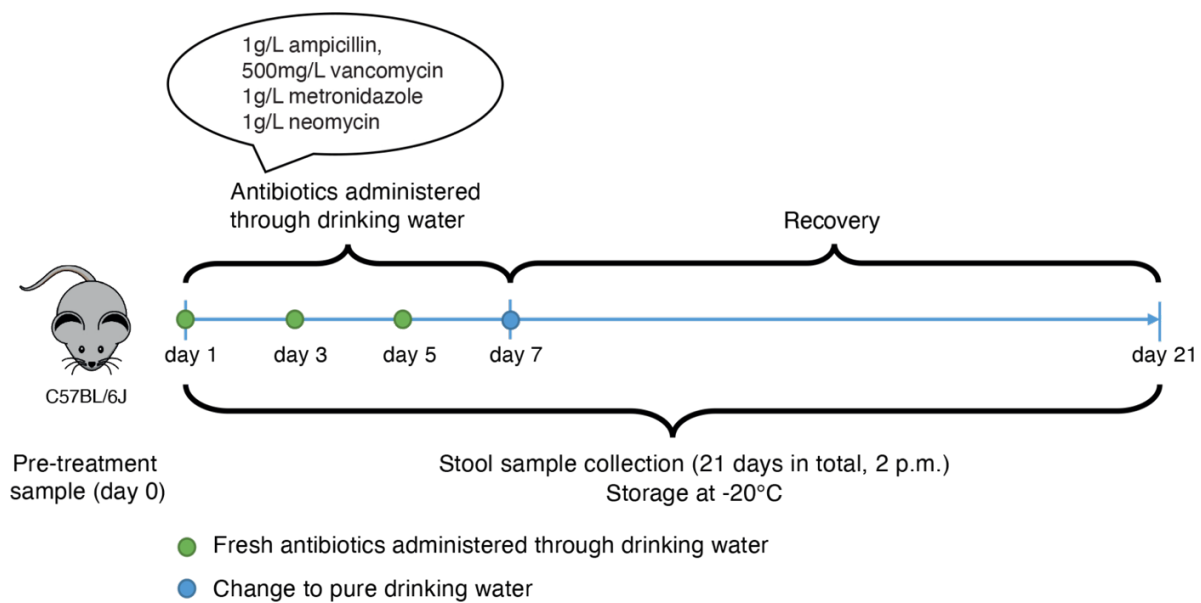


Figure 4.1: Representative illustration of the experimental design. Project overview, including dates and collections of treatment, and fecal sample collection. Eight single-housed mice per group were treated with an antibiotic cocktail (treatment) or water (control) longitudinally. Fecal samples collected at days 0, 7 and 21 were used for microbiome and antimicrobial resistance profiling.

4.2.3 Faecal processing and nucleic acid extraction

A total of 48 faecal samples were acquired across three timepoints, i.e., prior to treatment: day 0, immediately after treatment: day 7, and after recovery: day 21, from each of the mice. 50 mg of frozen stool samples were aseptically weighed into sterile vials. Genomic DNA was isolated with the DNeasy PowerSoil Kit (Qiagen, Luxembourg) including an additional incubation step at 65 °C and milling, as described previously [347]. A minimum of 200 ng of total DNA was recovered from all the samples, yielding sufficient DNA for metagenomic sequencing including high-resolution, artefact-curated metagenomic data for subsequent analyses [165]. DNA extracted from all timepoints was thereafter stored at -80 °C until further use.

4.2.4 DNA sequencing

All DNA samples were subjected to random shotgun sequencing. Briefly, 200 ng of DNA was used for metagenomic library preparation using the Westburg NGS DNA Library Prep Kit (cat. no. WB 9096). The genomic DNA was enzymatically fragmented for 12 min and DNA libraries were prepared without PCR amplification. The average insert size of libraries was 480bp. Prepared libraries were quantified using Qubit (Invitrogen) and quality checked on a

Bioanalyzer instrument (Agilent). Sequencing was performed on NextSeq500 instruments using 2x150 bp read lengths.

4.2.5 Data processing for metagenomics, including genome reconstruction

The Integrated Meta-omic Pipeline (IMP) [231] was used to process paired forward and reverse reads using the built-in metagenomic workflow as previously described [211]. The workflow includes pre-processing, assembly, genome reconstruction and functional annotation of genes based on custom databases in a reproducible manner. After trimming the adapters, the preprocessing step included the removal of *Mus musculus* (GRCm38.p6 (GCA_000001635.8); retrieved on 16-May-2020 from https://www.ensembl.org/Mus_musculus/Info/Index) reads. Thereafter the *de novo* assembly was performed using the MEGAHIT (version 2.0) assembler [276]. Default IMP parameters were retained for all samples. Metagenomic operational taxonomic unit (mOTU) profiles were generated from the trimmed and preprocessed reads to generate microbiome profiles for the control and treatment groups using mOTUs v2.5.1 [281]. Concurrently, we used MetaBAT2 [277] and MaxBin2 [278] for binning in addition to an in-house binning methodology previously described [211] for genome reconstructions, i.e. metagenome-assembled genomes (MAGs). Subsequently, we obtained a non-redundant set of MAGs using DASTool [280] with a score threshold of 0.7 for downstream analyses, and those with a minimum completion of 90% and less than 5% contamination as assessed by CheckM [282]. Taxonomy was assigned to the MAGs using the extensive database packaged with gtdbtk [283]. To generate pangenomes, we collected all the bins taxonomically identified as *Akkermansia muciniphila* and used the anvio-based pangenome workflow described by Meren *et al.* (<http://merenlab.org/2016/11/08/pangenomics-v2/>) [348]. One of the treated mice (#16), was excluded from the pangenome analyses due to the unavailability of MAGs.

4.2.5 Identification of antimicrobial resistance genes and association with mobile genetic elements

We used PathoFact [349], a pipeline for the prediction of virulence factors and antimicrobial resistance genes, to predict and identify ARGs within our metagenomes. The assembly files from individual samples were used as input for the AMR analyses. ARGs were then collapsed into categories based on the Comprehensive Antibiotic Resistance Database (CARD) [43] and identified using PathoFact [349]. Thereafter, the relative abundance of the ARGs was calculated using the Rnum_Gi method described by Hu *et al.* [233].

Identified ARGs and their categories were linked to associated bacterial taxonomy using the metagenomic bin classifications. Furthermore, utilizing PathoFact, ARGs were linked to predicted mobile genetic elements (MGEs: phages and plasmids) to identify probable transmission of AMR between taxa. More specifically, to link both the MGEs and the taxonomy to the ARGs, we mapped the genes to assembled contigs, followed by identifying the corresponding bins (MAGs) to which the contigs belonged. By considering all different predictions of MGEs, a final classification was made based on the genomic contexts of the ARGs encoded on plasmids, phages or chromosomes, including classification of those that could not be resolved (ambiguous). The ARGs that could not be assigned to either the MGEs or bacterial chromosomes were further referred to as unclassified genomic elements. Certain ARGs were encoded on both the bacterial chromosome and phage genomes.

4.2.6 Linking ARGs with integrons

The assemblies generated via IMP were used to assess the presence and abundance of integrons within the metagenomes. Briefly, *attC* sites were identified by HattCI[350] while for the annotation of the *intl* sites a BLAST database was created using the *intl* variant sequences from the UniProt database[226]. Only those contigs where both the signature genetic regions (*intl* and *attC*) were found were annotated as having 'complete' integron elements. We also identified the MAGs along with which the integrons were binned, thus linking the integrons to the reconstructed genomes. The ARG information was overlaid onto this to identify contigs where integrons were linked with ARGs. Furthermore, we used sequence coordinates to identify integron localization, i.e., chromosome, plasmid or phage localization of gene cassettes, incomplete and complete integrons on MGEs. This information was used for downstream differential analyses.

4.2.7 Data analysis

Figures for the study including visualizations derived from the taxonomic and functional, were created using version 3.6 of the R statistical software package. GraphPad [351] was used to generate the figures for describing the longitudinal weight measurements of the mice. DESeq2 [234] and Wilcoxon rank-sum tests with FDR-adjustments for multiple testing were used to assess significant differences for the AMR and taxonomic analyses whereas a paired two-way ANOVA (Analysis of Variance) within the *nlme* package was used for identifying statistically

significant differences in the integron profiles. Chord diagrams for the HGT events were generated using scripts found within the MetaCHIP package [288] while the pangenome visualizations were obtained using anvi'o [352].

4.3 Results

4.3.1 Selection of specific taxa due to antibiotic-mediated depletion of the gut microbiome

To assess the selective pressure on AMR evolution, two groups of mice were housed in specific pathogen-free conditions, where one of the groups was treated with an antibiotic cocktail (ampicillin, vancomycin, metronidazole and neomycin) representing broad-spectrum antibiotic treatment regimens used in preoperative procedures and in animal models. Longitudinal faecal samples were collected from the control and antibiotic-treated mice prior to treatment, i.e., day 0, immediately after treatment, i.e. day 7, and after recovery, i.e. day 21 (**Figure 4.1**).

Although there was a drop in weight during the antibiotic treatment phase, mice treated with antibiotics did not show any significant differences in weight compared to the control mice (**Appendix B.3: Supplementary figure 4.2**). We assessed the overall microbiome profile at days 0, 7 and 21 and observed a major shift in the community profiles of the antibiotic-treated mice at both days 7 and 21 (**Appendix B.3: Supplementary figure 4.3**). At the level of metagenome-assembled genomes (MAG), taking only into account high-quality genomes, i.e., > 90% complete and < 5% contamination, we found a significant enrichment in *Akkermansia muciniphila* on day 21 after antibiotic treatment, despite a near-total depletion of the microbiota at day 7 in the treated mice (**Figure 4.2**). Simultaneously, several genera such as *Alistipes*, *Odoribacter*, members of Muribaculaceae, and *Prevotella* (including CAG-95, CAG-485, CAG-873) were significantly decreased or depleted entirely at days 7 and 21 (**Figure 4.2**), demonstrating the potency of the antibiotic cocktail. Concomitantly, the overall functional potential of the metagenomes of the antibiotic-treated mice was altered, whereby we found a shift in the functional profile due to the treatment (**Appendix B.3: Supplementary figure 4.4a**). The functional variation demonstrated a significant enrichment in pathways relating to signaling molecules at days 7 and 21 compared to the controls (**Appendix B.3: Supplementary figure 4.4b**). Interestingly, at day 7, we also observed a significant decrease in genes involved in the biosynthesis of secondary metabolites in the antibiotic-treated mice, which may be associated with the depletion of the microbial community owing to the antibiotic treatment.

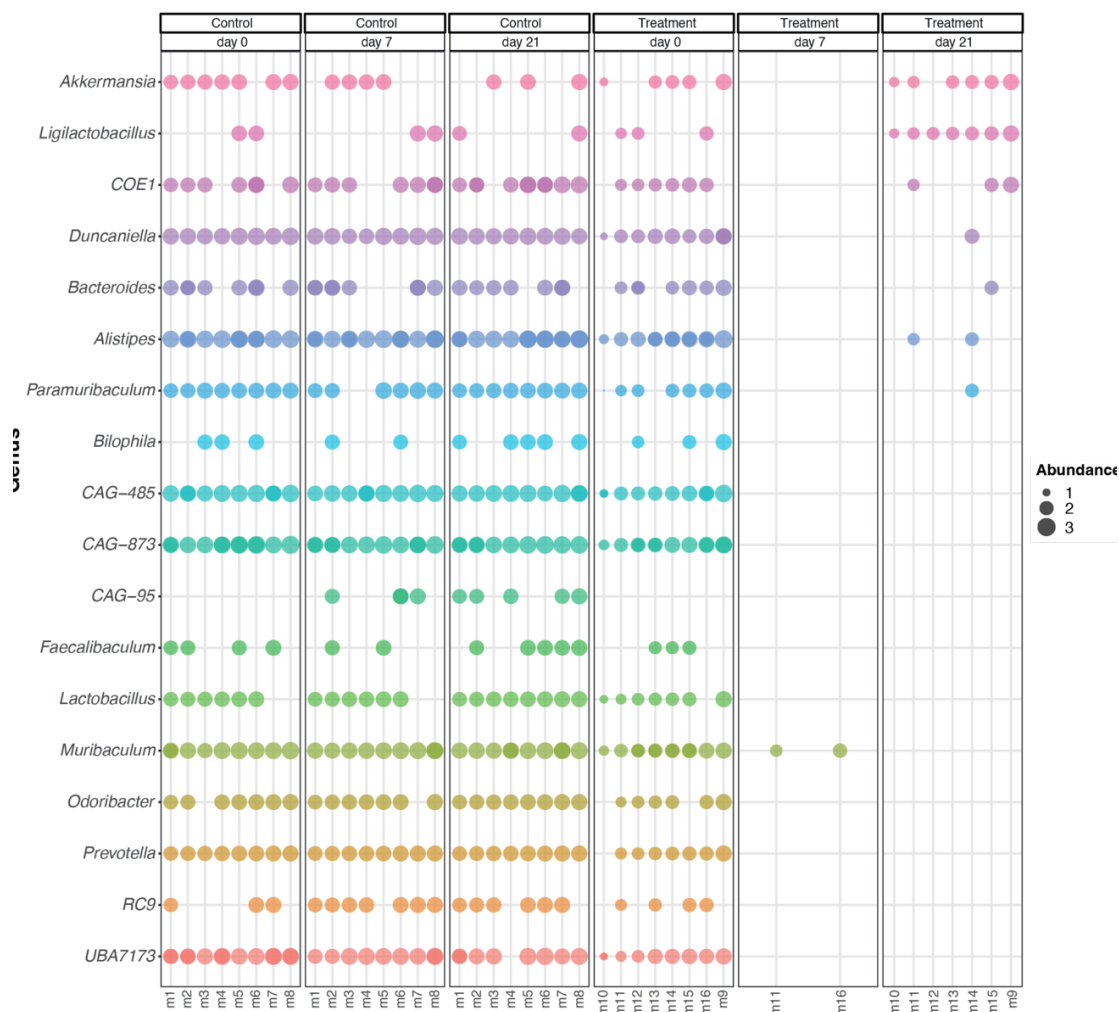
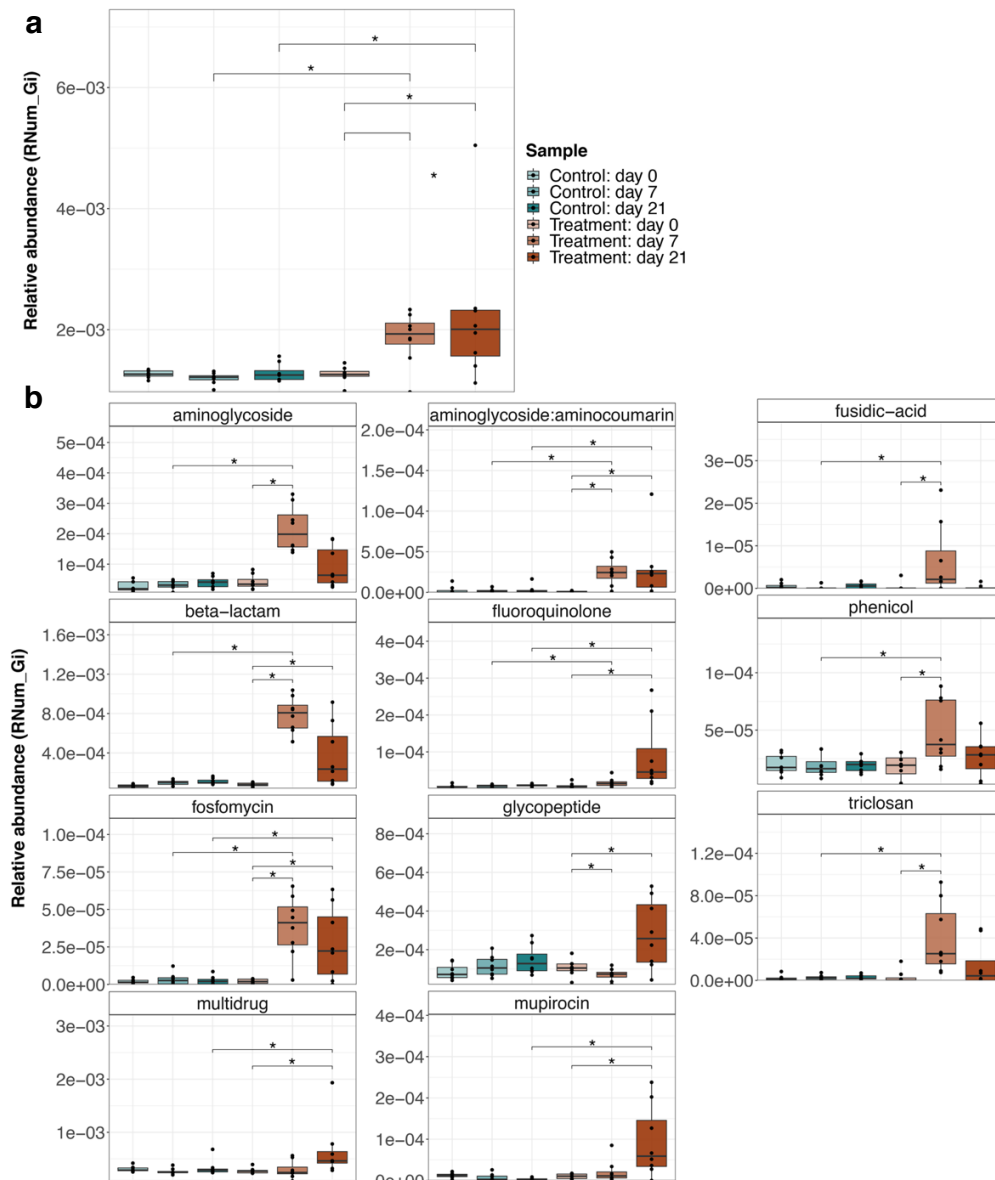


Figure 4.2: Metagenome-assembled genome profiles. Genus level representation of the MAGs recovered from control and treatment groups – pre and post (day 0, day 7 and day 21 respectively) antibiotic administration.

4.3.2 Resistome after antibiotic treatment

Due to the differences observed in the functional potential associated with interactions and secondary metabolites, we further assessed the abundance of ARGs in the control and treated mice. We observed a significant increase in ARGs within the treated group (adj. $p < 0.05$, Wilcoxon rank-sum tests; **Figure 4.3a**). The antibiotic-treated mice had significantly enriched AMR abundances compared to the control mice directly after antibiotic treatment at day 7 which was maintained during recovery until day 21. We further found that within the treated mice the overall AMR abundance was significantly higher after treatment at days 7 and 21, compared to pre-treatment at day 0 (adj. $p < 0.05$, Wilcoxon rank-sum tests). Specifically, we found antimicrobial resistance against aminoglycoside, aminoglycoside:aminocoumarin, beta-lactam,

fluoroquinolone, fosfomycin, glycopeptide, fusidic-acid, phenicol, mupirocin, triclosan and multidrug to be significantly enriched within the treated group (**Figure 4.3b**). Of these, three resistance categories, namely aminoglycoside (neomycin), beta-lactam (ampicillin) and glycopeptide (vancomycin), could be directly linked back to the administered antibiotics. While the other resistance categories are not associated with the administered antibiotics, it is likely that they were indirectly selected due to their co-localization along with other resistance genes [353,354].



4.3.3 Antibiotic-induced changes in taxonomic composition

Since the metagenomes revealed an enrichment in different AMR categories, we investigated which taxa harbored these ARGs. We linked ARGs to individual genomes by identifying contigs encoding ARGs and their corresponding assignment to MAGs including taxonomic classification of the MAGs using gtdbtk [283]. We subsequently compared taxa contributing to AMR between the groups, including mice treated with antibiotics and those without. While taxa contributing to AMR within the control group remained constant, within the treatment group a shift of AMR associated taxa was observed after recovery at day 21 (**Figure 4.4a**). Alongside the increase in the abundance of several taxa (**Figure 4.4a**), we found that an abundance of overall ARGs was increased in taxa belonging to the Akkermansiaceae, Enterococcaceae and Lactobacillaceae families across all treated mice, as well as compared to the control group, at day 21 (**Figure 4.4b**). Given the enrichment in ARGs at day 21, as opposed to day 7, it is likely that the observed ARGs were acquired over time, rather than being encoded as intrinsic resistance mechanisms.

4.3.4 MGEs linked to AMR dissemination

Mobile genetic elements are an established mechanism for the dissemination of AMR. To determine the function of MGEs in conferring the resolved ARGs under selective pressure, we analyzed the genomic context of the ARGs. The majority of the resistance genes were encoded on the bacterial chromosome, while ARGs were found to a lesser extent to be encoded on both phages and plasmids (**Figure 4.5a**). Interestingly, a depletion in the general abundance of plasmids was observed at day 7, which subsequently recovered at day 21. In contrast, phages linked to AMR were significantly enriched in the treated mice at day 7 compared to both pre-treatment at day 0 and the controls at day 7 (**Figure 4.5a-b**). Moreover, when analyzing the specific categories of the resistome, we found an increase in the abundance of phage sequences linked to aminoglycoside, aminoglycoside:aminocoumarin and beta-lactam resistance, in conjunction with the administered antibiotic cocktail (**Figure 4.5b**). Significant differences of these phage-associated AMR categories were observed in the treated group compared to the controls at day 7 and also within the group when comparing phage-mediated AMR levels between days 0 and 7 (**Figure 4.5c**).

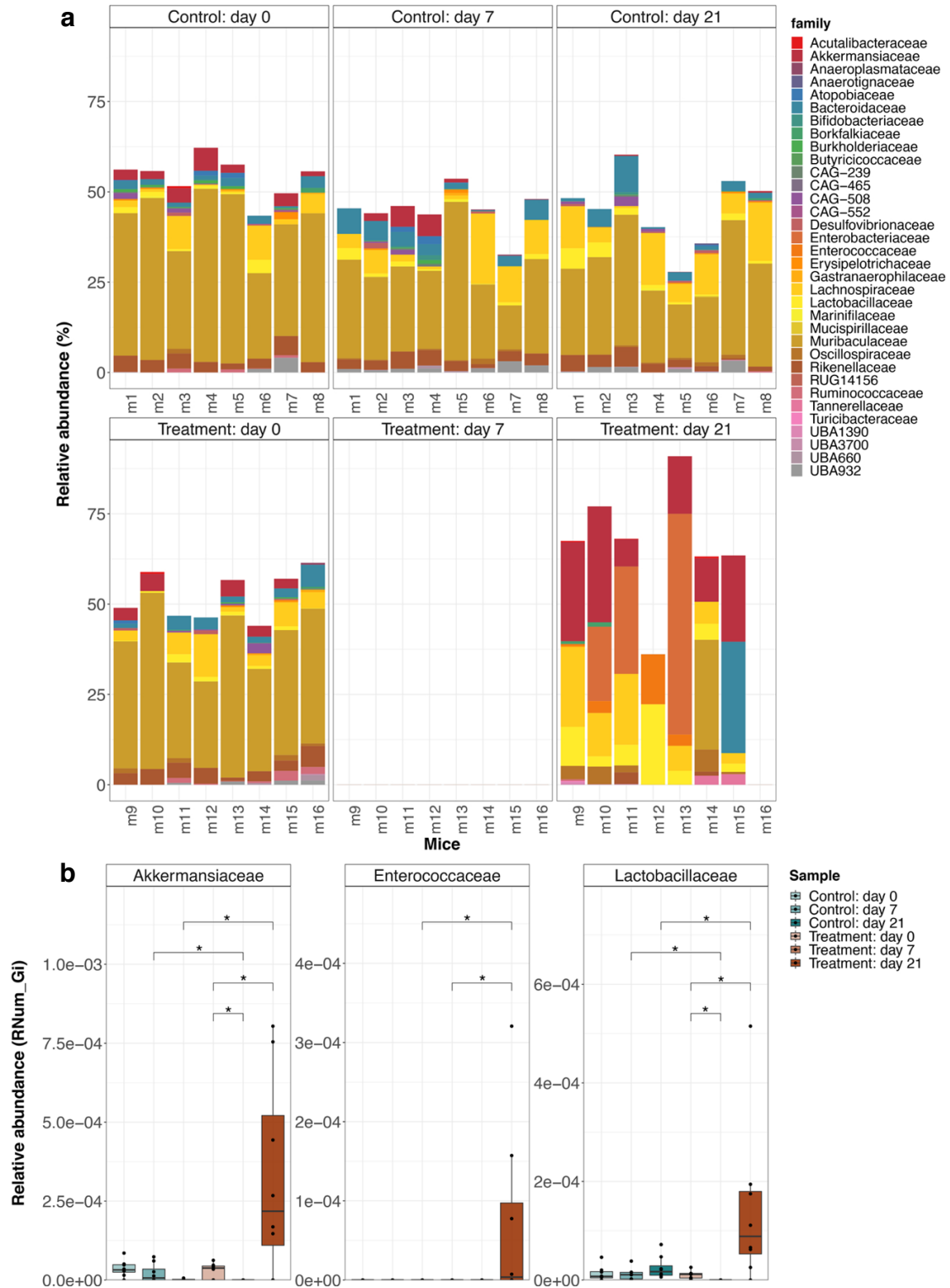


Figure 4.4: AMR-associated taxonomy. **a.** Barplots depicting the relative abundance of ARGs associated with MAGs (Family level) in each sample. **b.** Relative abundance of ARGs associated with Akkermansiaceae, Enterococcaceae and Lactobacillaceae in the control and treated mice. *adjusted p -value < 0.05 (Wilcoxon rank-sum test).

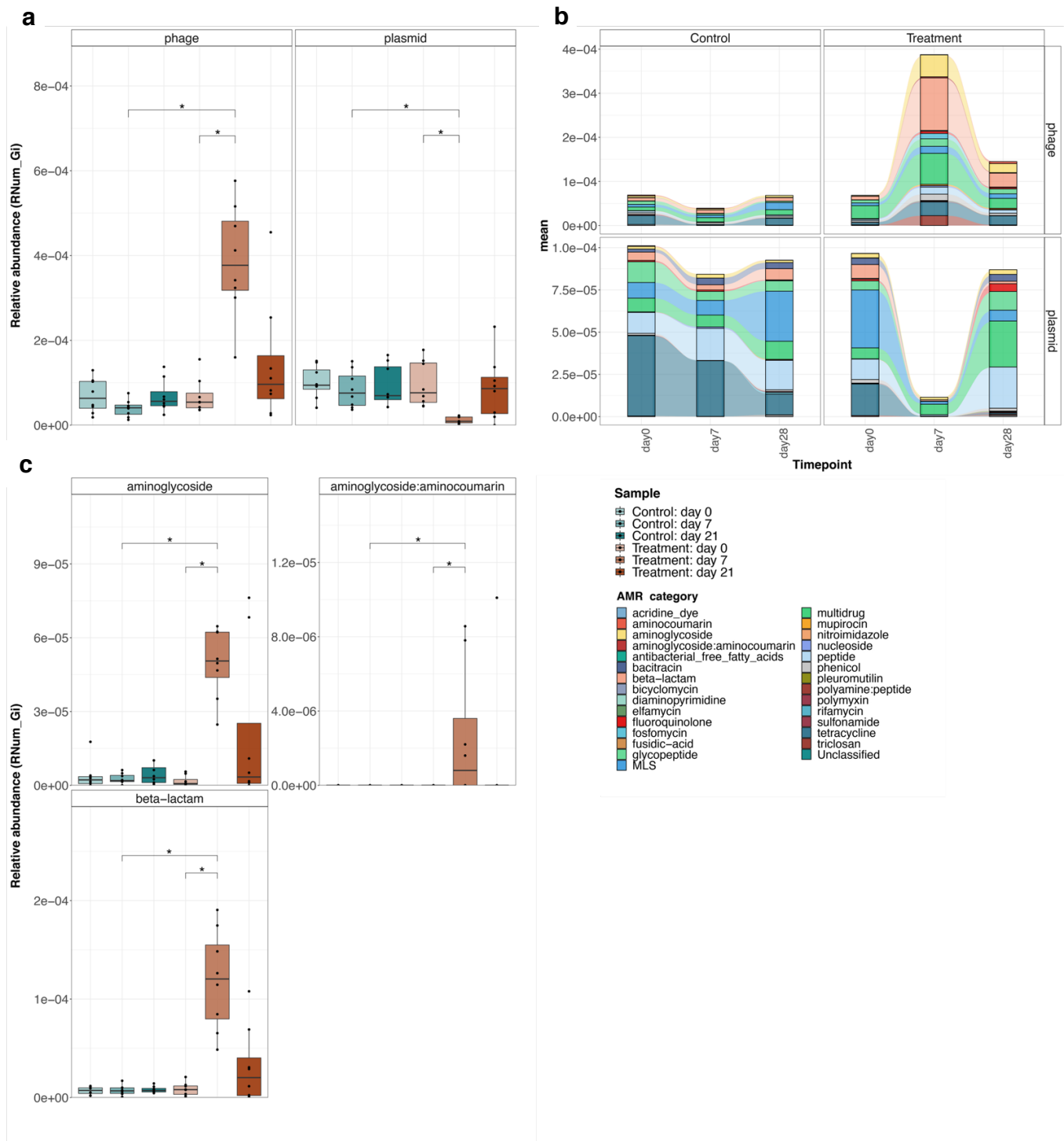


Figure 4.5: Abundance levels of AMR categories associated with MGEs. **a.** Relative abundance of AMR mediated via MGEs, i.e. phage and plasmid. **b.** Abundance levels of AMR categories disseminated via phages and plasmids. Categories pre- and post-treatment (day 0, day 7, day 21). **c.** Abundance levels of aminoglycoside, aminoglycoside:aminocoumarin and beta-lactam resistance genes mediated via phages. * $adj.p < 0.05$ (Wilcoxon rank-sum test).

4.3.5 Integrins mediate AMR in antibiotic-treated mice

To further investigate the effect of antibiotic treatment on the evolution of AMR within the microbiota, we assessed the pangenomes of the significantly enriched and recalcitrant taxa in the treated mice, i.e. *Akkermansia muciniphila* and *Ligilactobacillus* spp.. Interestingly, pangenome analyses of *Akkermansia muciniphila* revealed the acquisition of several genes, including those mediated by integrases (Figure 4.6).

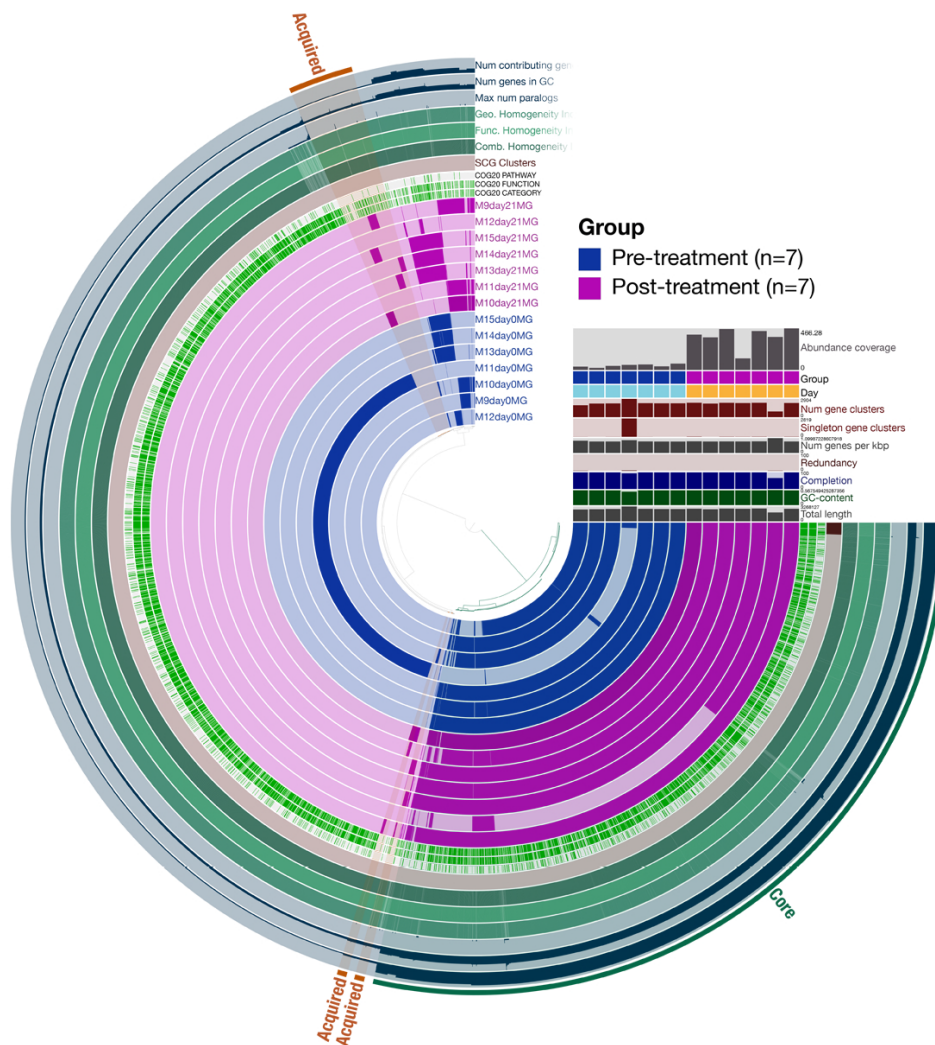


Figure 4.6: Comparison of *Akkermansia muciniphila* genomes. Anvi'o based visualization of the *Akkermansia muciniphila* genomes from the pre- and post- treatment (day 0 and day 21 respectively) samples in blue and purple respectively. Annotation on the outer ring indicate the 'Core' genome and the 'Acquired' genome partially (75%) mediated by integrins.

Horizontal gene transfer is typically attributed to plasmids and phages in metagenomes. However, integrons, often overlooked, play a key role in AMR dissemination and prevalence [355]. To evaluate the role of integrons in AMR, we assessed the abundance of *attC* sites and *intI* genes, both of which are required for efficient integron-mediated activity. We estimated the abundance of these genes on the same contig, including those that were associated with AMR categories. Overall, we found that ARGs abundant in antibiotic-treated mice were transferred via integrons (**Figure 4.7a**). Of these, there was a significant enrichment in ARGs associated with complete (presence of *attC* and *intI* genes) integrons in mice at day 21 compared to day 0 and also when compared to the controls (**Figure 4.7a**).

Additionally, these integron-mediated ARGs (complete, gene cassettes and incomplete) were further analyzed to identify their putative genomic locations on phages or plasmids, since they are known to be carriers of integrons, thus elaborating on the method of integron-mediated AMR transmission. Interestingly, we identified several integron-mediated ARG cassettes encoded on plasmids at day 21 in the antibiotic-treated mice (**Figure 4.7b** and **Appendix B.3: Supplementary figure 4.5a**).

As we identified antibiotic-induced changes of the microbial composition, we further investigated the association of AMR-encoding integrons within the microbial community. For this, we linked the AMR-associated ‘complete’ integrons with the reconstructed genomes and found that a substantial number was associated with genomes from families including Akkermansiaceae, Lachnospiraceae and Enterobacteriaceae (**Figure 4.7c**, **Appendix B.3: Supplementary figure 4.5b**). This finding reinforces our earlier findings with respect to enriched taxa and potential ARG-mediated mechanisms of resistance through integrases.

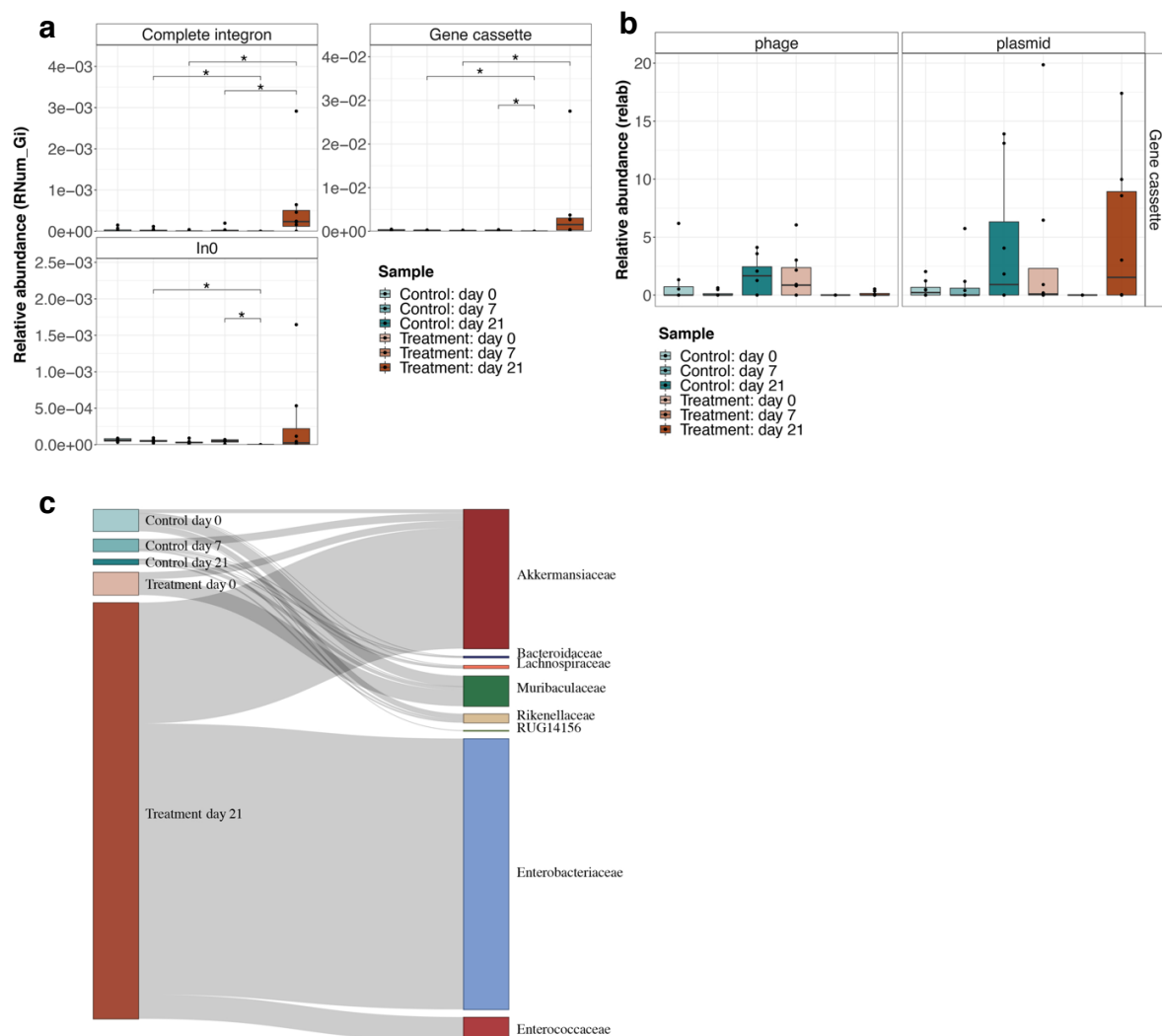


Figure 4.7: AMR-mediated via integrons in mice administered with antibiotics. a. Boxplots showing the relative abundance of AMR linked to integrons. **b.** Relative abundance of gene cassettes found on phages and plasmids across different groups and timepoints. **d.** Alluvial diagram visualizing the abundance of ‘complete’ integrons linked with observed ARGs in specific taxonomic families. The flow (grey) bars indicate the abundance of AMR-linked integrons found in each group.

4.4 Discussion

The classes and uses of antibiotics have been extensively developed since the fateful discovery of “mold juice” by Alexander Fleming [356,357]. Antibiotic consumption is increasingly driving antimicrobial resistance emergence due to both overuse and misuse [358,359]. The use including their overuse has led to unrecoverable and irreversible states of resistance [360], resulting in an “arms race” where newer and more potent molecules [361] are becoming a necessity to ward off otherwise-susceptible bacteria. Even though antibiotics may result in the

emergence of multi-resistant pathogens, their expanding use in medicine, especially as a means of modulating the gut microbiome, remains unquestionable [362]. For example, they have also been proposed as prophylactics for treating cancers [333]^[363] and modulating the gut microbiota [335,342]. Antibiotics not targeting *Clostridioides difficile* infection are commonly used within the first weeks of a fecal microbiota transplantation (FMT) as a standard therapy [364]. In other cases, antibiotics have been administered prior to studying the efficacy of FMT [365,366]. Similarly, preoperative antibiotic prophylaxis in humans is a common practice and involves a cocktail of three antibiotics (cefazolin, vancomycin, and gentamicin) [367,368]. Given the wide increase in uses, it is therefore important to understand the mechanisms and speed of the acquisition of antimicrobial resistance, especially amongst gut commensals. Here, we hypothesized that antibiotic treatment will lead to an evolution of antimicrobial resistance in the commensal gut microbiome population within a single animal generation and tested our hypothesis in a wild-type mouse cohort.

The antibiotics (ampicillin, vancomycin, metronidazole and neomycin) were chosen given their utility in several mouse models [344,345] and in varying combinations in line with some clinical procedures [346]. We observed that *Akkermansia muciniphila* was significantly enriched whilst most taxa were depleted in mice post-treatment, which is in line with other reports [369–371] where vancomycin treatment alone led to propagation of *A. muciniphila* or its dominance in the resistant commensal population. The resistance of this taxon can specifically be attributed to the presence of β -lactamase and nitroimidazole resistance genes reported by van Passel *et al.* [372]. These finding also agree with the report by Palleja *et al.* in which the authors found that species harboring β -lactam resistance genes were positively selected during antibiotic exposure [373], which is likely the case in our study since we observed higher ARG abundance at day 21 and not prior to antibiotic treatment. Alternatively, the observed ARGs could also be due to the acquisition of resistance genes possibly via lateral or horizontal gene transfer as reported previously by Guo *et al.* [374].

In addition to an overall enrichment in *A. muciniphila*, we also observed an enrichment in the functional complement of *A. muciniphila* with respect to signaling molecules, specifically quorum sensing and cyclic dinucleotide signaling. Microbial communities are characterized by emergent properties that themselves are primarily shaped by microbial interactions [375]. These interactions include intra- and extracellular signaling, such as quorum-sensing (QS) and cyclic dinucleotide sensing, as a means of adapting to internal and external stimuli [376]. Due to the paucity of external stimuli from other bacteria following antibiotic treatment it is plausible

that QS functions were selected for as a means for recalcitrant community members to ramp up signaling functions to induce antibiotic tolerance. This is in line with reports of collective antibiotic tolerance [377], contributing towards AMR, which may be mediated via QS molecules leading to bacterial resistance gene expression in a density dependent manner [377]. Furthermore, QS molecules have also been reported to regulate secondary metabolite synthesis [378]. However, we observed a depletion in genes involved in secondary metabolite synthesis following antibiotic treatment which is expected since a majority of the endogenous population is depleted. Alternatively, this phenomenon also suggests that selection for genes involved in signaling and secondary metabolite synthesis are somewhat uncoupled in our experimental murine model and, thus, are subject to different selective sweeps. Overall, our results highlight the role of key functions conferred by specific bacterial taxa in antibiotic-exposed communities and shed light on the shorter-term evolutionary processes shaping community assembly and composition.

Given the nature of the antibiotic cocktail treatment, we found several related ARGs in the metagenomes of the treated mice. More importantly, we observed significantly increased resistance against three out of the four antibiotics used in our study protocol: *aminoglycoside* (neomycin), *beta-lactam* (ampicillin) and *glycopeptide* (vancomycin). We, however, did not recover any resistance genes against nitroimidazoles (metronidazole). Additionally, we found that several taxa in mice treated with antibiotics were directly linked to the resistance categories of the antibiotics that they were treated with. This suggests that the selective pressure of the administered antibiotics may lead to real-time evolution of AMR within the gut microbiome. This is in line with a recent report by Xu *et al.* [379] albeit in a different mouse model (Balb/c), who found that treatment with single antibiotics lead to increased abundance of resistance genes. Intriguingly, they only noted considerable levels of MGE-mediated horizontal gene transfer for fosfomycin but did not observe high levels of AMR associated with integrases. In our study, we found that while phages and plasmids contributed to both HGT and AMR, integrons were also a key factor in AMR dissemination. Despite the decrease in overall diversity and number of taxa in antibiotic-treated mice, we observed significantly increased prevalence of 'complete integron'-mediated AMR. In accordance with previous reports [380], we found that gene cassettes encoding ARGs are localized and mediated via MGEs, specifically plasmids. This has further implications since mobility allows penetration and potential integration of AMR into new taxa. Importantly, in conjunction with our findings about the survival and resistance of *A. muciniphila*, we found that several integrons associated with AMR were directly linked to this taxon.

Concerted knowledge has concentrated on pathogenic bacteria, and lately the emergence of antimicrobial resistance. Although the genome of *A. muciniphila* is thought to be plastic [374], our analyses emphasize the role of integrons in mediating AMR in commensals, including within this taxon. Given the association of *A. muciniphila* with Parkinson's and other chronic diseases, our findings highlight the need to understand the role of integrons in mediating AMR within and beyond this taxon. Taken together, our data show that AMR is relevant to studies involving antibiotic-treatment especially within the commensal gut microbiome population. Our study is built on a systemic, and longitudinal design to understand the stage at which the resistance genes are acquired following antibiotic treatment. Simultaneously, it is unclear whether the acquired ARGs are expressed, requiring the need for both experimental and cross-validation using methods such as metatranscriptomics to observe gene expression. Overall, we highlight the need for understanding the real-time evolution of AMR in microbiome research, including functional and evolutionary consequences of integron-mediated AMR.

Chapter 5. Mobilome-mediated segregation of antimicrobial resistance in a wastewater treatment plant

This chapter is based on the following publication submitted for peer-review:

Laura de Nies, Susheel Bhanu Busi, Benoit Kunath, Patrick May and Paul Wilmes (2021). Mobilome-driven segregation of the resistome in biological wastewater treatment. *eLife* **in review** [Appendix A.5]

5.1 Introduction

Throughout human history, bacterial infections have been a major cause of both disease and mortality [381]. The discovery as well as the subsequent development and medical use of antibiotics have provided effective treatment options which limited the development and spread of bacterial pathogens. However, the use of antibiotics has exacerbated the emergence of antimicrobial resistance (AMR) in both commensal and pathogenic bacteria [2]. As a result, AMR, as the "silent pandemic", has become a prevalent threat to human health [5,6,382].

From a public health perspective, biological wastewater treatment plants (BWWTPs) are considered hotspots of AMR due to the convergence of antibiotics with resistant, potentially pathogenic microorganisms originating from both the general population as well as agriculture and healthcare services [129,130] (**section 1.4.3**). Additionally, the mobilization of antimicrobial resistance genes (ARGs) through rampant horizontal gene transfer (HGT) promotes the dissemination of AMR within the BWWTP microbial community [16] (**section 1.4.3**). Therefore, BWWTPs represent an environment exceptionally suited for the evolution and subsequent spread of AMR [383,384]. To date, more than 32 studies have documented the role of BWWTPs as key reservoirs of AMR [385]. Furthermore, BWWTPs generally do not contain the necessary infrastructure to remove either ARGs or resistant bacteria, which are released into the receiving water via the effluent, promoting its spread in the environment at large [130]. Most often these are surface water bodies such as rivers, which contribute to the further dissemination of AMR and resistant bacteria among environmental microorganisms [386]. Acquired resistance may in turn be carried over to humans and animals using these water resources. In fact, there is strong evidence suggesting that ARGs from environmental bacteria can be taken up by human-associated and pathogenic bacteria [59,148]. From an epidemiological and surveillance perspective, BWWTPs also provide samples representative of entire populations [177]. As such, BWWTPs have recently been crucial for the monitoring of SARS-CoV-2 within the human population [387]. Overall, to increase our understanding of the dissemination of AMR and the underlying mechanisms as well as its general prevalence, it is necessary to map the resistome of various environments starting with biological BWWTPs because it is critical to unravel the extent to which they act as reservoirs for the dissemination of antimicrobial resistance genes (ARGs) to bacterial pathogens. Moreover, understanding the community-level overviews of the ARG potential and its expression, coupled with population-level linking, including to pathogens, may allow for efficient monitoring of pathogenic and AMR potential with broad impacts on human health.

The conditions such as the presence of resistance genes [388], and sub-inhibitory antibiotic selection pressure from various sources [16] facilitate HGT of ARGs into new hosts through the mobilome, i.e. mobile genetic elements (MGEs). Acquisition of ARGs via MGEs primarily occurs through two mechanisms: conjugation or transduction [20] (**section 1.2**). Of these mechanisms, conjugation is often thought to have the greatest influence on the dissemination of ARGs, while transduction is deemed less important [16]. In general terms, studies concerning AMR and its dissemination focus either on phage [389,390] or plasmids solely [391]. Alternatively, the two are treated collectively [130,392] without a comprehensive comparative analysis. This circumstance has created a knowledge gap whereby the contributions of plasmids and phages as independent entities to AMR transmission within complex communities, such as those found in biological BWWTPs, is largely unknown.

To shed light on the evolution, dissemination and potential segregation of AMR within MGEs in a WTP microbial community, we leveraged longitudinal meta-omics data (metagenomics, metatranscriptomics and metaproteomics). Samples collected for 51 consecutive weeks over a period of 1.5 years, were used to characterize the resistome. We found that several bacterial orders such as Acidimicrobiales, Burkholderiales and Pseudomonadales were associated with 29 AMR categories across all timepoints. Our longitudinal analysis demonstrated that MGEs are important drivers of AMR dissemination within BWWTPs and that assessing the activity of the ARGs is critical for understanding the underlying mechanisms. More importantly, we reveal that MGEs, i.e. plasmidomes and phageomes, contribute differentially to AMR dissemination. Furthermore, we observed this phenomenon in clinically-relevant taxa such as the ESKAPEE pathogens [393], for which plasmids and phages were exclusively associated with specific ARGs. Collectively, our data suggest that BWWTPs are critical reservoirs of AMR which show clear evidence for the segregation of distinct ARGs within MGEs especially in complex microbial communities. In general, we believe that these findings may provide crucial insights into the segregation of the resistome via the mobilome in any and all reservoirs of AMR, including but not limited to animals, humans, and other environmental systems.

5.2 Methods

5.2.1 Sampling and biomolecular extraction

From within the anoxic tank of the Schifflange municipal biological wastewater treatment plant (located in Esch-sur-Alzette, Luxembourg; 49° 30' 48.29" N; 6° 1' 4.53" E) individual floating sludge islets were sampled according to previous described protocols [394]. Sampling was performed starting on 21-03-2012 till 03-05-2012 in approximately one-week intervals resulting in a total of 51 samples. DNA, RNA and proteins were extracted from the samples in a sequential co-isolation procedure as previously described [395].

5.2.2 Sequencing and data processing for metagenomics and metatranscriptomics

Paired-end libraries were generated for metagenomics with the AMPure XP/Size Select Buffer Protocol following a size selection step recommended by the standard protocol. Libraries for metatranscriptomics were prepared from RNA after washing stored extractions with ethanol and depletion of rRNAs with the Ribo-Zero Meta-Bacteria rRNA Removal Kit (Epicenter). Subsequently, the ScriptSeq v2 RNA-seq library preparation kit (Epicenter) was used for cDNA library preparation, followed by sequencing on an Illumina Genome Analyses Iix instrument with 100-bps paired-end protocol. Processing and assembly of metagenomic and metatranscriptomic reads was done using the Integrated Meta-omic Pipeline [231] (IMP v1.3; available at <https://r3lab.uni.lu/web/imp/>). For the IMP processing, Illumina Truseq2 adapters were trimmed, and reads of human origin were filtered out, followed by a de novo assembly with MEGAHIT [276] v1.0.6. Both metagenomic and metatranscriptomic reads were coassembled to increase contiguity of the assemblies [231].

5.2.3 Identification of antimicrobial resistance genes and association with mobile genetic elements

The assembled contigs from IMP were used as input for PathoFact [396], for the prediction of antimicrobial resistance genes, and to annotate MGEs. ARGs were further collapsed into their respective AMR categories, as identified by PathoFact in accordance with those provided by the Comprehensive Antibiotic Resistance Database (CARD) [43]. Thereafter, the raw read counts per ORF, as given by PathoFact, were determined with FeatureCounts [397]. The relative abundance of the ARGs was calculated using the RNum_Gi method described by Hu *et al.* [233]. This method was done using the BAM files generated by mapping using samtools

[398], both for the metagenomic and metatranscriptomic reads independently, to extract gene copy number and transcriptome expression respectively, per sample.

Identified ARGs and their categories were further linked to associated bacterial taxonomies using the taxonomic classification system Kraken2 [399]. Kraken2 was run on the contigs using the `maxikraken2_1903_140GB` (March 2019, 140GB) (https://lomanlab.github.io/mockcommunity/mc_databases.html) database [399]. Furthermore, utilizing PathoFact, ARGs were linked to predicted mobile genetic elements (i.e. plasmids and phages) to identify probable transmission of AMR between taxa. Specifically, to link both the MGEs and the taxonomy to the ARGs, we mapped the genes to assembled contigs. By considering all different predictions of MGEs, a final classification was made based on the genomic contexts of the ARGs encoded on plasmids, phages or chromosomes, including classification of those that could not be resolved (ambiguous). The ARGs that could not be assigned to either the MGEs or bacterial chromosomes were further referred to as *unclassified* genomic elements.

5.2.4 Metaproteomic sequencing and data analyses

Raw mass spectrometry files were converted to MGF format using MSconvert [400] with default parameters. Metaproteomic search was performed using SearchGUI / PeptideShaker [401] for each time point. To generate the databases, each predicted protein sequence file was concatenated with the cRAP database of contaminants (*common Repository of Adventitious Proteins*, v 2012.01.01; The Global Proteome Machine) and with the human UniProtKB Reference Proteome [402]. In addition, reversed sequences of all protein entries were concatenated to the databases for the estimation of false discovery rates (FDRs). The search was performed using SearchGUI-3.3.20 [403] with the X!Tandem [404], MS-GF+ [405] and Comet [406] search engines and the following parameters. Trypsin was used as the digestion enzyme and a maximum of two missed cleavages was allowed. The tolerance levels for identification were 10 ppm for MS1 and 15ppm for MS2. Carbamidomethylation of cysteine residues was set as a fixed modification and oxidation of methionine's was allowed as variable modification. Peptides with length between 7 and 60 amino acids and with a charge state composed between +2 and +4 were considered for identification. The results from SearchGUI were merged using PeptideShaker-1.16.45 [401] and all identifications were filtered in order to achieve a peptide and protein FDR of 1%.

Each predicted protein sequence corresponds with the predicted ORFs generated by the Prodigal (version 2.6.3) [204] predictions included in PathoFact. As such predicted protein sequences matched the ARG annotation of the ORFs as given by PathoFact.

5.2.5 Multi-omic integration

To further improve upon the understanding of the AMR expression and assess its stability across time, we estimated the normalized protein index (NPI) per gene, by integrating the multi-omic data. To estimate the NPI, we first normalized the metatranscriptomic abundance based on per gene copy numbers obtained via the metagenomic abundance:

$$NPI = \frac{N_{metaproteome}}{N_{metatranscriptome} / N_{metagenome}}$$

This, the normalized expression of genes, yields the per copy expression of ARGs within each AMR category. Subsequently, the normalized expression was used to standardize the metaproteomic abundances for those genes where the necessary data was available.

5.2.6 MGE partition assessment

To assess the segregation of MGEs through AMR we determined niche regions and overlap using the nicheROVER R package [407]. nicheROVER uses Bayesian methods to calculate niche regions and pairwise niche overlap using multidimensional niche indicator data (i.e. stable isotopes, environmental variables). As such, using AMR as the indicator data, we extended the application of nicheROVER to calculate the probability for the size of the niche area of one MGE inside that of the other, and vice versa. We calculated the segregation size estimate for each MGE and additionally generated the posterior distributions of μ (population mean) for each AMR category in all omics. We further computed the niche overlap estimates between MGEs with a 95% confidence interval over 10 000 iterations.

5.2.7 Data analysis

Figures for the study including visualizations derived from the taxonomic and functional analyses were created using version 3.6 of the R statistical software package [408]. A paired two-way ANOVA (Analysis of Variance) within the *nlme* package was used for identifying statistically significant differences for the AMR and taxonomic analyses. Tripartite and Bipartite networks were generated using the *SpiecEasi* [409] R package where a weighted adjacency matrix was generated using the Meinhausen and Buhlmann (*mb*) algorithm, with a λ of

40, and lambda minimum ratio at 0.001. The analyses were bootstrapped with n=999 to avoid overfitting, autocorrelations and false network associations. The network was further refined, selecting for positive edges, with a degree greater than the mean-degree of the initial network. The *igraph* [410] package was used in R to render the graphics for the network.

5.3 Results

5.3.1 Longitudinal assessment of the resistome within a WWTP

To characterize the BWWTP resistome, we sampled a municipal BWWTP on a weekly basis over a 1.5 year period (ranging from 21-03-2011 to 03-05-2012) [394,411]. Utilizing the PathoFact pipeline we resolved the BWWTP resistome. This analysis revealed the presence of 29 different categories of AMR within the BWWTP. Subsequent longitudinal analyses highlighted enrichments in aminoglycoside, beta-lactam and multidrug resistance genes (**Figure 5.1a**). Concomitantly, we observed specific shifts in the AMR profiles over time. For example, a shift at two timepoints (13-05-2011, 08-02-2012) highlighted a steep increase in resistance genes corresponding to glycopeptide resistance. Other AMR categories, such as diaminopyrimidine resistance, exhibited a less drastic but more fluid change in longitudinal abundance observable over multiple timepoints.

Additionally, AMR categories were found to persist variable over time (**Figure 5.1b**). A core group of 15 AMR categories in total were identified and found to be present across the 1.5 year sampling period. These included aminoglycoside, beta-lactam and multidrug resistance genes, which contributed the most to the pool of ARGs. A further six (aminocoumarin, aminoglycoside:aminocoumarin, elfamycin, nucleoside, triclosan and unclassified) AMR categories were found to be prevalent (>75% of all timepoints), while another three AMR categories were moderately (50 - 75% of all timepoints) present over time (**Figure 5.1b**). Five other categories were rarely present within the BWWTP, with resistance corresponding to acridine dye only present at six of the timepoints. Altogether, this emphasized that the BWWTP resistome varies over time, substantiating the requirement for a longitudinal analysis to obtain an accurate overview of the community's overall resistome.

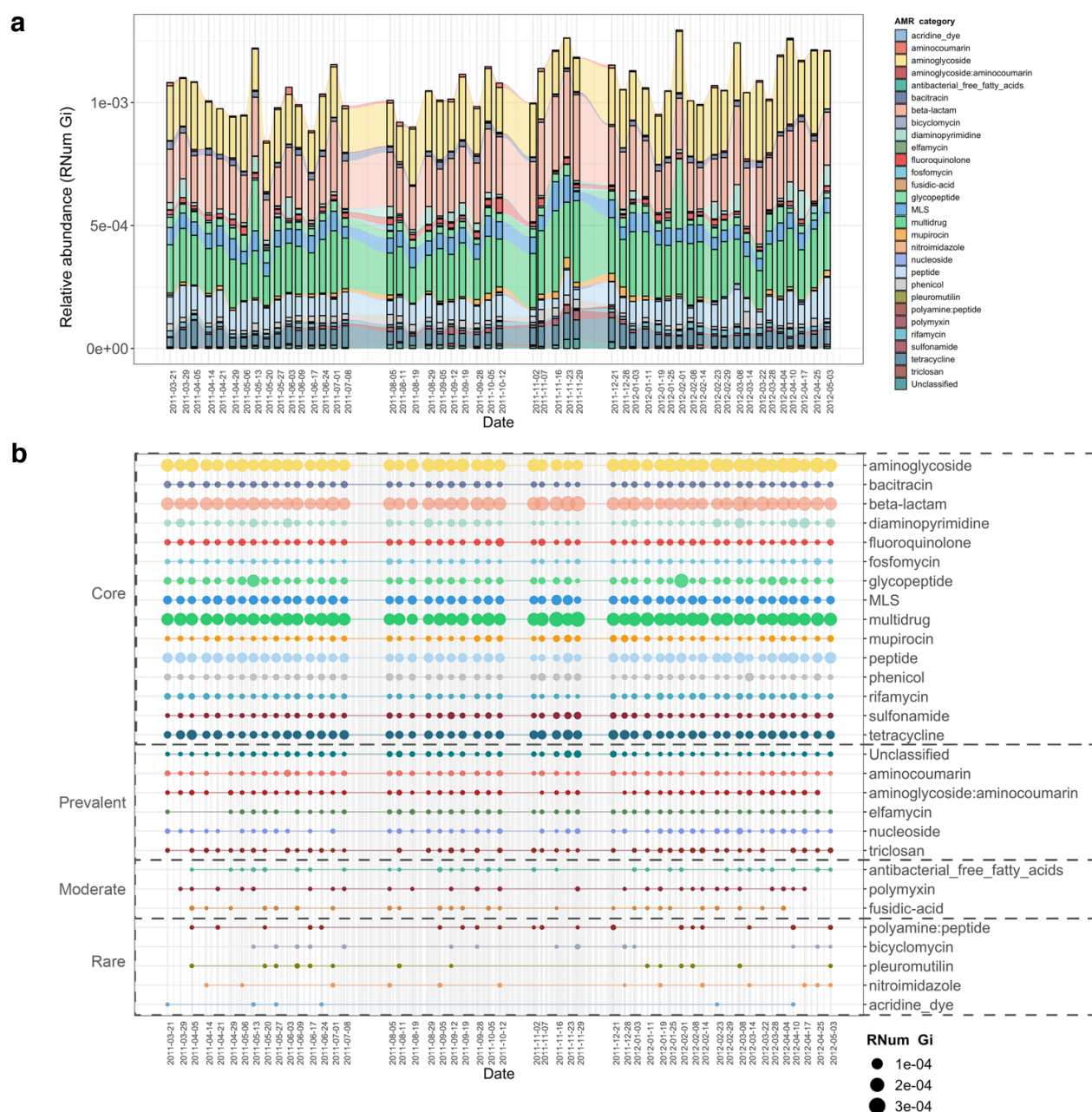


Figure 5.1: Longitudinal metagenomic assessment of AMR. a. ARG relative abundances over time within the BWTP. **b.** ARG categories at various timepoints categorized in 4 distinct groups based on presence/absence: Core (all timepoints), Prevalent (>75% of timepoints), Moderate (50-75% of timepoints) and Rare (<50% of all timepoints).

Although the data thus far provided a clear overview of the BWTP from a metagenomic perspective, it did not provide any information regarding AMR expression. We therefore utilized the corresponding metatranscriptomic dataset to investigate the expression of identified ARGs

and monitor their changes, within the BWWTP, over time. In contrast to the metagenomic data, we observed a difference in AMR expression levels for several categories. Aminoglycoside, beta-lactam, and multidrug resistance identified at high levels in metagenomic information were also highly expressed within the BWWTP (**Figure 5.2**). However, peptide resistance demonstrated the highest expression levels of all the AMR categories. We further investigated which ARG subtypes contributed to the identified peptide resistance category and found that ~90% of the expressed peptide resistance was directly contributed by a single resistance gene, *YojI* (**Appendix B.4: Supplementary figure 5.1**), typically associated with resistance to microcins [412] a potential adaptive strategy amongst the microbial populations in the BWWTP against these specific stressors.

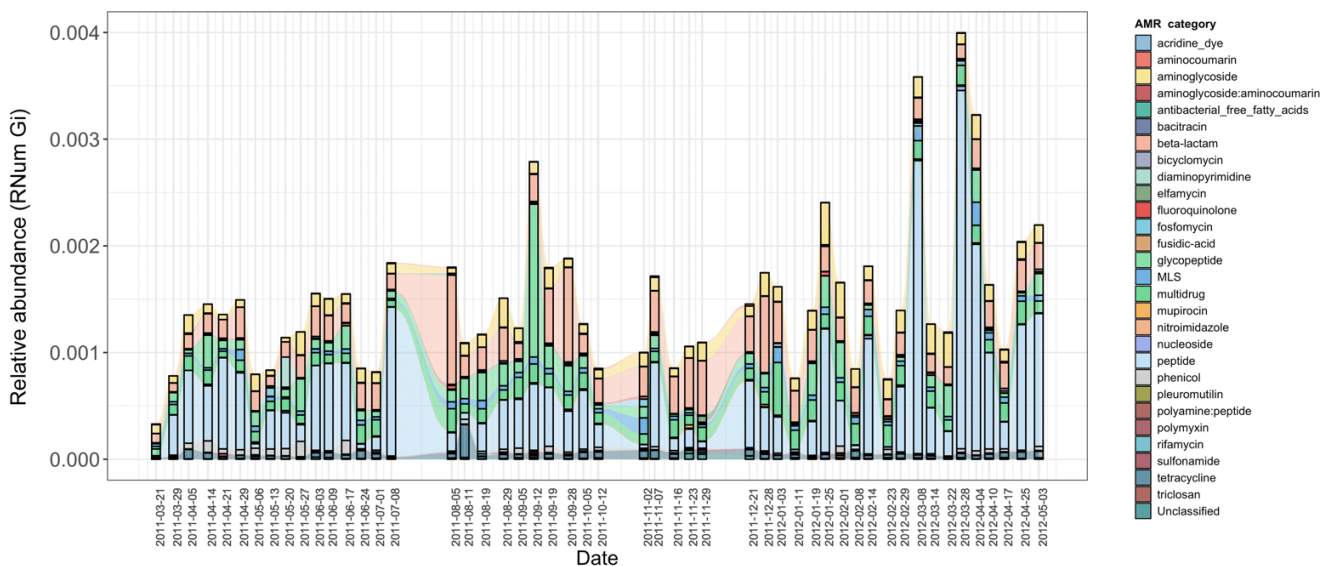


Figure 5.2: Longitudinal metatranscriptomic assessment of AMR. Relative abundance levels of expressed AMR categories over time within the BWWTP

5.3.2 Microbial populations and co-occurrence patterns of AMR

Based on the previously identified microbial community [394], we hypothesized that the abundant and prevalent bacterial orders such as Acidimicrobiales were major contributors to the abundance in ARGs observable via metagenomics. To further investigate the contribution to AMR by the distinct microbial populations, we linked ARGs to the contig-based taxonomic annotations of the assemblies. Herein, we identified a wide variety of taxonomic orders contributing to AMR, with multiple orders often contributing to the same resistance categories (**Appendix B.4: Supplementary figure 5.2**). Overall, taxa belonging to Acidimicrobiales, followed by Burkholderiales, were found to encode most of the ARGs (**Figure 5.3a**).

Additionally, the abundance of ARGs linked to taxonomy varied over time. This was most noticeable during a five-week period (autumn: 02-11-2011 to 29-11-2011), where a decrease in abundance in ARGs linked to Acidimicrobiales and Bacteroidales was observed coinciding with an increase in ARG abundance in Pseudomonadales and Lactobacillales.

Since the family Acidimicrobiales was found to be linked to the highest abundance in ARGs, we further resolved the taxonomic affiliation and identified the species *Candidatus* Microthrix parvicella (hereafter known as *M. parvicella*) to be the main contributor to AMR. *M. parvicella* was previously found to dominate this microbial community [411] and is a well-characterized bacterium commonly occurring in the BWWTP [413]. Overall, aminoglycoside, beta-lactam, multidrug and peptide resistance were found to be abundant in this species (**Figure 5.3b**), with aminoglycoside resistance demonstrating the highest expression levels as confirmed through metatranscriptomic analysis (**Figure 5.3c**). Although it was not surprising to find a high abundance of ARGs linked to this species, the longitudinal variation in the abundances of these ARGs was nevertheless surprising (**Figure 5.3b**). Furthermore, coupled to a decrease in the abundance of *M. parvicella* itself [411], we observed an almost complete decrease in ARGs at two timepoints (23-11-2011 and 29-11-2011). However, the *M. parvicella* population recovered to levels resembling the earlier timepoints in conjunction with the abundances in ARGs towards the end of the sampling period (**Figure 5.3a-b**), underlining their overall contribution to AMR within this BWWTP. Alternatively, it is plausible that the dominance of *M. parvicella* is attributable to the encoded ARGs, which in turn, may confer a fitness advantage.

In order to determine whether the abundances in ARGs may be directly associated with the community composition and population sizes over time, co-occurrence patterns between ARG subtypes and taxa (genus level) were explored using the metagenomic data. Bipartite network analyses (**Figure 5.3c**) demonstrated that ARGs, within or across ARG types and microbial taxa, showed clear and distinct co-occurrence patterns within the BWWTP. These patterns indicated a strong segregation of distinct, taxa specific ARG subtypes within the BWWTP community over time. One clear example was that of *M. parvicella* which encoded different aminoglycoside resistance genes (**Appendix B.4: Supplementary figure 5.3**). Thus, the abundance of this bacterium along with the aminoglycoside ARGs were highly correlated.

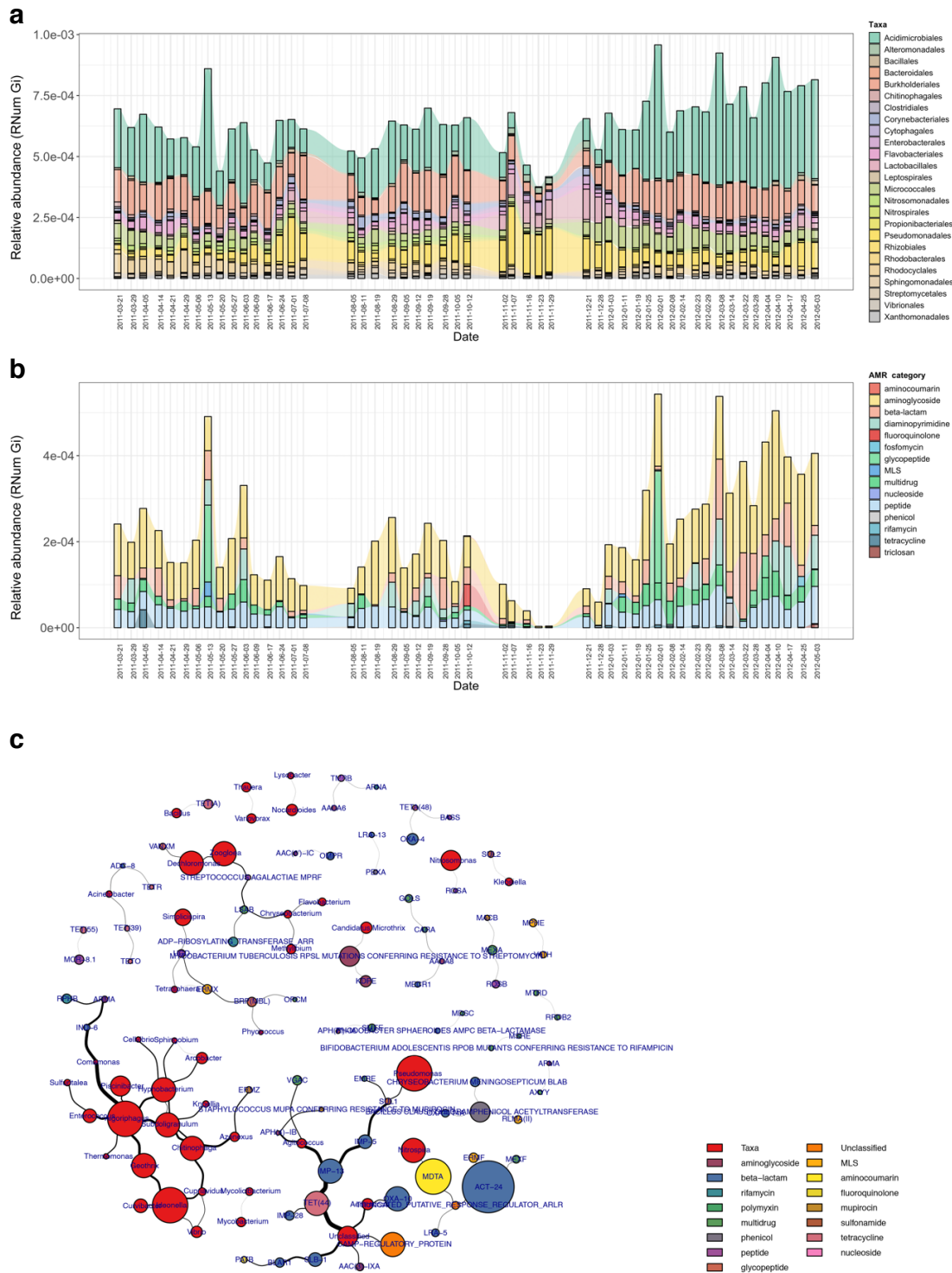


Figure 5.3: Microbial population-linked AMR. **a.** Longitudinal ARG relative abundance levels linked to their corresponding microbial taxa (order level). **b.** Relative abundance of ARG categories linked to *Candidatus* *Microthrix parvicella*. **c.** Bi-partite network depicting co-occurrence patterns of individual ARGs and microbial taxa on genus level.

5.3.3 Monitoring pathogenic microorganisms within BWWTPs

In conjunction with the families observed within BWWTPs, we also found that certain ESKAPEE pathogens [393], such as *Klebsiella* spp. and *Pseudomonas* spp., demonstrated co-occurring patterns with ARGs (**Figure 5.3c**).

As previously mentioned, BWWTPs represent a collection of potentially pathogenic microorganisms originating from, among others, the human population. Moreover, evidence suggests that ARGs from environmental and commensal bacteria can spread to pathogenic bacteria through HGT [20]. Therefore, we assessed the acquisition and dissemination of AMR in the extended priority list of pathogens (Table 5.1), characterized as such by the WHO [414], using both metagenomics and metatranscriptomics.

Table 5.1: WHO priority list for research and development of new antibiotics for antibiotics-resistant bacteria [414].

Bacteria	Priority	Organism detected	Resistance detected
<i>Acinetobacter baumannii</i>	Critical	+	+
<i>Pseudomonas aeruginosa</i>	Critical	+	+
<i>Enterobacteriaceae</i>	Critical	+	+
<i>Enterococcus faecium</i>	high	+	+
<i>Staphylococcus aureus</i>	high	+	+
<i>Helicobacter pylori</i>	high	+	+
<i>Campylobacter</i> spp	high	+	-
<i>Salmonella</i> spp	high	+	+
<i>Neisseria gonorrhoeae</i>	high	+	-
<i>Streptococcus pneumoniae</i>	medium	+	+
<i>Haemophilus influenzae</i>	medium	+	-
<i>Shigella</i> spp	medium	+	+

Of the identified pathogens (Table 5.1), we found that *Pseudomonas aeruginosa*, both encoded and expressed the highest abundance of ARGs, followed by *Acinetobacter baumannii*, over

time within the BWWTP (**Figure 5.4**). Moreover, an increase in ARG abundance and expression was observed in *Pseudomonas aeruginosa* during the time period, during which the otherwise dominant *M. parvicella* demonstrated reduced abundance (**Figure 5.3a**, **Figure 5.4**).

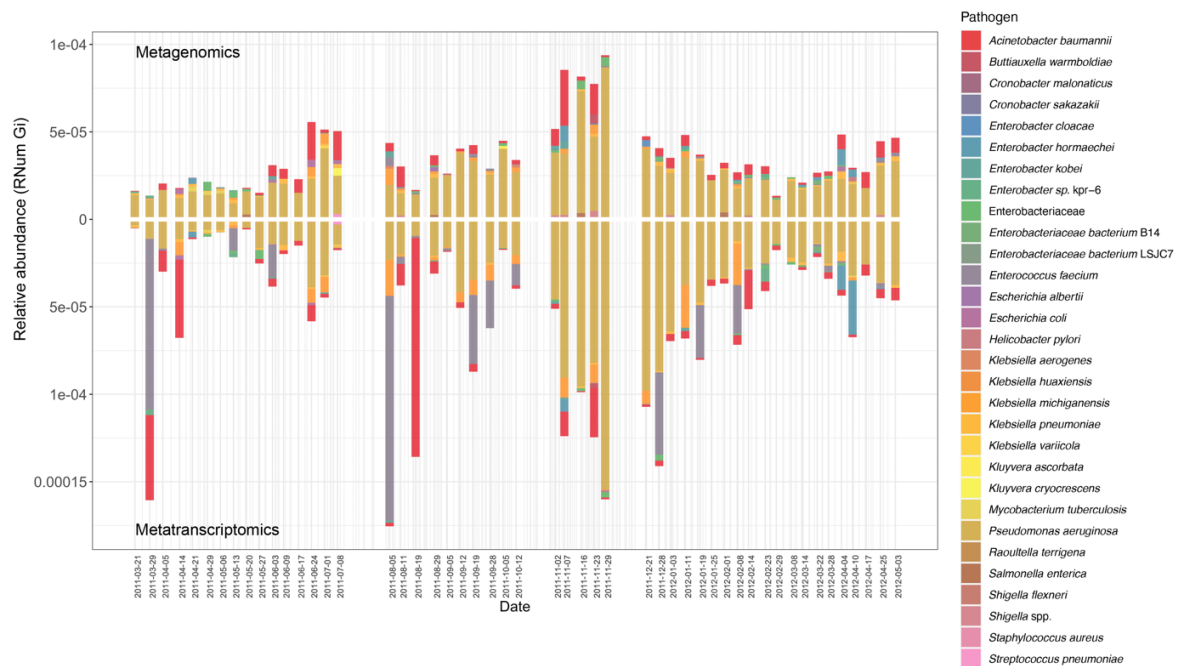


Figure 5.4: Assessment of AMR associated with clinical pathogens. ARG relative abundance encoded and expressed by clinical pathogens over time within the BWWTP.

5.3.4 Differential transmission of antimicrobial resistance via mobile genetic elements

As previously described [182,415], the mobilome is a major contributor to the dissemination of AMR within a microbial community. Consequently, to understand (i) the role of MGE-mediated AMR transfer within the BWWTP, and (ii) to identify differential contribution of the mobilome to the dissemination of AMR, we identified both plasmids and phages within the metagenome and linked these to the respective ARGs. Overall, we found that plasmids contributed to an average of 10.8% of all ARGs, while phage contributed to an average of 6.8% of all resistance genes, confirming the general hypothesis that conjugation has the greatest influence on the dissemination of ARGs [20]. This phenomenon, however, varied across time within the BWWTP (**Figure 5.5a**).

When investigating the dissemination of AMR via MGEs, most reports typically focus on either phages or plasmids individually, or both as collective contributors to transmission [416]. To date and to our knowledge, the respective contributions of phage and plasmid to AMR transmission

have not been subjected to a comprehensive comparative analysis. To facilitate a systematic, comparative view of MGE-mediated AMR, we assessed the segregation of MGEs with respect to AMR and found that phages and plasmids contributed differentially to AMR. Specifically, we found a significant difference in six AMR categories when comparing ARGs encoded by phages and plasmids (**Figure 5.5b**). Aminoglycoside, bacitracin, MLS (i.e. macrolide, lincosamide and streptogramin) and sulfonamide resistance were found to be primarily encoded by plasmids, whereas fosfomycin and peptide resistance were found to be associated with phages.

To further understand AMR in relation to the community dynamics, we investigated the abundance and segregation of the above-mentioned significant resistance categories at different timepoints within the BWWTP. We observed ARG abundances varied over time both in phages (**Figure 5.5c**) as well as plasmids (**Figure 5.5d**). For instance, the abundance in aminoglycoside and sulfonamide resistance, which was encoded primarily by plasmids (**Appendix B.4: Supplementary figure 5.4a**), fluctuated widely over time in both phages and plasmids (**Figure 5.5c**). Additionally, plasmid-mediated sulfonamide resistance was reduced at 23-11-2011, followed by its highest abundance a week later (20-11-2011), while subsequently again decreasing. Similarly, in line with the above observations, fosfomycin and peptide resistance genes, while segregating within phages, demonstrated significant fluctuations over time (**Figure 5.5d**). In addition to the metagenome, we also contextualized the localization of the expressed ARGs within MGEs based on the metatranscriptomic information. Specifically, we found that plasmids demonstrated a significantly increased expression of aminoglycoside along with bacitracin and sulfonamide resistance genes, while the expression of glycopeptide, mupirocin and peptide resistance genes were primarily enriched in phages (**Figure 5.6a**). These observations pertaining to plasmid-mediated AMR were in line with the metagenomic findings (**Figure 5.5b**). Only peptide resistance was observed to be expressed via phages in contrast to the differential enrichment of fosfomycin resistance observable in the metagenomic data.

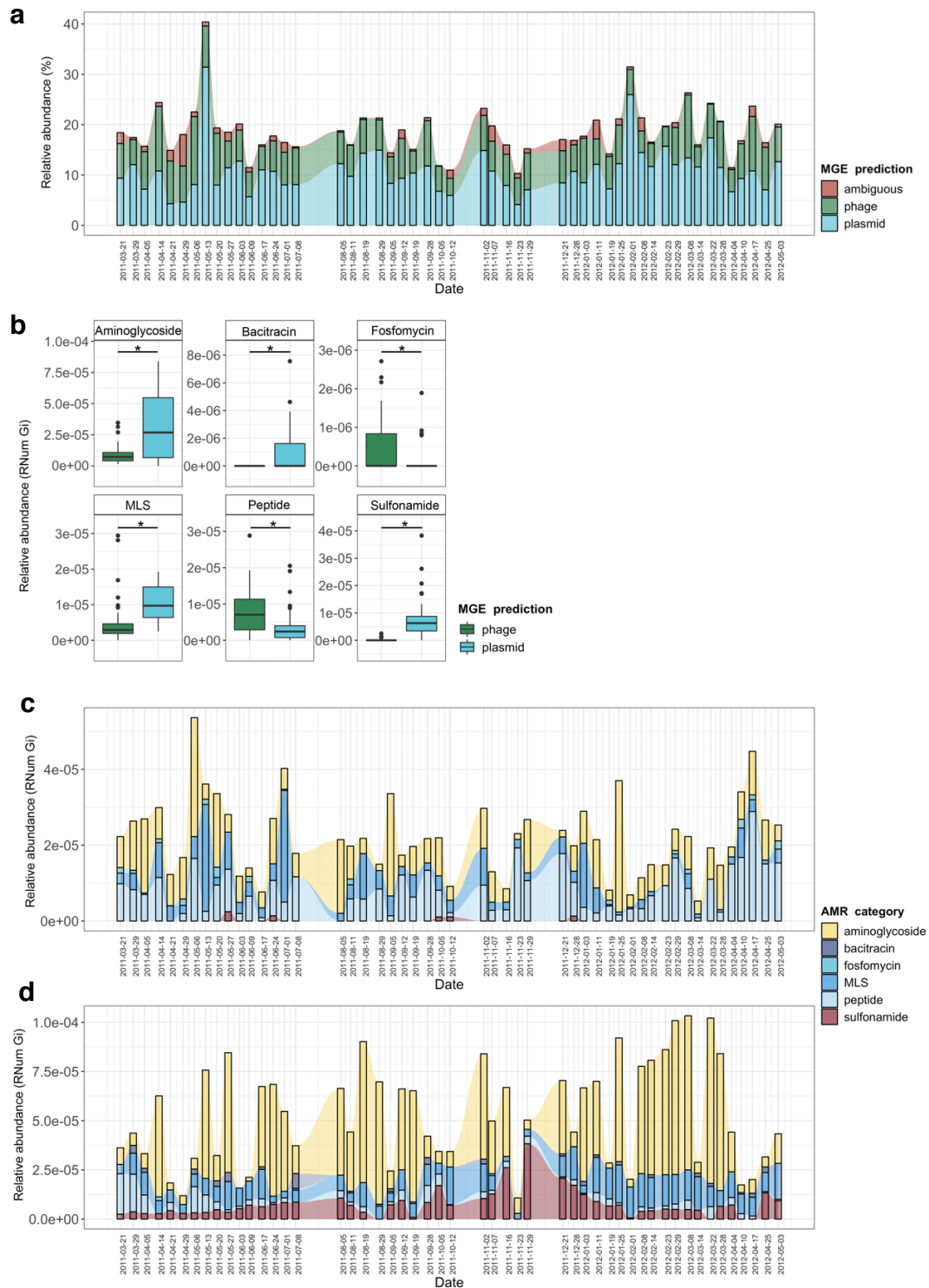


Figure 4.5: MGE-derived AMR within the BWTP resistome. a. Overall relative abundance of MGEs encoding ARGs. **b.** Boxplots depicting significant ($adj.p < 0.05$, Two-way ANOVA) differential abundances of ARGs encoded by plasmids vs phages. **c.** Relative abundance of the 6 significantly different AMR categories encoded on phages over time. **d.** Relative abundance of the 6 significantly different AMR categories encoded on plasmids over time.

5.3.5 Taxonomic affiliations of MGE-derived resistance genes

When assessing the differential contributions of MGEs to AMR, we found congruency between plasmids and phages to the AMR categories and taxonomic affiliations (**Figure 5.6b**). For example, in the metagenomic data MGEs (phage and plasmid) were predominantly associated with the same AMR category and subsequently the same taxa. However, some exceptions were observed with specific taxa associated with AMR either through plasmids or phages. For instance, MLS resistance in Bacteroidales and Nostocales was mediated solely through plasmids, whereas the same resistance category was mediated by phage in Bifidobacteriales, indicating a mechanistic basis for the segregation of AMR between taxa and MGEs.

As most bacteria harbor MGEs, we queried whether the MGE-mediated AMR categories were linked to the abundance of some of the earlier reported taxa. Interestingly, we found that peptide resistance encoded by *M. parvicella* was solely associated with phages, while aminoglycoside resistance was primarily correlated with plasmids (**Appendix B.4: Supplementary figure 5.4b**). Other highly abundant taxa such as *Pseudomonas* and *Comamonas* (**Appendix B.4: Supplementary figure 5.4c-d**), on the other hand, were correlated with sulfonamide resistance in addition to aminoglycoside resistance encoded on plasmids (**Figure 5.6b**). This was further reflected within the metatranscriptome data where in taxa such as Acidimicrobiales the expression levels of aminoglycoside resistance were solely associated with plasmids (**Appendix B.4: Supplementary figure 5.5a**). Additionally, in the Burkholderiales family, peptide and glycopeptide resistance were found to be expressed through phages (**Appendix B.4: Supplementary figure 5.5b**).

We also found a clear segregation of the mobilome with respect to individual pathogens in the metagenome. Interestingly, plasmids were exclusively associated with AMR in six out of the fourteen relevant taxa (**Figure 5.6c**). These included *Streptococcus pneumoniae*, *Staphylococcus aureus*, *Shigella flexneri*, *Klebsiella pneumoniae*, *Enterobacter kobei* and *Enterobacter hormaechei*. Furthermore, the plasmids were also associated with conferring peptide, multidrug, MLS, beta-lactam, fluoroquinolone, bacitracin, aminoglycoside, aminoglycoside:aminocumarin and sulfonamide resistance. Phages were exclusively associated with glycopeptide and aminoglycoside resistance in *Salmonella enterica*. Overall, our results revealed for the first time the key segregation patterns of AMR via the mobilome in taxa that are of relevance to human health and disease. Moreover, substantiating the metagenomic data, the pathogenic bacteria *S. pneumoniae*, *S. aureus*, *K. pneumoniae*, *E.*

kobei and *E. hormaeche* were found to express ARGs solely associated with plasmids (**Figure 5.6c**). Collectively, these findings represent an imminent threat to global health due to their potential for dissemination across reservoirs.

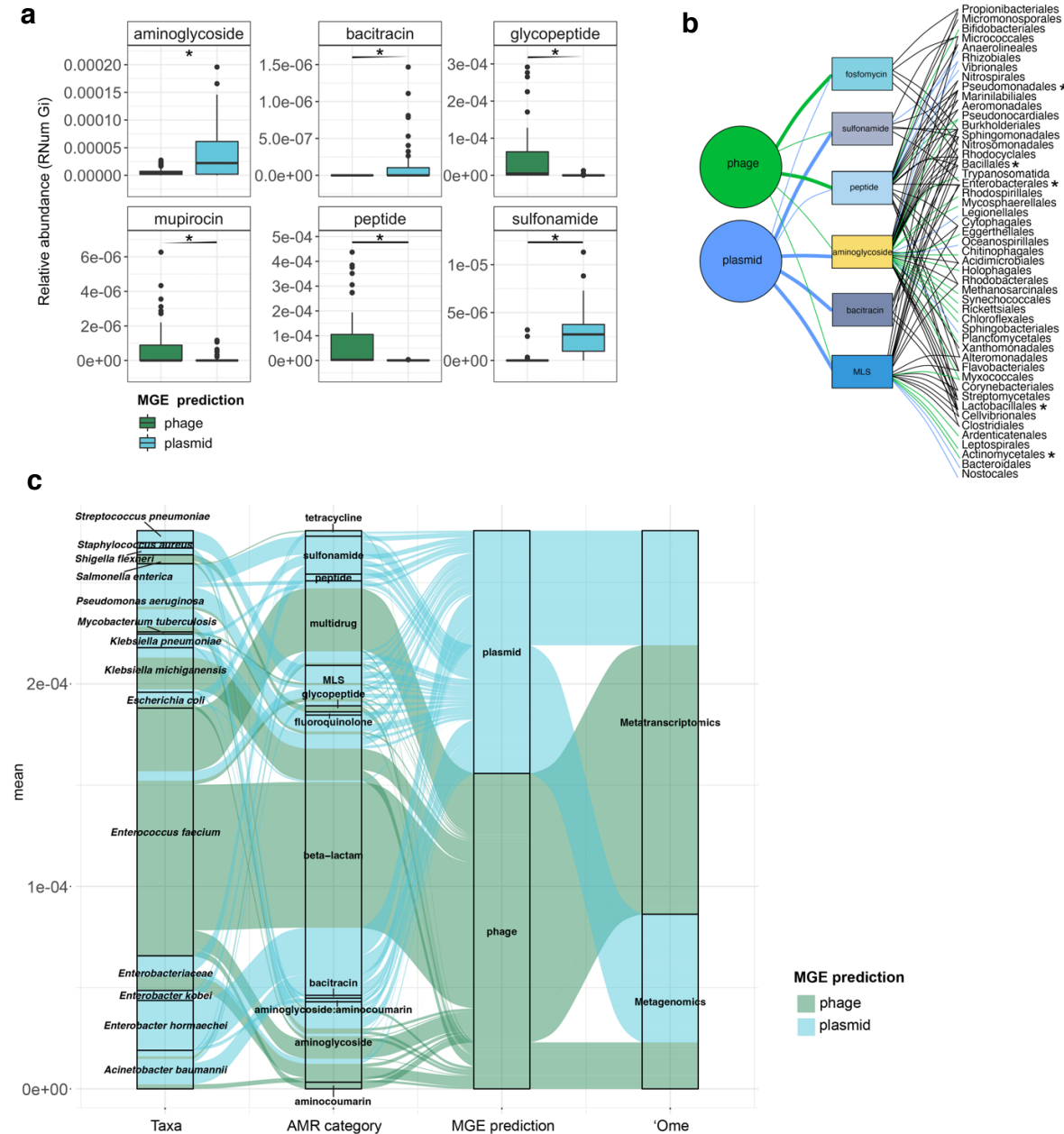


Figure 5.6: Taxonomic affiliations of MGE-derived resistance genes. **a.** Boxplot depicting significant differential abundance ($adj.p < 0.05$, Two-way ANOVA) of ARGs expressed in plasmids vs phages. **b.** Tripartite network assessing the association of MGE-derived ARGs with the microbial taxa, Thickness of the lines representing potential niche-partitioning of the ARG category to one MGE over the other. Color of the line representing which MGE the ARG is linked to: green (phage), blue (plasmid) or black (both phage and plasmid) Asterisk (*) denotes taxonomic orders which include known clinical pathogens. **c.**

Alluvial plot depicting relative abundance of MGE-derived ARGs encoded (metagenome) and/or expressed (metatranscriptome) by clinical pathogens.

5.3.6 Metaproteomic validation of AMR abundance and expression

In order to validate our findings with the expression (metatranscriptomic) analyses on the BWWTP, we further used the corresponding metaproteomic data to offer complementary information at the protein level. Similar to the metagenome data we found protein expression linked to aminoglycoside, beta-lactam and multidrug resistance, over time within the BWWTP (**Appendix B.4: Supplementary figure 5.6**). Proteins linked to multidrug resistance especially were found to increase over time.

To further improve upon the understanding of the AMR expression and assess its stability across the time, we estimated the normalized protein index (NPI) per gene, as discussed in the Methods, by integrating all of the multi-omic data. The estimated NPI demonstrated stable levels of aminoglycoside and multidrug resistance within the BWWTP (**Figure 5.7a**). Specifically, proteins conferring multidrug resistance were found to increase over time, which is in line with the gene- and expression-level observations. Furthermore, we contextualized the normalized proteins conferring AMR to their localization on MGEs. We identified five resistance categories, i.e. aminoglycoside, beta-lactam, sulfonamide, multidrug and tetracycline resistance, to be expressed through MGEs (**Figure 5.7b**). Of these categories we found that aminoglycoside resistance, in concordance with the gene and expression levels, was significantly higher mediated through plasmids compared to phages. We further found that the MGE-mediated AMR categories were associated with specific microbial taxa. with plasmid-mediated aminoglycoside resistance found to be strongly associated with the previously mentioned *M. parvicella* (**Figure 5.7b**). On the other hand, we did not identify any peptides associated with the ESKAPEE pathogens via metaproteomics.

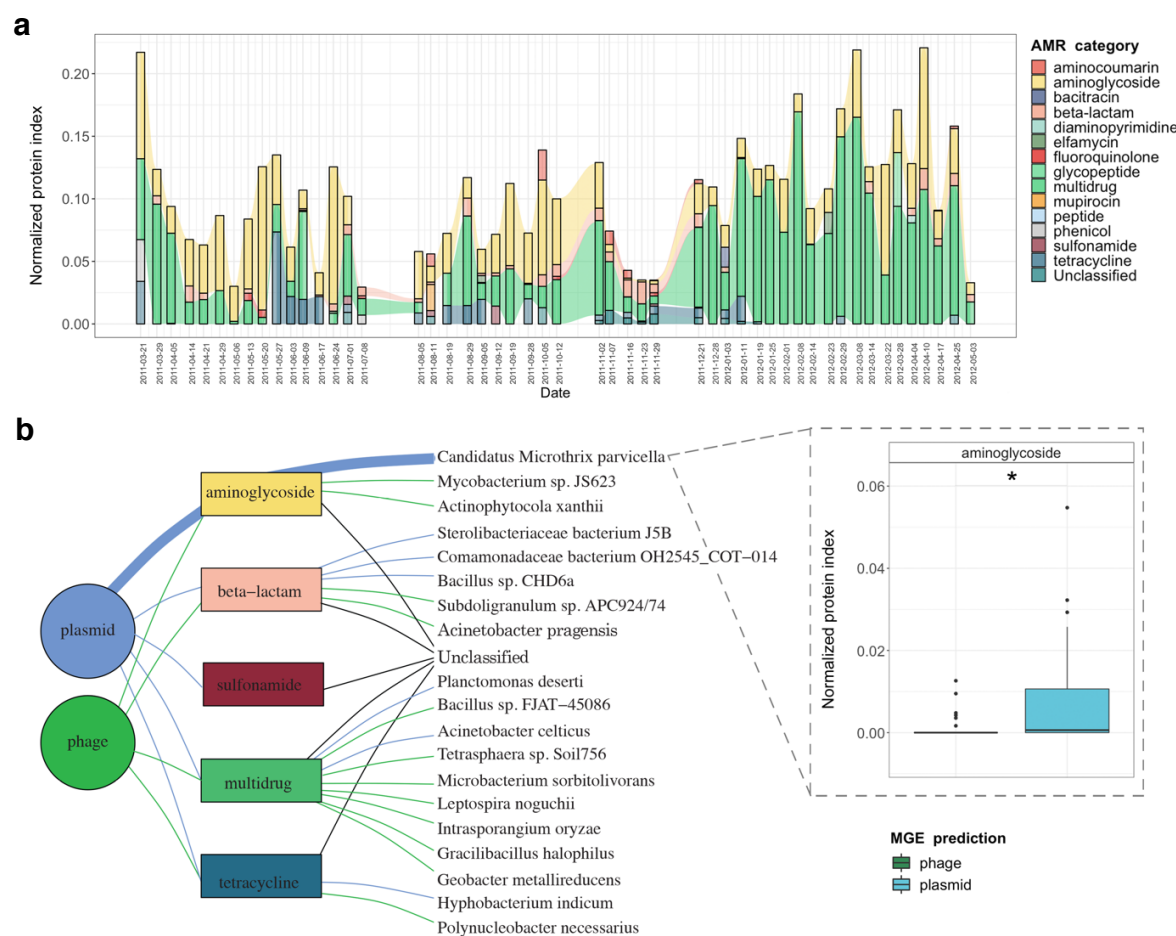


Figure 5.7: Integrative multi-omic assessment of AMR. a. Metagenomic and metatranscriptomic normalized protein levels linked to AMR within the WWTP over time. **b.** Tripartite network assessing the normalized protein levels derived from MGEs and associated taxa. Boxplots depicting significant differential ($adj.p < 0.05$, Two-Way ANOVA) abundance of aminoglycoside resistance in plasmid versus phage in *Candidatus Microthrix parvicella*.

5.4 Discussion

The surveillance of wastewaters for the identification of microbial molecular factors is a critical tool for identifying potential pathogens. This has been highlighted recently with the tracking of SARS-CoV-2 within wastewater treatment plants to assess viral prevalence and load within a given community [417]. Such approaches have also been employed for screening for antimicrobial resistance at a population level [418,419]. So far, several studies [177,383,420,421] have characterized the proliferation of ARGs and antibiotic resistant bacteria in BWWTP. Szczepanowski *et al.* [420] identified 140 clinically relevant plasmid-derived ARGs in a BWWTP metagenome while Parsley *et al.* [421] characterized ARGs from bacterial chromosomes, plasmids and in viral metagenomes found in a BWWTP. Further studies have

shown that conventional BWWTP processes at best only partially remove ARGs from the effluent and may find their way into the urban water cycle [422–424]. Wastewater treatment plants, therefore, are crucial reservoirs of AMR, whose monitoring may allow for early-detection of AMR within the human population feeding into the system. Here, we leveraged a systematic and longitudinal sampling scheme from a BWWTP to identify diverse AMR categories prevalent within the BWWTP microbial community. In line with the studies by Szczepanowski *et al.* [420] and Parsley *et al.* [421], we found up to 29 AMR categories with several ARGs within the BWWTP. More importantly, and unlike the previous studies, we linked the identified ARGs to clinically-relevant ESKAPEE pathogens, which represent a growing global threat to human health.

In our BWWTP samples, we identified a core group of 15 AMR categories that were ubiquitous at all timepoints. In line with the above-mentioned reports, the observed core resistance categories may reflect their abundance in the surrounding human population [425]. This has previously been reported by Pärnänen *et al.* [426], Su *et al.* [427] and Hendriksen *et al.* [177] where they showed that BWWTP AMR profiles correlate with clinical antibiotic usage as well as other socio-economic and environmental factors. On the other hand, bacteria are known to have innate defense mechanisms against inhibitory bacteriocins from other taxa [428]. Therefore, one must be cognizant of the phenomenon that the observed core group of AMR categories may also be a proxy for the abundance of specific resistant bacteria. Despite this observation, it is plausible that both anthropogenic and microbial sources for AMR play a role in the observed resistance categories within the BWWTP. Interestingly, we found that several AMR categories, including ancillary (prevalent, moderate, and rare) groups, were associated with *M. parvicella* within the BWWTP. Similar to the findings by Munck *et al.* [150], we found a wide range of bacteria associated with AMR categories including Acidimicrobiales, Burkholderiales and Rhodocyclales. On the other hand, we report that taxa, including ESKAPEE pathogens, belonging to 25 bacterial orders were associated with 29 AMR categories, compared to the eight bacterial orders reported previously.

It is important to note that the mobilome plays a critical role in the dissemination of AMR within microbial communities. AMR from resistant bacteria within the BWWTP can quickly disseminate within the BWWTP [385,392], including transmission from pathogenic to commensal species [66,67]. As a result, mediated through HGT, the BWWTP becomes a hotspot for resistant bacteria, which are then released back into the receiving environment

[429], and eventually the human population [385,430]. Therefore, to limit the dissemination of AMR, it is important to understand the role of MGEs within the BWWTP. Our comprehensive analyses identified the differential contributions of AMR transmission mediated via phage and plasmid (**Figure 5.8**). Specifically, we identified clear segregation of aminoglycoside, bacitracin, MLS and sulfonamide resistance categories with plasmids, while fosfomycin and peptide resistance were increasingly encoded and conferred via phages. While the association between these AMR categories and plasmids [431–434] or phages [323] are in line with previously reported results, differential analysis between MGEs has not been previously reported and has not been performed on multi-omic levels. As such, in this study we report for the first time the systematic and extensive comparison of AMR encoded and expressed by phages versus plasmids. Our results indicating the segregation of ARGs within the ESKAPEE taxa via the MGEs further provide insights into potential modes of AMR transmission among pathogens. Though one cannot exclude the possibility of transmission of the above-mentioned ARGs via other MGEs, identifying potential segregation of MGEs in the transmission of ARGs brings us one step closer to identifying specific transmission paths and limiting the spread of AMR. For example, some studies have reported plasmid “curing”, the process by which plasmids are removed from bacterial populations, as a strategy against dissemination of AMR [435,436]. As described by Buckner *et al.* [437] plasmid curing, as well as other anti-plasmid strategies, could both reduce AMR prevalence, and (re-)sensitize bacteria to antibiotics [437]. Combining these strategies with AMR categorization according to preference for specific MGEs will give us novel strategies for removing MGE-mediated resistance in the fight against AMR.

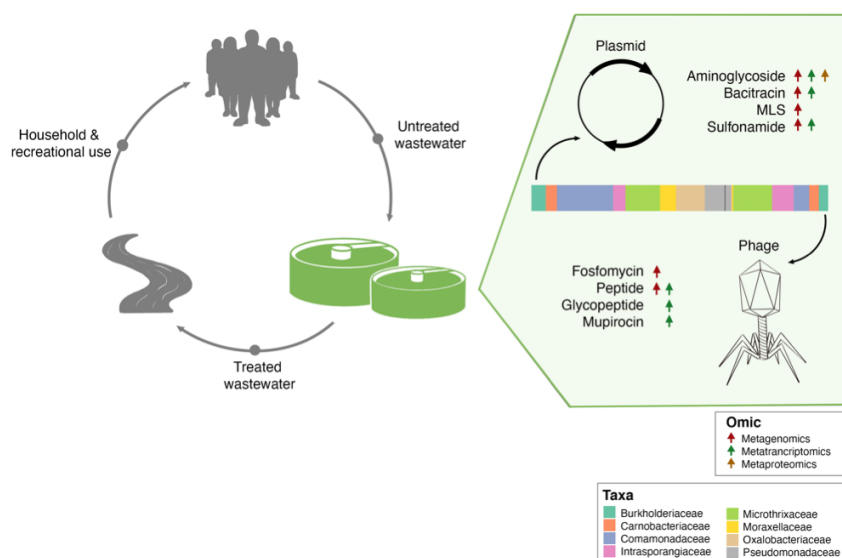


Figure 5.8: Separation of MGE-derived AMR within the BWWTP. A graphical summary highlighting AMR categories found significantly increased in phage versus plasmid in all three omes.

By complementing the metagenomic analyses, metatranscriptomics conferred essential information regarding gene expression within the resistome. For instance, when comparing AMR expression levels of aminoglycoside, bacitracin, and sulfonamide mediated via MGEs, it is noticeable that expression levels in plasmids mirror the genomic content, i.e. they exhibited higher levels of expression when compared to phage. On the other hand, glycopeptide and mupirocin resistance genes which were highly expressed in phages were not reflected within the metagenomic data. Additionally, we found the *YojI* resistance gene to be more highly expressed than any other ARGs. To facilitate resistance against the peptide antibiotic microcin J25, the outer membrane protein, TolC, in combination with *YojI* is required to export the antibiotic out of the cell [412]. Microcin J25 belongs to the group of ribosomally synthesized and post-translationally modified peptides (RiPPs) and has antimicrobial activity against pathogenic genera such as *Salmonella* spp. and *Shigella* spp. [438]. Interestingly, it has only recently been proposed as a treatment option against *Salmonella enterica* and has been discussed in recent years as a potential novel antibiotic [439]. Based on these results, by considering that BWWTPs may reflect both the presence of AMR within the human population as well as be a hotspot of dissemination and generation of new AMR, surveillance of BWWTPs must be emphasized when developing new antibiotics. Our findings collectively suggest that the differential capacity of MGEs to disseminate AMR, coupled with longitudinal and expression-level analyses are crucial for monitoring human health conditions. More importantly, we report for the first time that BWWTP monitoring for AMR may allow for early detection of previously undescribed and previously undescribed resistance mechanisms.

Finally, we applied an integrated multi-omic approach to improve our knowledge on the functional potential of AMR and simultaneously validate the abundance and expression findings of the ARGs. By normalizing the metaproteomic results with the normalized expression of genes we were able to assess the stability of expressed AMR across time. We find that our methodology allows for an unbiased assessment of overall expression accounting for gene copy abundance and expression. These findings support the notion that the ARGs may serve as sentinels or indicators of the presence of particular antimicrobial agents. However, it is plausible that we are only identifying the most abundant proteins and/or proteins that are more stable over time, and do not capture the entirety of the proteome profiles. Factors such as protein decay rates [440] among others, may additionally influence this assessment. Irrespective of these observations, we identified segregation of AMR categories with respect to

plasmids and phages. Our findings also highlighted the potential for identifying segregation of AMR via specific MGEs with an aim towards possible therapeutic and mitigation strategies via for example plasmid curing. Furthermore, we demonstrate that longitudinal analyses are required to survey AMR within BWWTPs due to the variations in the resistome across time. These shifts may either be representative of a shift within the human population itself, which in turn could be associated with the concurrent use of antibiotics at a given time, or competition within the microbial community. In any case, an independent or static analysis of the various time points may show an incomplete view of the BWWTP resistome, thus underlining the importance of our longitudinal resistome analyses. Overall, our findings suggest that BWWTPs are critical reservoirs of AMR, potentially allowing for early detection and monitoring of pathogens and novel resistance mechanisms linked to the introduction of new antimicrobials, whilst serving as a model for understanding the separation of MGEs through AMR.

Chapter 6. Diversity of the resistome and biosynthetic gene clusters in glacier-fed stream biofilms

This chapter is based on the following publication submitted for peer-review:

Susheel Bhanu Busi*, **Laura de Nies***, Paraskevi Pramateftaki, Massimo Bourquin, Leïla Ezzat, Tyler J. Kohler, Stilianos Fodelianakis, Grégoire Michoud, Hannes Peter, Michail Styllas, Matteo Tolosano, Vincent De Staercke, Martina Schön, Valentina Galata, Tom Battin, and Paul Wilmes (2021). Glacier-fed stream biofilms harbour diverse resistomes and biosynthetic gene clusters. *Microbiome in review* [**Appendix A.6**]

* Co-first author

6.1 Introduction

Today, antimicrobial resistance (AMR) has become a well-known threat to human health with an estimated number of 700,000 people per year dying of drug-resistant infections [68]. The dramatic rise of antimicrobial resistance over the past decade has even led to the moniker, “silent pandemic” [441]. Therefore, AMR is often directly associated with human impacted environments with a global increase in resistant bacteria linked to the over- and mis-use of antibiotics [2]. However, contrary to public perception, AMR is a natural phenomenon, which has existed for billions of years [1]. Long before the rather recent use of antibiotics in the clinical setting, microorganisms have used these, along with corresponding protective mechanisms, to establish competitive advantages over other microbes contending for the same environment and/or resources [134].

Microbes, in general, produce a range of secondary metabolites with diverse chemical structures which in turn confer a variety of functions, including antibiotics [442]. Such secondary metabolites including metal transporters and quorum sensing molecules [443,444] are not directly associated with the growth of microorganisms themselves but instead are known to provide benefits by acting as growth inhibitors against competing bacteria. Consequently, many of these natural products have found their uses in industrial settings as well as in human medicine as anti-infective drugs [443,445,446]. The biosynthetic pathways responsible for producing these specialized metabolites are encoded by locally clustered groups of genes known as ‘biosynthetic gene clusters’ (BGCs). Typically, BGCs include genes for expression control, self-resistance, and metabolite export [447]. They can, however, be further divided into various classes including non-ribosomal peptide synthetases (NRPSs), type I and type II polyketide synthases (PKSs), terpenes, and bacteriocins alongside others [446]. NRPSs and PKSs specifically have been of interest due to their known synthesis of putative antibiotics [448,449]. Furthermore, evidence suggests that within these BGCs at least one resistance gene conferring resistance can be found as a self-defense mechanism against the potentially harmful secondary metabolites encoded by the BGC [63]. For instance, the tylosin-biosynthetic gene cluster of *Streptomyces fradiae* also encodes three resistance genes (tlrB, tlrC and tlrD) [450], while in another example, *Streptomyces toyacaensis*, the vanHAX resistance cassette is proximal to the vancomycin biosynthesis gene cluster, thereby encoding inherent resistance [451].

Remote and pristine microbial communities provide a rich genetic resource to explore the historical evolutionary origins of naturally occurring antibiotic resistance from the pre-antibiotic era. Only in few pristine environments with limited anthropogenic influence (e.g., permafrost, glaciers, deep sea, and polar regions) can remnants of the above-described ancient biological warfare mechanisms still be detected. These ARGs and resistant bacteria evolving in pristine environments may therefore be considered the inherent antibiotic resistance present in the environment [134].

We have recently reported the genomic and metabolic adaptations of epilithic biofilms to windows of opportunities in glacier-fed streams (GFSs) [452]. For example, given the short flow season during glacial melt, i.e. summer, the incentive to reproduce quickly while conditions are favorable, is high. During these windows of opportunity, the necessity for taxa to not only acquire physical niches, but also appropriate resources yields a competitive environment. Within these biofilms, we observe complex cross-domain interactions between microorganisms to potentially mitigate the harsh nutrient and environmental conditions of the GFSs. Additionally, owing to their complex biodiversity [453] and generally oligotrophic conditions [454], epilithic biofilms are ideal model systems for understanding BGCs and AMR. While oligotrophy may provide the basis for competition over resources amongst microorganisms such as prokaryotes and (micro-)eukaryotes. Our previous insights revealed that taxa such as *Polaromonas*, *Acidobacteria*, and *Methylobacter* have strong interactions with eukaryotes such as algae and fungi [452]. The inherent diversity allows for understanding the influence of AMR in microbial interactions. For example, the accidental discovery of penicillin by Alexander Fleming in 1928 based on bacterial-fungal interactions, [455], has since been expanded upon by Netzker *et al.* [456]. They reported that microbial interactions lead to the production of bioactive compounds including antibiotics that may shape the microbial consortia within a community.

Here, to shed light on the role of AMR in shaping microbial communities within (relatively) pristine environments, we used high-resolution metagenomics to investigate twenty-one epilithic biofilms from glacier-fed streams. These samples were collected from 8 GFSs spread across the Southern Alps in New Zealand and the Caucasus in Russia. Herein, we found 29 categories of ARGs within the GFSs across both bacterial and eukaryotic domains. Importantly, most of the AMR was found in bacteria. We also identified antibacterial BGCs that were encoded both in bacterial and eukaryotes suggesting extensive intra- and inter-domain

competition. Our findings demonstrate that microorganisms within biofilms from pristine environments not only encode ARGs, but that they may potentially influence several features of epilithic biofilms such as biofilm formation, community assembly and/or maintenance, including conferring mechanisms for competitive advantages under extreme conditions.

6.2 Methods

6.2.1 Sampling and biomolecular extraction

Eight GFSs were sampled in early- to mid-2019 from the New Zealand Southern Alps and the Russian Caucasus, respectively, for a total of 21 epilithic biofilms. The biofilm samples were collected from each stream reach due to biofilms ranging from abundant to absent, depending on stream geomorphology. One to three biofilm samples were collected per reach, taken using sterilized metal spatulas to scrape rocks, followed by their immediate transfer to cryovials. Samples were immediately flash-frozen in liquid nitrogen and stored at -80 °C until DNA was extracted. DNA from the epilithic biofilms was extracted using a previously established protocol [457] adapted to a smaller scale due to relatively high DNA concentrations. DNA quantification was performed for all samples with the Qubit dsDNA HS kit (Invitrogen).

6.2.2 Sequencing and data processing for metagenomics

Random shotgun sequencing was performed on all epilithic biofilm DNA samples, after library preparation using the NEBNext Ultra II FS library kit. 50 ng of DNA was enzymatically fragmented for 12.5 mins and libraries were prepared with 6 PCR amplification cycles. An average insert of 450 bp was maintained for all libraries. Qubit was used to quantify the libraries followed by sequencing at the Functional Genomics Centre Zurich on a NovaSeq (Illumina) using a S4 flowcell. The metagenomic data was processed using the Integrated Meta-omic Pipeline (IMP v3.0; commit# 9672c874 available at <https://git-r3lab.uni.lu/IMP/imp3>) [231]. IMP's workflow includes pre-processing, contig assembly, genome reconstruction (metagenome-assembled genomes, i.e. MAGs) and additional functional analysis of genes based on custom databases in a reproducible manner [231].

6.2.3 Identification of antimicrobial resistance genes, antibiotics biosynthesis pathways and BGCs

For the prediction of ARGs the IMP-generated contigs were used as input for PathoFact [396]. Identified ARGs were further collapsed into their respective AMR categories in accordance with

the Comprehensive Antibiotic Resistance Database (CARD) [43]. PathoFact uses an HMM-based search to identify homologous sequences across genomic data, therefore possibly also detecting resistance genes within eukaryotic genomic fragments. Subsequently, the raw read counts per ORF, obtained from PathoFact, were determined using FeatureCounts [232].

To identify pathways for the biosynthesis of antibiotics, we assigned KEGG orthology (KOs) identifiers to the ORFs using a hidden Markov model [285] (HMM) approach using *hmmsearch* from HMMER 3.1 [458] with a minimum bit score of 40. Additionally, we linked the identified KOs to their corresponding KEGG orthology pathways and extracted the pathways annotated as antibiotic biosynthesis pathways by KEGG. Both the identified ARGs and KEGG pathways were then further linked to associated bacterial taxonomies. The bacterial and eukaryotic taxonomies were assigned using the PhyloDB and MMETSP databases associated with EUKulele (commit# fb8726a; available at <https://github.com/AlexanderLabWHOI/EUKulele>). Consensus taxonomy per contig was then used for downstream analyses including association with ARGs.

We further identified BGCs within the MAGs using antiSMASH (ANTibiotics & Secondary Metabolite Analysis SHell) [459] and annotated these using deepBGC [460]. To link BGCs and ARGs, we linked the resistance genes to their associated assembled contigs, followed by identifying the corresponding bins (MAGs) to which said contigs belonged.

6.2.4 Data analysis

The relative abundance of the ORFs was calculated based on the RNum_Gi method described by Hu *et al.* [233]. Figures for the study including visualizations derived from the taxonomic and functional analyses were created using version 3.6 of the R statistical software package [289] and using the *tidyverse* package [461]. Alluvial plots were generated using the *ggalluvial* package [462] while heatmaps were generated using the *ComplexHeatmap* package [463] developed for R.

6.3 Results

6.3.1 Antimicrobial resistance in a pristine environment

We characterized the resistomes of GFS epilithic biofilms and assessed the distribution of AMR in twenty-one epilithic biofilm samples, across 8 individual glaciers originating from the

Southern Alps in New-Zealand (SA1, SA2, SA3 and SA4) and the Caucasus in Russia (CU1, CU2, CU3, CU4). In total, we identified a high number (n=1840) of ARGs within 29 categories of AMR, with similar AMR profiles observed across all GFSs (**Figure 6.1, Appendix B.5: Supplementary figure 6.1**), except for SA2 and SA3 where the differences were driven by elevated fluoroquinolone, glycopeptide and phenicol resistance, respectively. It is to be noted that while ARGs refer to the genes encoding specific resistance, AMR categories derived from metagenomic data in this context, typically reflect the functional potential associated with respect to the resistance encoded. Of the identified AMR categories, beta-lactam and multidrug resistance (i.e. resistance conferring protection against multiple antibiotic classes), followed by aminoglycoside resistance, were found to be highly abundant in all samples. We subsequently analyzed the diversity of ARGs within the various resistance categories and found beta-lactam resistance to represent the largest resistance category, contributing 930 unique ARGs to the resistome. This was followed by multidrug (179 ARGs) and aminoglycoside (176 ARGs) resistance (Supp. Table 2). In contrast, some resistance categories such as polymyxin and pleuromutilin resistance were only detected at very low levels within the epilithic biofilm resistomes.

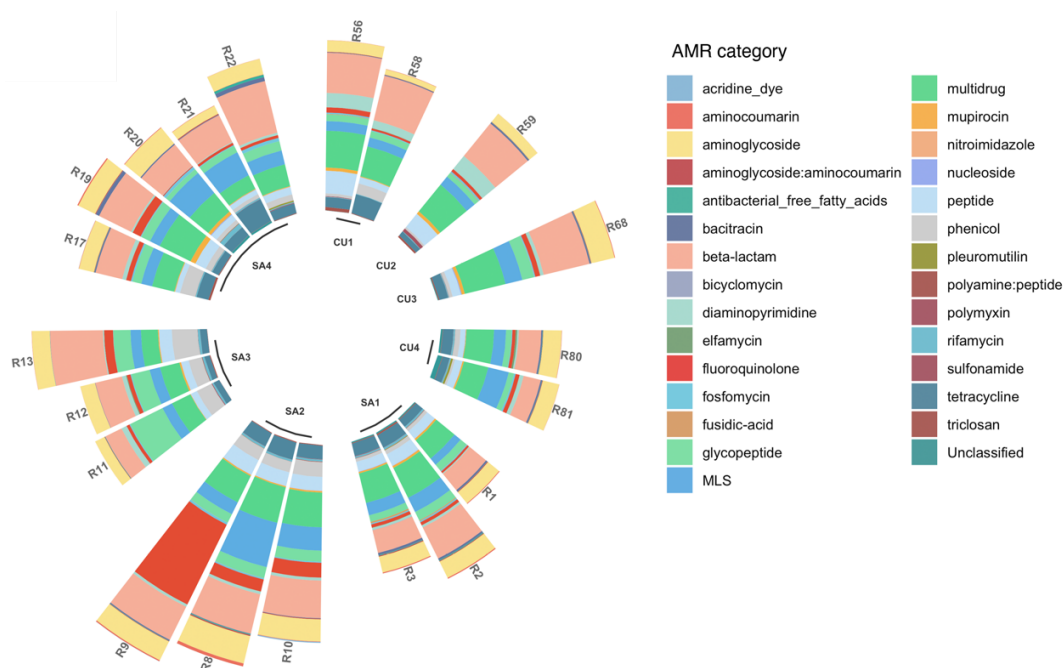


Figure 6.1: Epilithic biofilms in GFSs harbors a diverse resistome. Relative abundance of 29 AMR categories within 21 epilithic biofilms collected from four New Zealand Southern Alps (SA) and four Russian Caucasus (CU) GFSs.

We further investigated the contribution of microbial populations to the resistome and found contributions from both prokaryotes and eukaryotes (**Figure 6.2a**). Prokaryotes within this study refer to bacteria alone, since archaea encoded for an infinitesimal number of ARGs ($<0.000001\%$ RNum_Gi; Methods), and therefore were excluded from further analyses. Among the eukaryotes, the phylum Ochrophyta (algae) was the dominant contributor and encoded most of the AMR categories (**Figure 6.2b, Appendix B.5: Supplementary figure 6.2a**). In bacteria, AMR was more evenly distributed with most of the phyla encoding ARGs across all categories (**Figure 6.2b**). However, members of the Alphaproteobacteria, Betaproteobacteria, and the Bacteroidetes/Chlorobi group encoded the highest overall ARG abundance (**Figure 6.2b, Appendix B.5: Supplementary figure 6.2b**). Additionally, AMR categories such as aminoglycoside, beta-lactam, glycopeptide and rifamycin resistance (among others) were widely distributed in both bacteria as well as among the eukaryotes. On the other hand, categories such as aminocoumarin, bacitracin, and diaminopyrimidine resistance were found to be primarily encoded by bacteria.

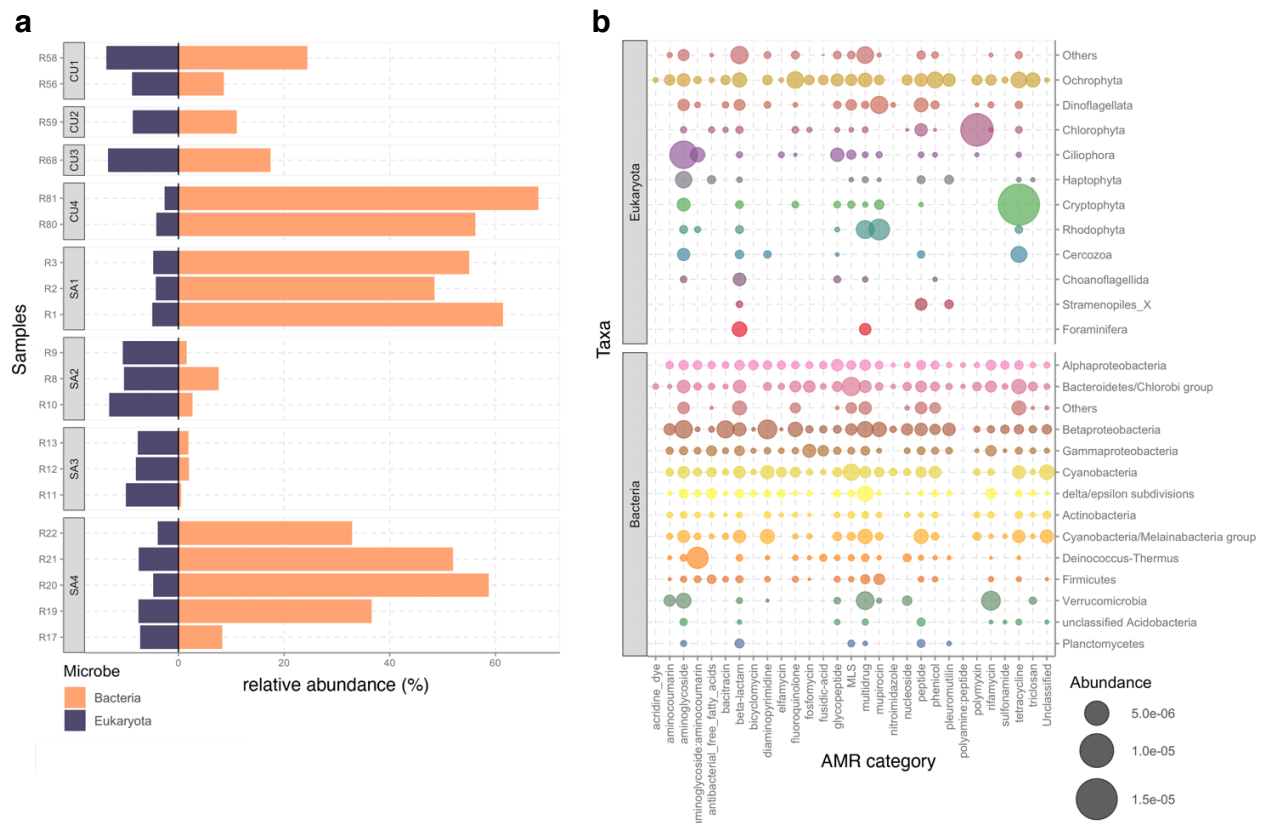


Figure 6. 2: Taxonomic affiliation of AMR in GFSSs epilithic biofilms. a. Bar plots depicting the relative abundance of bacteria and eukaryotes encoding ARGs. **b.** Phylum-level representation of the AMR abundances across bacteria and eukaryotes. Size of the closed circle indicated the normalized relative abundance (Rnum_Gi; see *Methods*), whereby the color represents individual phyla.

6.3.2 Antibiotic biosynthesis pathways and biosynthetic gene clusters

As described above, beta-lactam, multidrug and aminoglycoside resistance were the most abundant resistance categories within GFS epilithic biofilms. This was not surprising as beta-lactams and aminoglycosides are natural and prevalent compounds [464,465]. Furthermore, multidrug resistance is typically conferred via efflux machineries which were also common in the GFS epilithic biofilms. These typically serve dual purposes in particular for protein export within most bacteria [466]. Based on these results, it is therefore highly likely that pristine environments such as GFSs potentially reflect the spectrum of natural antibiotics and their resistance mechanisms, reinforcing their capacity to serve as natural baselines for assessing enrichments and spread of AMR.

To further understand if these encoded resistance genes reflected natural antibiotic pressure, we investigated pathways associated with antibiotic biosynthesis using the KEGG database [214]. In total, we identified seven different pathways corresponding to the biosynthesis of macrolides (MLS), ansamycins, glycopeptides (vancomycin), beta-lactams (monobactam, penicillin and cephalosporin), aminoglycosides (streptomycin), and tetracyclines, which were present in various abundances in all samples (**Appendix B.5: Supplementary figure 6.3a**). Importantly, the identified antibiotic synthesis genes thereby corresponded to the resistance categories identified within the epilithic biofilms. Interestingly, in most of the GFSs, antibiotic biosynthesis was primarily encoded by bacteria spanning multiple phyla (**Appendix B.5: Supplementary figure 6.3b-c**). Exceptions to these were GL11 and GL15 in which biosynthesis pathways were equally distributed among eukaryotes, specifically Ochrophyta, in addition to bacteria.

To further validate our observations, we assessed the abundance of BGCs, which are known to encode genes for secondary metabolite synthesis, including antibiotics. We found six different structural classes of BGCs by annotating 537 medium-to-high quality (>50% completion and <10% contamination) bacterial and 30 eukaryotic MAGs using antiSmash [459] and DeepBGC [460]. Using this ensemble approach, we identified one or more BGCs in most bacterial (n=490, ~91% of all bacterial MAGs) and eukaryotic (n=28) MAGs. Of these BGCs, those annotated with an antibacterial function were dominant across the microbial populations, represented here by the MAGs, and were found across all phyla (**Figure 6.3a**). Overall, a wider variety of BGCs associated with cytotoxic activity, inhibitory, and antifungal mechanisms were also identified in bacteria. Eukaryotes, on the other hand, encoded a high prevalence of

antibacterial BGCs (~93% of all eukaryotic MAGs) (**Figure 6.3a**). We further annotated those BGCs identified as antibacterial to determine their subtypes and found that most of them were ‘unknown’ (**Figure 6.3b**). However, other identified subtypes include ribosomally synthesized and post-translationally modified peptides (RiPPs) such as bacteriocins, along with NRPs, PKs, and terpenes.

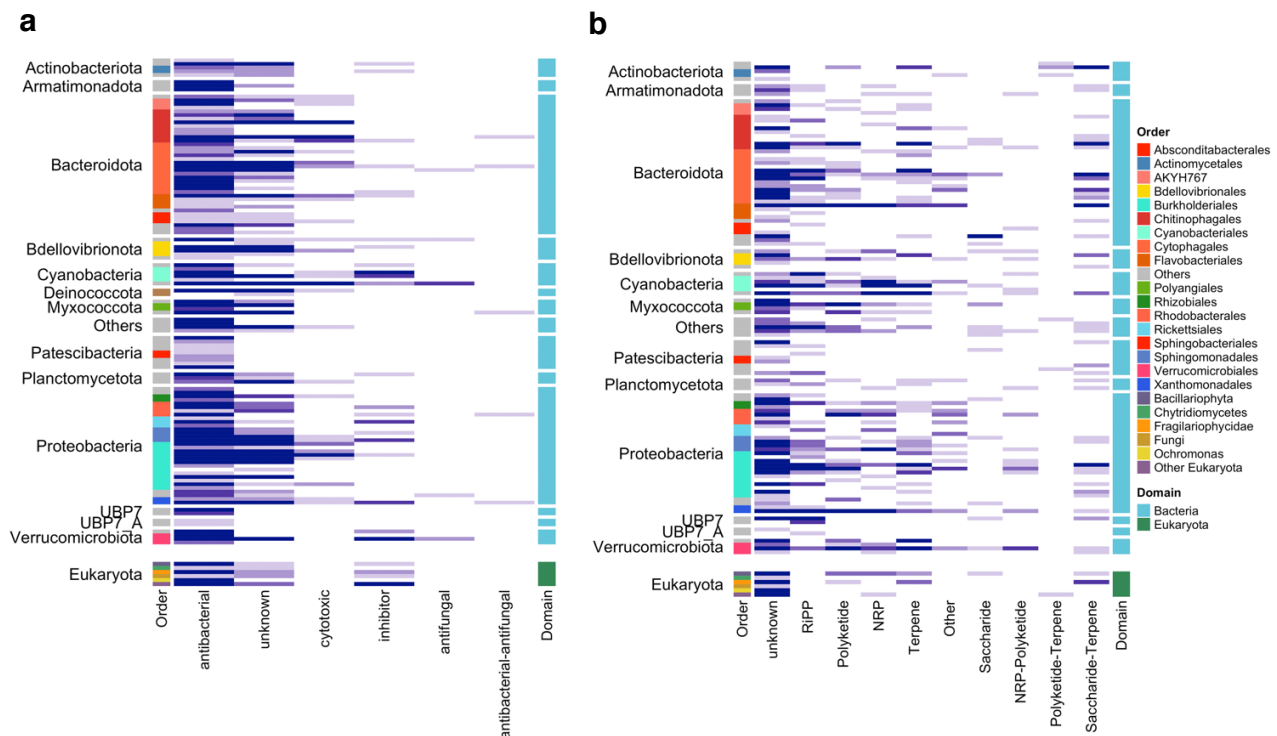


Figure 6.3: Biosynthetic gene clusters indicate the resistome potential. a. Heatmap depicting the overall abundance of BGCs identified across bacterial and eukaryotic MAGs. The respective phyla are listed on the left while the colored legend represents the taxonomic order. **b.** In-depth characterization of the ‘antibacterial’ BGCs found within all phyla and orders across medium-to-high quality MAGs.

According to the resistance hypothesis [63], within or close to, each BGC there is at least one gene conferring resistance to its encoded secondary metabolite. To test this, we assessed whether the MAGs encoding a BGC also encoded corresponding ARGs. In line with this hypothesis, we identified BGCs and their respective resistance genes in close proximity to each other through their localization on the same contig. Consequently, we identified various BGCs encoded together with ARGs in both the bacterial and eukaryotic MAGs. For example, we found that an antibacterial BGC was encoded by *Flavobacterium* spp. on the same contig as both MLS (macrolides, lincosamides and streptogramin) and beta-lactam resistance genes (**Figure 6.4**). Incidentally, we also found that a candidate phyla radiation (CPR) bacterium (*Aalborg-*

AAW-1; phylum Patescibacteria) also encoded both antibacterial BGC and MLS resistance on the same contig.

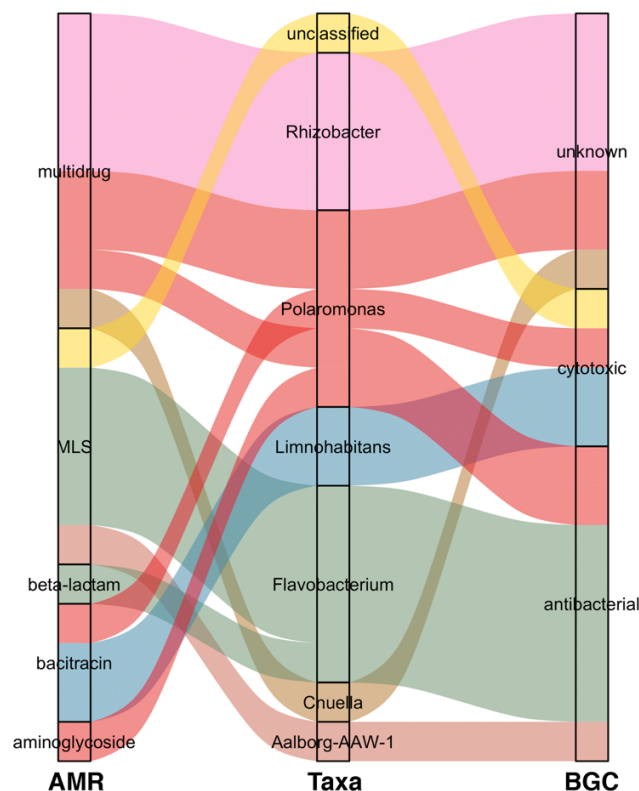


Figure 6.4: Association of BGCs with AMR. Alluvial plots depicting the taxa where both BGCs and AMR were found adjacently on the same contig. Colors indicate genera associated with the MAGs.

6.4 Discussion

Microbial reservoirs in pristine environments, with little to no impact from anthropogenic selection pressures, provide the opportunity to investigate the natural propensity and linked evolutionary origins of AMR. Here, by leveraging high-resolution metagenomics on twenty-one epilithic biofilms, we assessed the resistomes of eight individual GFS epilithic biofilms.

To date, while many studies have looked for novel antibiotics and resistance genes in pristine environments such as the deep sea [467] or the polar regions [146], few have explored the full diversity of antibiotic resistance in such environments [468,469]. Van Goethem *et al.* [470] identified 117 naturally occurring ARGs associated with multidrug, aminoglycoside and beta-lactam resistance in pristine Antarctic soils. Similarly, D’Costa *et al.* [1] identified a collection of

ARGs encoding resistance to beta-lactams as well as tetracyclines and glycopeptides in 30,000-year-old Beringian permafrost sediments. In agreement with these previous studies, we identified 29 AMR categories, including the previously mentioned resistance categories, in the studied biofilm communities. Among these, the highest ARG abundance was associated with aminoglycoside and beta-lactam resistance. Our study further suggests that although the overall abundance differs, the epilithic resistome was highly similar in all GFSs, independent of origin (i.e. New Zealand or Russia). Furthermore, our results agree with the results obtained in other resistomes identified in pristine environments such as Antarctic soils and permafrost in terms of the identified ARGs. Unlike previous studies, where ARGs were primarily associated with bacteria, we report for the first time that AMR was associated with both bacteria and eukaryotes in various abundances in environmental samples including GFSs. A previous study by Brown *et al.* [471] reported that the IRS-HR (isoleucyl-tRNA synthetase - high resistance) type gene conferring resistance against mupirocin was identified in *Staphylococcus aureus*. More importantly, they suggested that horizontal gene transfer led to the acquisition of IRS-HR genes by bacteria from eukaryotes [471]. Despite these early reports, the contribution of eukaryotes to most resistomes, including from pristine environments, has largely been unexplored thus far. An exception to this was the report by Fairlamb *et al.* [472] who identified eukaryotic drug resistance, especially encoded by fungi (*Candida* and *Aspergillus*) and parasites (*Plasmodium* and *Trypanosoma*). However, most of these modes of resistance were highly specific towards particular drug treatments [472]. Our results specifically revealed that taxa from the phylum Ochrophyta encoded resistance to 28 AMR categories and this was also reflected in other (micro-)eukaryotes.

Apart from encoded resistance mechanisms, microalgae such as Ochrophyta have been of interest as a source of (new) antimicrobial compounds [473,474]. In line with this, Martins *et al.* suggested that extracts from different microalgae may potentially serve not only as antimicrobial agents, but also as anti-cancer therapeutics. However, our present results suggest that these taxa may also serve as environmental reservoirs for AMR itself. It is however presently unclear whether this phenomenon confers advantages with respect to niche occupation and protection against bacterial infection as well as whether the eukaryotes are sensitive to the antibiotics produced by them.

Studies delving into the origins of AMR have reported that fecal pollution may explain ARG abundances in anthropogenically impacted environments [475]. This phenomenon was also

observed by Antelo *et al.* [476] and others [477] who detected ARGs in soils in Antarctica, especially in proximity to scientific bases. Although it is plausible that some of the GFSs sampled in our study may indeed be under anthropogenic influence, in pristine environments, AMR is most likely derived from natural antibiotics produced by microorganisms as a competitive advantage. Microorganisms acquire resistance either as a protective measure against other microorganisms [7,61] or as a self-defense mechanism to prevent inadvertent suicide by damaging metabolites [63]. Accordingly, we found both antibiotic biosynthesis pathways and BGCs within the epilithic resistomes. We identified pathways for the biosynthesis of glycopeptides, beta-lactams, and aminoglycosides, among others, concurrent with the high abundance of ARGs against said antibiotics. Additionally, we identified BGCs with a predicted antibacterial function in both eukaryotes and bacteria. While a limited number of studies such as Waschulin *et al.* [478] and Liao *et al.* [479], have shown BGCs in pristine environments, none of these studies have contextualized the co-occurrence of BGCs with AMR. Hence, we not only found that most of our MAGs contain BGCs, of which many have an antibacterial function, but also found all MAGs to encode multiple resistance genes. Additionally, we found several BGCs closely localized to ARGs on the same contig, thereby indicating an immediate self-defense mechanism against the encoded secondary metabolites. This agrees with the resistance hypothesis highlighted by Tran *et al.* stating that a gene conferring resistance to potentially harmful metabolites produced by the organism are to be found within the BGC-encoding operons [63]. We also observed that the recently identified CPR bacteria [480] (in our case, phylum Patescibacteria) not only encoded for AMR but also harbored genes associated with the production of molecules with antibacterial effects. Although Patescibacteria have been identified in oligotrophic environments [481,482] with carbon and/or nutrient limitations similar to those observed for GFSs, it is plausible that their ability to survive with minimal biosynthetic and metabolic pathways may indeed depend on the expression of BGCs and AMR. At the time of writing, a preprint by Maatouk *et al.* [483], described the presence of ARGs across publicly available CPR bacterial genomes. In addition, we report the identification of AMR within GFS-derived CPR genomes, likely as a means of competitive inhibition against other taxa. Alternatively, biofilms may also allow for collective resistance, tolerance, and exposure protection to antibacterial compounds [484]. The AMR and BGCs encoded by most phyla may therefore affect cooperation and/or interactions associated with nutrient exchange, leading to the privatization of public goods [484]. Such a phenomenon may be achieved due to the competition within taxa, both at the intra- and inter-species levels, via secretion of toxins [61] and occupying spatial niches [485,486] thereafter. Furthermore, Stubbendieck and Straight

previously highlighted the multifaceted effects of bacterial competition which include the potential taxation and subsequent increase in bacterial fitness [487]. Thus, the in-situ competition within multi-species biofilms may allow for cross-phyla and cross-domain interactions whilst simultaneously increasing the overall fitness of the endogenous epilithic microbial community. Alternatively, these interactions or lack thereof may shape the overall community including spatial organization [488], especially in energy limited systems such as the GFSSs.

6.5 Conclusions

Epilithic biofilms are an integral and key mode of survival in extreme environments such as glacier-fed stream ecosystems. Herein, we report that these biofilms provide critical insights into the naturally occurring resistome. Our findings demonstrate that intra- and inter-domain competition and survival mechanisms shed light on the ecological dimension of microbial communities. Furthermore, we reveal the congruence of genes encoding for both BGCs and AMR, in both bacteria and eukaryotes. More importantly, we highlight for the first time the comprehensive AMR profile of CPR bacteria and of (micro-)eukaryotes. Collectively, our results highlight underlying resistance mechanisms, including BGCs, employed in 'biological warfare' in oligotrophic and challenging glacier-fed stream ecosystems.

Chapter 7. General conclusion and future perspectives

Parts of this chapter are based on the following publications submitted for peer-review:

Laura de Nies, Susheel Bhanu Busi, Paul Wilmes (2021). Reservoirs of antimicrobial resistance in the context of One Health
Current Microbiology in review [**Appendix A.1**]

Valentina Galata, Susheel Bhanu Busi, Benoit Josef Kunath, **Laura de Nies**, Magdalena Calusinska, Rashi Halder, Patrick May, Paul Wilmes, Cedric Christian Laczny (2021). Functional meta-omics provide critical insights into long- and short-read assemblies
Briefings in Bioinformatics [**Appendix A.7**]

7.1 General overview

Antimicrobial resistance is an ever-present challenge, not necessarily due to the use of antibiotics alone, but also due to the role of mobile genetic elements. It is therefore necessary to understand the dissemination of antibiotic resistance by characterizing the resistome within various environments and to unravel how they act as a reservoir for bacterial pathogens. A One Health perspective integrating research on AMR as well as resistant microbes, circulating in humans, animals and the environment is therefore crucial to enhance our understanding of the complex epidemiology of AMR. In recent years, many studies have used different techniques to sample the resistomes of soils, wastewater, as well as human and animal microbiota. While many of these studies are focused on specific pathogens or resistance categories, research utilizing sequence-based metagenomics provides a comprehensive perspective on all ARGs within different microbial reservoirs (**section 1.5**). Currently, various tools exist for the prediction of ARGs in metagenomes, with other tools focusing on the independent prediction of MGEs. Consequently, we developed PathoFact, a pipeline for the prediction of antimicrobial resistance and virulence factors and their subsequent contextualization to MGEs (**Chapter 2**). However, few metagenomic studies are focused on multiple microbial reservoirs or target only one side of the One Health triad. This work presents extensive metagenomic analyses on different microbial reservoirs of antimicrobial resistance.

In **Chapter 3** we investigated the infant gut resistome and found that the abundance of ARGs against (semi-)synthetic agents were increased in infants born via cesarean section compared to those born via vaginal delivery at five days after birth. Additionally, we identified horizontal gene transfer events, mediated through phage and plasmid, of antimicrobial resistance at an early age. In **Chapter 4** we further assessed the evolution and consecutive dissemination of AMR within the commensal gut microbiome, utilizing a mouse model and a single course treatment with an antibiotic cocktail. While plasmids and phages were found to contribute to the spread of AMR, we found that integrons represented the primary factors mediating AMR in the antibiotic-treated mice. Concurrently, we observed an increase of multidrug resistant *Akkermansia muciniphila* and members of the *Lachnospiraceae* family. Finally, in **Chapters 5 and 6**, to complete the One Health triad, we investigated the environmental resistome, comprising both the urban environments, i.e., the WWTP, and a natural environment, i.e. GFS biofilms. Utilizing a multi-omics approach, we investigated the WWTP resistome over a 1.5 years' time series and found that a core group of fifteen AMR categories were always present. Additionally, we found a significant difference in AMR categories encoded on phages versus

plasmids indicating that the MGEs contributed differentially to the dissemination of AMR. On the other hand, the GFS biofilms represent pristine environments with limited anthropogenic influences. Therein, we found that eukaryotes, as well as prokaryotes, may serve as AMR reservoirs owing to their potential for encoding ARGs. In addition to our identification of biosynthetic gene clusters encoding antibacterial secondary metabolites, our findings highlight the constant intra- and inter-domain competition and the underlying mechanisms influencing microbial survival in GFS epilithic biofilms.

7.2 AMR within and across biomes

Humans, animals, sewage and sludge are considered important reservoirs for ARGs because abundant ARGs have been frequently detected in these environments. Over the last decades, antibiotic usage, in particular, has increased the prevalence of ARGs in the human and animal microbiome. In the natural environment across the globe, resistance is ancient with several ARGs found in pristine environments with minimal anthropogenic impact. Consequently, the in-depth investigation of the diversity and abundances of ARGs in various environments is central to establishing the overall picture that is essential for management decision frameworks for controlling antibiotic resistance. To enhance our understanding of the evolution and dissemination of AMR we systematically resolved different reservoirs of antimicrobial resistance as described in the below mentioned chapters.

In the previous chapters we have described the resistome of the infant gut (**chapter 3**), the WWTP (**chapter 5**), and GFS biofilms (**chapter 6**). By identifying the resistome within individual biomes, we are able to subsequently compare their identity and prevalence against each other. This can be further supplemented with publicly available datasets of the adult microbiome and livestock (i.e. cow, pig and chicken). Overall human and livestock resistomes demonstrated a higher abundance of ARGs compared to environmental samples (**Figure 7.1a**). Interestingly, the infant resistome at 1 year of age demonstrated comparable AMR abundances to the adult resistome. Additionally, together with the GFS biofilms, the urban WWTP demonstrated the lowest abundance of AMR. WWTPs have long been known as hotspots of AMR encoding a wide variety of AMR categories (**section 5.3.1**). However, few studies to date have investigated the overall AMR abundances between WWTPs and human microbiomes. Nonetheless, concordant with our findings both Li *et al.* [181] and Zhou *et al.* [489] identified a higher absolute abundance of ARGs in human samples compared to WWTP samples. In addition to the overall AMR abundances, the similarity of ARG compositions in the different samples, evaluated using

non-metric multidimensional scaling (NMDS), revealed that the grouping pattern was primarily influenced by the types of environments, with samples of the same environment grouping together (**Figure 7.1b**). Human and environmental samples demonstrated the greatest distance, while livestock patterns were closer to human samples.

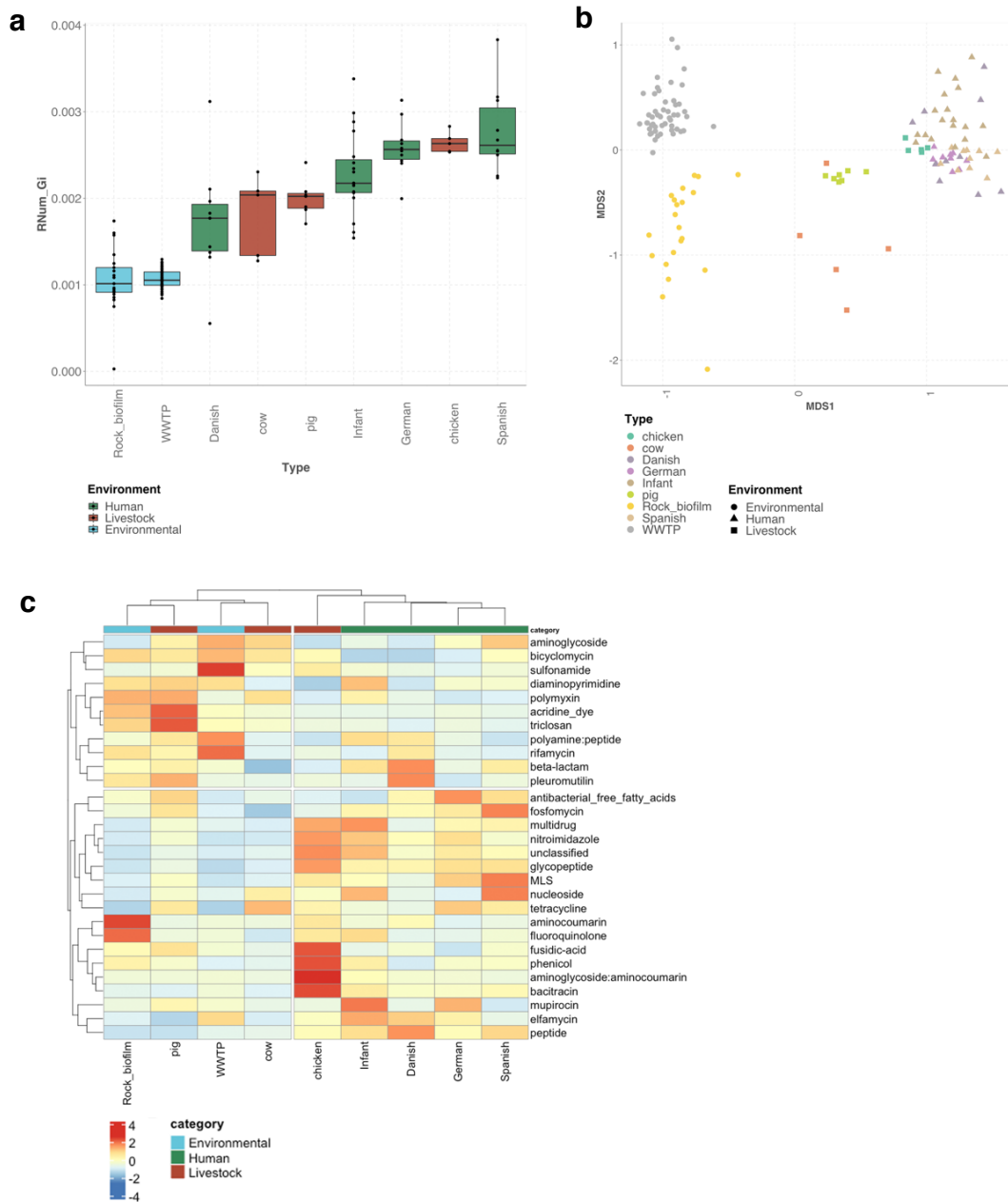


Figure 7.1: Comparison of the human, animal, and environmental resistome. a. Bar plot depicting the relative abundance of AMR in human, animal, and environmental reservoirs. **b.** Evaluation of the ARG composition of the human, animal, and environmental resistome using non-metric multidimensional scaling (NMDS). **c.** Heatmap comparing the abundance of various AMR categories between different microbial reservoirs.

Within each microbial reservoir we identified multiple AMR categories and identified which resistance categories dominated each respective resistome. Aminoglycoside and beta-lactam resistance were found to be prevalent both in the GFS biofilms (**section 6.3.1**) and WWTP (**section 5.3.1**), while diaminopyrimidine and glycopeptide resistance, among others, was found to be abundant in the infant microbiome at 1 year of age (**section 3.3.4**). Additionally, in all reservoirs, multidrug resistance mechanisms were found to be highly abundant. Yet, although less prevalent, less abundant resistance categories should not be disregarded. When comparing the various resistance categories within different resistomes we observe that aminocoumarin and fluoroquinolone resistance are more prevalent in GFS biofilms compared to any other microbial reservoir. The same can be observed for sulfonamide and rifamycin resistance in the WWTP, while multidrug and mupirocin are especially abundant within the infant gut microbiome (**Figure 7.1c**). Interestingly, Livestock resistomes show similarities with both the human and environmental resistomes.

To further compare the similarity of ARGs composition on all three sides of the One Health triad, we determine the shared ARGs between all microbial reservoirs. In total 187 ARGs were found in every single microbial reservoir. Interestingly, a total of 400 ARGs were identified to be unique to the environmental resistome, with 181 unique to GFS biofilms and 89 unique to the WWTP (**Figure 7.2**). Of these resistance categories the majority corresponds to resistance against beta-lactams.

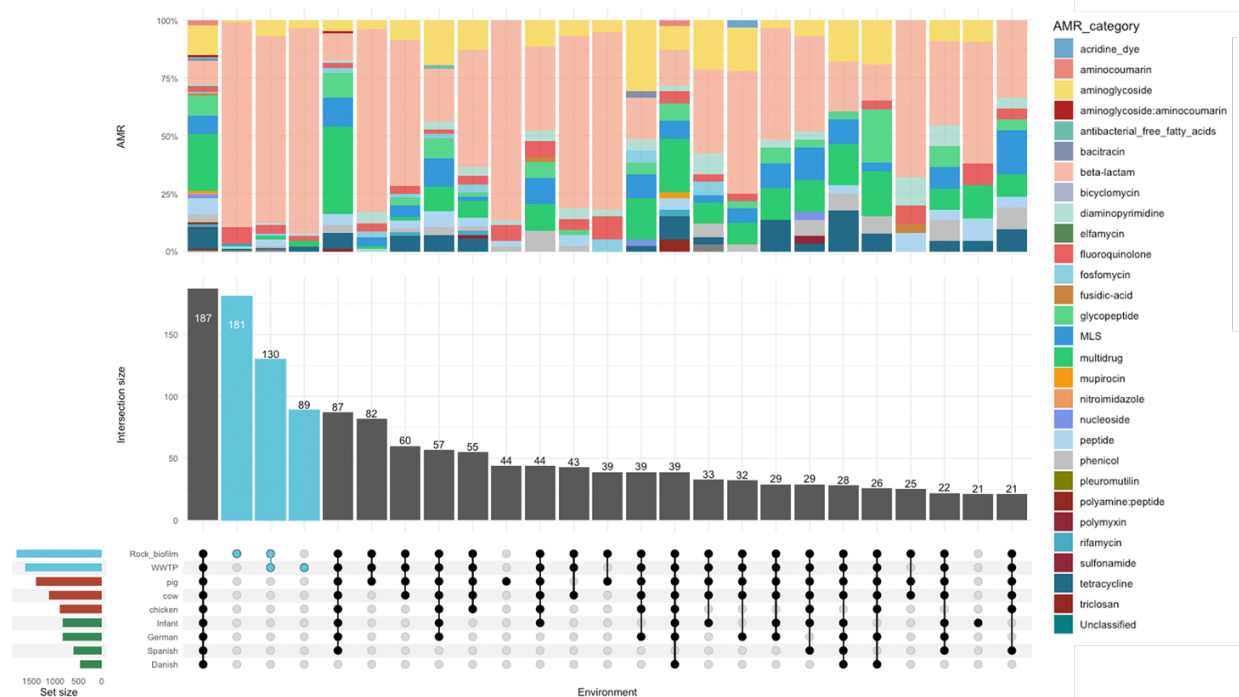


Figure 7.2: ARG diversity in different microbial reservoirs. Upset plot depicting the unique number of antimicrobial resistance genes (ARGs) in different microbial reservoirs. Blue bars representing ARGs unique to environmental reservoirs.

Environmental microbial communities tend to be more complex compared to human and animal microbiomes. Additionally, as described in **section 6.1** microorganisms, especially within complex communities, naturally produce and use antibiotics as competitive mechanisms against other microbes competing for the same environment and/or resources. Beta-lactams in specific are a large group of natural antibiotics which readily can be found in many environments. Allen *et al.* [490], Fonseca *et al.* [491] and Goethem *et al.* [470] all reported an abundance of known beta-lactam resistance genes and unique homologs in pristine environments. Meanwhile, Piotrowska *et al.* [492], identified a wide diversity, including new variants, of beta-lactam resistance genes in WWTPs. Simultaneously, in line with our findings in **section 5.3.1-5** Majeed *et al.* [493] reported that a majority of the ARGs conferred beta-lactam resistance. Our observations suggest that the human and livestock microbiomes have the highest absolute abundance of AMR, while the environmental reservoirs have the highest diversity of ARGs. Consequently, both the abundance and diversity of resistance genes need to be considered when addressing AMR in a One Health context.

7.3 Dissemination of MGE-derived AMR

The development of AMR and the subsequent spread of ARGs within microbial populations has to a large extent been enabled by the recruitment of ARG via MGEs. For instance, multiple studies such as Botts *et al.* [494], Wang *et al.* [495] and Mathers *et al.* [496] have reported multiple transmissible multi-drug plasmids, while Vintov *et al.* [497] and Gomez-Gomes *et al.* [317] describe phage-mediated AMR transmission. Therefore, it is crucial to identify the relationship between ARGs and MGEs. Consequently, in **chapter 2** we developed the PathoFact pipeline which allows us to contextualize the identified ARGs to their localization on phages and plasmids in metagenomic datasets. In **chapter 3** we found that both phage- and plasmid- encoded ARG were dominant in the infant microbiome during the first year of life, independent of birth mode. Additionally, in **chapter 5** we found a significant difference between plasmid- and phage- derived ARGs within the WWTP.

Anthropogenic factors are important contributors to AMR, as described by Perry and Wright [498], who posited the role of the former in the mobilization and dissemination of ARGs. Comparing the diverse resistomes, we find that plasmid-encoded ARG is less abundant in GFS biofilms compared to either the WWTP or the infant microbiome, or the public datasets of the adult and livestock microbiome (**Figure 7.3**). This finding is in line with the study by Hughes and Datta [499] who concluded that plasmids isolated from pathogenic bacteria predating the antibiotic era only rarely encode any ARGs. As GFS biofilms are environments with limited anthropogenic contamination it is expected to find a lower abundance of plasmid-encoded ARGs compared to anthropogenic environments, i.e. humans and livestock.

In **chapter 4** we further assessed the consecutive dissemination of AMR within the commensal gut microbiome after a single course antibiotic treatment. Similar to **chapters 3** and **5** we identified both phage- as well as plasmid- encoded ARGs. However, no significant difference was observed in the general abundance of either phage- or plasmid- derived AMR between the pre- and post-treatment groups. On the other hand, we found that ARGs that were abundant in antibiotic-treated mice were transferred via integrons, of which some were indeed encoded on plasmids. Therefore, it can be concluded that integrons, in addition to phages and plasmids, are important in understanding dissemination mechanisms of AMR. Supporting this hypothesis, Rao *et al.* [500] reported a significant association between the class 1 integron and specific multidrug resistance patterns among *Salmonella* isolates from human, bovine and swine. While Amos *et al.* [501] found that class 1 integrons conferring resistance were distributed across a

diverse range of bacteria. In this regard, though studies have started to characterize AMR encoded on integrons, efforts have mainly been limited to amplicons generated by PCR.

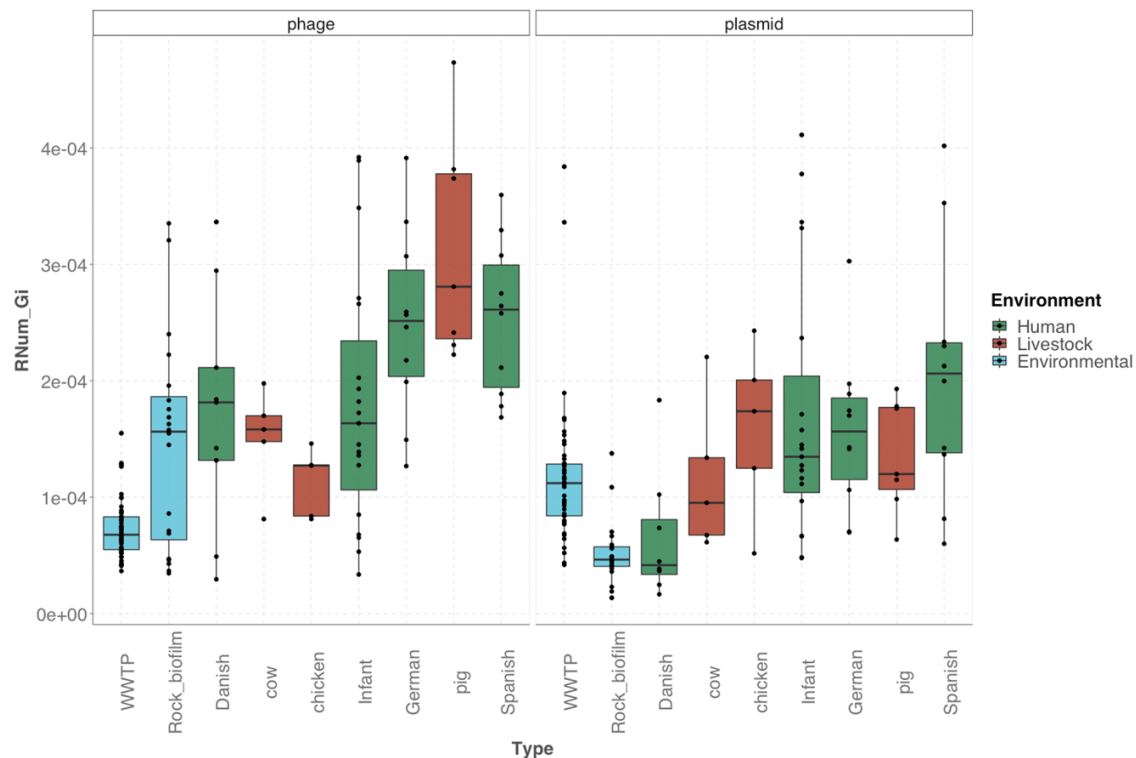


Figure 7.3: Association of MGEs with AMR in different microbial reservoirs. Box plot depicting the relative abundance of ARGs derived from MGEs, i.e., phages and plasmids.

7.4 The risk of antimicrobial resistance

Understanding the environment as a source and dissemination route for ARGs is fundamental to identifying risk scenarios for human health. Environments with a large diversity of resistance genes not generally present in the human microbiome are potential sources for recruitment of ARGs to pathogens. Additionally, other factors, such as gene mobility and host pathogenicity, are important to consider when evaluating the risk to human health. For instance, intrinsic resistance to the important last-resort antibiotic colistin has long been identified, yet low mobility via HGT has limited the clinical impact. On the other hand, the mobilized colistin resistance gene, *mcr-1*, driven by plasmid-mediated HGT has rapidly spread to various pathogenic species prevalent in both human and livestock samples [502,503]. As such, based on the framework proposed by Zhang *et al.* [502] to assess the risk of individual ARGs to human health one should consider the following criteria: enrichment in human-associated environments, gene mobility, and host pathogenicity (**Figure 7.4a**).

Therefore, to assess the risk of AMR in a One Health perspective, we need to compare ARGs for those present within human microbial reservoirs, human-associated ARGs, with those present only in the environment or in livestock. Moreover, we need to identify which ARGs are encoded on MGEs (rank II), and subsequently which MGE-derived ARGs are hosted by pathogenic bacteria such as the ESKAPEE pathogens (rank I). When we apply this to the microbial reservoirs described within this work, along with publicly available metagenomic datasets of livestock and human microbiome, the majority of the resistance genes identified are found to be human-associated (**Figure 7.4b**). Interestingly, a small number of ARGs were identified in rank I that were not found to be associated with the human microbiome. However, their association to both pathogens and MGEs makes them targets of interest for further study as they may easily become a risk to human health. Additionally, a number of ARGs were identified that could not be linked to MGEs in any of the microbial reservoirs (rank III). These immobile ARGs were found to be mainly environmentally-associated and rarely identified in the human microbiome.

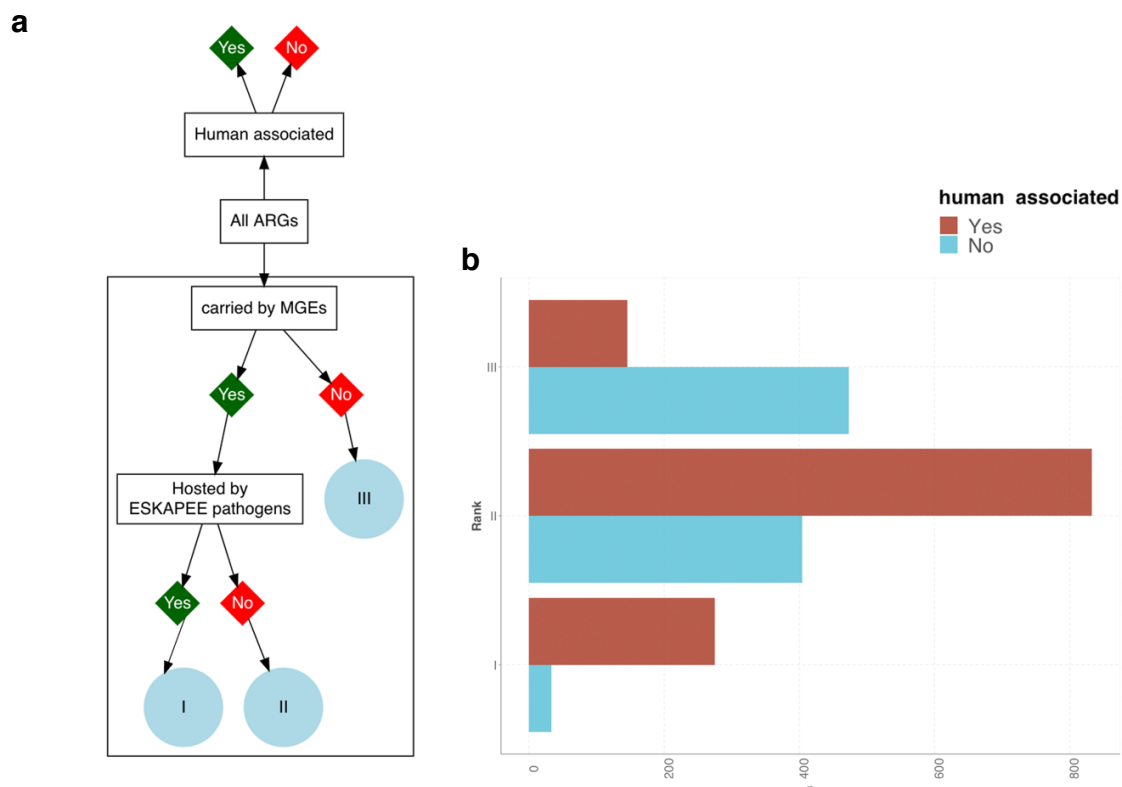


Figure 7.4: Risk of AMR. a. Decision tree for ranking the risk of AMR, with Rank III corresponding to the lowest risk and Rank I to the highest. Risk is determined by considering the localization of an ARG with MGEs and association with pathogens. **b.** Bar plot depicting the number of ARGs in each risk category in addition to association of said ARGs with the human microbiome.

Furthermore, in addition to the above described ESKAPEE pathogens, we should consider other known pathogens representing a risk to human health. For example, the WHO provides an extended list of pathogens from medium to critical priority for research and development of new antibiotics. In addition to the ESKAPEE pathogens these include *Mycobacterium tuberculosis*, *Helicobacter pylori*, *Campylobacter* species, *Salmonella* species, *Neisseria gonorrhoeae*, *Streptococcus pneumonia*, *Haemophilus influenzae* and *Shigella* species. Interestingly, many of these pathogens have been found to encode resistance genes within our microbial reservoirs. For instance, beta-lactam (e.g. carbapenem and cephalosporin) and aminoglycoside resistant *Pseudomonas aeruginosa* was observed within the WWTP, while multidrug resistant Enterobacteriaceae genera were identified in the human microbiome, including the infant microbiome. Interestingly, vancomycin resistant *Enterococcus faecium* was found within the GFS biofilms, consequently revealing the presence of a through WHO classified high priority resistance pathogen within a pristine environment.

In conclusion, as we move into an era of extensive molecular surveillance of antimicrobial resistance, with an increase in metagenomic datasets regarding a variety of microbial reservoirs, it is important to interpret the risks of ARGs rather than simply document their presence and concentration. As such, by expanding on metagenomic studies, each further study can provide valuable applications by quantifying the risk of ARGs to effectively prevent the emergence and the transmission of ARGs into human pathogens.

7.5 Further perspectives on understanding One Health reservoirs

Sequence-based metagenomics provides a comprehensive perspective on all ARGs within different microbial reservoirs. However, for a comprehensive One Health perspective it is of importance to not only identify ARGs within individual biomes but to also compare the resistomes of different microbial reservoirs. Within this work we have investigated the resistome of various microbial reservoirs spanning the One Health triad and have made further inroads to compare the identity and prevalence of ARGs both within and across biomes.

Nonetheless, one major challenge still faced by all One Health studies is attributing the directionality of ARGs between various metagenomes. With rare exceptions it is impossible to accurately attribute directionality of transmission due to limitations of the existing methods. Studies have based evidence on similarity of bacterial and/or plasmid sequences and their

ARGs, on the co-occurrence. However, these overlapping patterns do not consider co-colonization from a shared source, nor allow for interpretation of directionality

As an exception, Mather *et al.* [504] quantified the relative contributions of animal- and human-derived multidrug resistant *Salmonella* isolates using the phylogenetic association of the bacterium and its antimicrobial resistance through the course of an epidemic. Consequently, they determined that there was only a limited transmission in either direction, while the bacterium and its resistance genes were largely maintained separately within animal and human populations [504]. Collectively, to accurately reconstruct patterns of transmission, especially the directionality of said transmission, one needs to further combine both (meta)genomic data analysis, including phylogenetic analysis, with epidemiological approaches [505]. What appears evident is that each AMR reservoir may affect another. This is further compounded by the recent discovery of giant extrachromosomal elements such as “borgs” in *Methanoperedens* archaea, which may be capable of augmenting microbial activity by encoding putative resistance genes and also via HGT [506]. Thereby, understanding the interactions/mechanisms and role of each component contributing to the spread of AMR is a critical step in monitoring this ultimate challenge to human health and wellbeing. Therefore, recognizing the One Health reservoirs of antimicrobial resistance is an important first-step towards this goal. Several methods exist both *in vitro* and *in silico* to identify the potential resistance genes and categories found in commensal microorganisms alongside well-characterized pathogens. However, future endeavors including molecular validation of identified AMR with the help of meta-omics will be required. Furthermore, combined methods incorporating the identity of ARGs, modes of transmission and integration into the individual reservoirs, alongside crossover mechanisms may be needed for comprehensive characterization of AMR dissemination mechanisms.

Finally, Given the possibility of interactions between humans, livestock, animals, and the environment, future studies on human health and disease will benefit from the consequential incorporation of One Health reservoirs into all aspects of studies regarding antimicrobial resistance.

Bibliography

1. D'Costa VM, King CE, Kalan L, Morar M, Sung WWL, Schwarz C, et al. Antibiotic resistance is ancient. *Nature*. 2011;477:457–61.
2. Wright GD. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nat Rev Microbiol*. 2007;5:175–86.
3. Zhang X-S, Li J, Krautkramer KA, Badri M, Battaglia T, Borbet TC, et al. Antibiotic-induced acceleration of type 1 diabetes alters maturation of innate intestinal immunity. *eLife Sciences Publications, Ltd*; 2018;7:e37816.
4. Roubaud-Baudron C, Ruiz VE, Swan AM Jr, Vallance BA, Ozkul C, Pei Z, et al. Long-Term Effects of Early-Life Antibiotic Exposure on Resistance to Subsequent Bacterial Infection. *MBio* [Internet]. 2019;10. Available from: <http://dx.doi.org/10.1128/mBio.02820-19>
5. O'Neill J. Antimicrobial resistance: tackling a crisis for the health and wealth of nations. Review on antimicrobial resistance. 2014;
6. Brogan DM, Mossialos E. A critical analysis of the review on antimicrobial resistance report and the infectious disease financing facility. *Global Health*. 2016;12:8.
7. Reygaert WC. An overview of the antimicrobial resistance mechanisms of bacteria. *AIMS Microbiol*. 2018;4:482–501.
8. Blair JMA, Richmond GE, Piddock LJV. Multidrug efflux pumps in Gram-negative bacteria and their role in antibiotic resistance. *Future Microbiol*. 2014;9:1165–77.
9. Karaman R, Jubeh B, Breijyeh Z. Resistance of Gram-Positive Bacteria to Current Antibacterial Agents and Overcoming Approaches. *Molecules* [Internet]. 2020;25. Available from: <http://dx.doi.org/10.3390/molecules25122888>
10. Kumar A, Schweizer HP. Bacterial resistance to antibiotics: active efflux and reduced uptake. *Adv Drug Deliv Rev*. 2005;57:1486–513.
11. Mah T-F. Biofilm-specific antibiotic resistance. *Future Microbiol*. 2012;7:1061–72.
12. Redgrave LS, Sutton SB, Webber MA, Piddock LJV. Fluoroquinolone resistance: mechanisms, impact on bacteria, and role in evolutionary success. *Trends Microbiol*. 2014;22:438–45.
13. Huovinen P, Sundström L, Swedberg G, Sköld O. Trimethoprim and sulfonamide resistance. *Antimicrob Agents Chemother*. 1995;39:279–89.
14. Martinez JL. General principles of antibiotic resistance in bacteria. *Drug Discov Today Technol*. 2014;11:33–9.
15. Davies J, Davies D. Origins and evolution of antibiotic resistance. *Microbiol Mol Biol Rev*. 2010;74:417–33.
16. von Wintersdorff CJH, Penders J, van Niekerk JM, Mills ND, Majumder S, van Alphen LB, et al. Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal

Gene Transfer. *Front Microbiol.* 2016;7:173.

17. Leclerc QJ, Lindsay JA, Knight GM. Mathematical modelling to study the horizontal transfer of antimicrobial resistance genes in bacteria: current state of the field and recommendations. *J R Soc Interface.* 2019;16:20190260.

18. Lopatkin AJ, Huang S, Smith RP, Srimani JK, Sysoeva TA, Bewick S, et al. Antibiotics as a selective driver for conjugation dynamics. *Nat Microbiol.* 2016;1:16044.

19. Bello-López JM, Cabrero-Martínez OA, Ibáñez-Cervantes G, Hernández-Cortez C, Pelcastre-Rodríguez LI, Gonzalez-Avila LU, et al. Horizontal Gene Transfer and Its Association with Antibiotic Resistance in the Genus *Aeromonas* spp. *Microorganisms* [Internet]. 2019;7. Available from: <http://dx.doi.org/10.3390/microorganisms7090363>

20. MacLean RC, San Millan A. The evolution of antibiotic resistance. *Science.* 2019;365:1082–3.

21. Carattoli A. Plasmids and the spread of resistance. *Int J Med Microbiol.* 2013;303:298–304.

22. Bassetti M, Righi E, Canelutti A, Graziano E, Russo A. Multidrug-resistant *Klebsiella pneumoniae*: challenges for treatment, prevention and infection control. *Expert Rev Anti Infect Ther.* 2018;16:749–61.

23. Huang T-W, Chen T-L, Chen Y-T, Lauderdale T-L, Liao T-L, Lee Y-T, et al. Copy Number Change of the NDM-1 sequence in a multidrug-resistant *Klebsiella pneumoniae* clinical isolate. *PLoS One.* 2013;8:e62774.

24. Salinas L, Cárdenas P, Johnson TJ, Vasco K, Graham J, Trueba G. Diverse Commensal *Escherichia coli* Clones and Plasmids Disseminate Antimicrobial Resistance Genes in Domestic Animals and Children in a Semirural Community in Ecuador. *mSphere* [Internet]. 2019;4. Available from: <http://dx.doi.org/10.1128/mSphere.00316-19>

25. Porse A, Schou TS, Munck C, Ellabaan MMH, Sommer MOA. Biochemical mechanisms determine the functional compatibility of heterologous genes. *Nat Commun.* 2018;9:522.

26. Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: building the web of life. *Nat Rev Genet.* 2015;16:472–82.

27. Partridge SR, Kwong SM, Firth N, Jensen SO. Mobile Genetic Elements Associated with Antimicrobial Resistance. *Clin Microbiol Rev* [Internet]. 2018;31. Available from: <http://dx.doi.org/10.1128/CMR.00088-17>

28. Stalder T, Barraud O, Casellas M, Dagot C, Ploy M-C. Integron involvement in environmental spread of antibiotic resistance. *Front Microbiol.* 2012;3:119.

29. Buongiorno Pereira M, Österlund T, Eriksson KM, Backhaus T, Axelson-Fisk M, Kristiansson E. A comprehensive survey of integron-associated genes present in metagenomes. *BMC Genomics.* 2020;21:495.

30. Liapis E, Bour M, Triponney P, Jové T, Zahar J-R, Valot B, et al. Identification of diverse integron and Plasmid structures carrying a novel carbapenemase among *Pseudomonas* species. *Front Microbiol.* *Frontiers Media SA;* 2019;10:404.

31. Rowe-Magnus DA, Guerout A-M, Ploncard P, Dychinco B, Davies J, Mazel D. The evolutionary history of chromosomal super-integrations provides an ancestry for multiresistant integrations. *Proc Natl Acad Sci U S A. National Academy of Sciences*; 2001;98:652–7.
32. Colavecchio A, Cadieux B, Lo A, Goodridge LD. Bacteriophages Contribute to the Spread of Antibiotic Resistance Genes among Foodborne Pathogens of the Enterobacteriaceae Family - A Review. *Front Microbiol.* 2017;8:1108.
33. Oh J-H, Alexander LM, Pan M, Schueler KL, Keller MP, Attie AD, et al. Dietary Fructose and Microbiota-Derived Short-Chain Fatty Acids Promote Bacteriophage Production in the Gut Symbiont *Lactobacillus reuteri*. *Cell Host Microbe.* 2019;25:273–84.e6.
34. McDaniel L, Paul JH. Effect of nutrient addition and environmental factors on prophage induction in natural populations of marine *synechococcus* species. *Appl Environ Microbiol.* 2005;71:842–50.
35. Knowles B, Silveira CB, Bailey BA, Barott K, Cantu VA, Cobián-Güemes AG, et al. Lytic to temperate switching of viral communities. *Nature.* 2016;531:466–70.
36. Dion MB, Oechslin F, Moineau S. Phage diversity, genomics and phylogeny. *Nat Rev Microbiol.* 2020;18:125–38.
37. Chiang YN, Penadés JR, Chen J. Genetic transduction by phages and chromosomal islands: The new and noncanonical. *PLoS Pathog.* 2019;15:e1007878.
38. Boolchandani M, D'Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. *Nat Rev Genet.* 2019;20:356–70.
39. Jorgensen JH, Ferraro MJ. Antimicrobial susceptibility testing: a review of general principles and contemporary practices. *Clin Infect Dis.* 2009;49:1749–55.
40. Lepage P, Leclerc MC, Joossens M, Mondot S, Blottière HM, Raes J, et al. A metagenomic insight into our gut's microbiome. *Gut.* 2013;62:146–58.
41. Hendriksen RS, Bortolaia V, Tate H, Tyson GH, Aarestrup FM, McDermott PF. Using Genomics to Track Global Antimicrobial Resistance. *Front Public Health.* 2019;7:242.
42. Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome.* 2018;6:23.
43. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, et al. CARD 2020: antibiotic resistance surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 2020;48:D517–25.
44. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother.* 2012;67:2640–4.
45. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob Agents Chemother.* 2014;58:212–20.

46. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, et al. Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrob Agents Chemother* [Internet]. 2019;63. Available from: <http://dx.doi.org/10.1128/AAC.00483-19>
47. Krawczyk PS, Lipinski L, Dziembowski A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res*. 2018;46:e35.
48. Robertson J, Nash JHE. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* [Internet]. 2018;4. Available from: <http://dx.doi.org/10.1099/mgen.0.000206>
49. Arredondo-Alonso S, Bootsma M, Hein Y, Rogers MRC, Corander J, Willems RJJ, et al. gplas: a comprehensive tool for plasmid analysis using short-read graphs. *Bioinformatics*. Oxford Academic; 2020;36:3874–6.
50. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, et al. Identifying viruses from metagenomic data by deep learning [Internet]. *arXiv [q-bio.GN]*. 2018. Available from: <http://arxiv.org/abs/1806.07810>
51. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. *PeerJ*. 2015;3:e985.
52. Amgarten D, Braga LPP, da Silva AM, Setubal JC. MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. *Front Genet*. 2018;9:304.
53. Fang Z, Tan J, Wu S, Li M, Xu C, Xie Z, et al. PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *Gigascience* [Internet]. Oxford Academic; 2019 [cited 2021 Feb 27];8. Available from: <https://academic.oup.com/gigascience/article/8/6/giz066/5521157>
54. Atlas, Ronald M. One Health: Its Origins and Future. In: Mackenzie JS, Jeggo M, Daszak P, Richt JA, editors. *One Health: The Human-Animal-Environment Interfaces in Emerging Infectious Diseases: The Concept and Examples of a One Health Approach*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013. p. 1–13.
55. Kim D-W, Cha C-J. Antibiotic resistome from the One-Health perspective: understanding and controlling antimicrobial resistance transmission. *Exp Mol Med*. 2021;53:301–9.
56. Kultima JR, Coelho LP, Forslund K, Huerta-Cepas J, Li SS, Driessen M, et al. MOCAT2: a metagenomic assembly, annotation and profiling framework. *Bioinformatics*. 2016;32:2520–3.
57. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* [Internet]. 2021;10. Available from: <http://dx.doi.org/10.7554/eLife.65088>
58. Berg G, Rybakova D, Fischer D, Cernava T, Vergès M-CC, Charles T, et al. Microbiome definition re-visited: old concepts and new challenges. *Microbiome*. 2020;8:103.
59. Trinh P, Zaneveld JR, Safranek S, Rabinowitz PM. One Health Relationships Between

- Human, Animal, and Environmental Microbiomes: A Mini-Review. *Front Public Health*. 2018;6:235.
60. Graham DW, Bergeron G, Bourassa MW, Dickson J, Gomes F, Howe A, et al. Complexities in understanding antimicrobial resistance across domesticated animal, human, and environmental systems. *Ann N Y Acad Sci*. 2019;1441:17–30.
61. Granato ET, Meiller-Legrand TA, Foster KR. The Evolution and Ecology of Bacterial Warfare. *Curr Biol*. 2019;29:R521–37.
62. Cundliffe E, Demain AL. Avoidance of suicide in antibiotic-producing microbes. *J Ind Microbiol Biotechnol*. 2010;37:643–72.
63. Tran PN, Yen M-R, Chiang C-Y, Lin H-C, Chen P-Y. Detecting and prioritizing biosynthetic gene clusters for bioactive compounds in bacteria and fungi. *Appl Microbiol Biotechnol*. 2019;103:3277–87.
64. Demain AL. Antibiotics: natural products essential to human health. *Med Res Rev*. 2009;29:821–42.
65. Penders J, Stobberingh EE, Savelkoul PHM, Wolffs PFG. The human microbiome as a reservoir of antimicrobial resistance. *Front Microbiol*. frontiersin.org; 2013;4:87.
66. Brinkac L, Voorhies A, Gomez A, Nelson KE. The Threat of Antimicrobial Resistance on the Human Microbiome. *Microb Ecol*. 2017;74:1001–8.
67. Blake DP, Hillman K, Fenlon DR, Low JC. Transfer of antibiotic resistance between commensal and pathogenic members of the Enterobacteriaceae under ileal conditions. *J Appl Microbiol*. 2003;95:428–36.
68. Stanton IC, Bethel A, Leonard AFC, Gaze WH, Garside R. What is the research evidence for antibiotic resistance exposure and transmission to humans from the environment? A systematic map protocol. *Environ Evid*. 2020;9:12.
69. Schmidt TS, Hayward MR, Coelho LP, Li SS, Costea PI, Voigt AY, et al. Extensive transmission of microbes along the gastrointestinal tract. *Elife* [Internet]. 2019;8. Available from: <http://dx.doi.org/10.7554/eLife.42693>
70. Carr VR, Witherden EA, Lee S, Shoaie S, Mullany P, Proctor GB, et al. Abundance and diversity of resistomes differ between healthy human oral cavities and gut. *Nat Commun*. 2020;11:693.
71. Ben Y, Fu C, Hu M, Liu L, Wong MH, Zheng C. Human health risk assessment of antibiotic resistance associated with antibiotic residues in the environment: A review. *Environ Res*. 2019;169:483–93.
72. Ben Maamar S, Hu J, Hartmann EM. Implications of indoor microbial ecology and evolution on antibiotic resistance. *J Expo Sci Environ Epidemiol*. Nature Publishing Group; 2019;30:1–15.
73. Graham DW, Giesen MJ, Bunce JT. Strategic Approach for Prioritising Local and Regional Sanitation Interventions for Reducing Global Antibiotic Resistance. *Water*. Multidisciplinary Digital Publishing Institute; 2018;11:27.

74. Li J, Cao J, Zhu Y-G, Chen Q-L, Shen F, Wu Y, et al. Global Survey of Antibiotic Resistance Genes in Air. *Environ Sci Technol*. 2018;52:10975–84.
75. Dueker ME, O'Mullan GD, Martínez JM, Juhl AR, Weathers KC. Onshore Wind Speed Modulates Microbial Aerosols along an Urban Waterfront. *Atmosphere*. Multidisciplinary Digital Publishing Institute; 2017;8:215.
76. Gilbert Y, Veillette M, Duchaine C. Airborne bacteria and antibiotic resistance genes in hospital rooms. *Aerobiologia*. 2010;26:185–94.
77. Yigit H, Queenan AM, Anderson GJ, Domenech-Sanchez A, Biddle JW, Steward CD, et al. Novel carbapenem-hydrolyzing beta-lactamase, KPC-1, from a carbapenem-resistant strain of *Klebsiella pneumoniae*. *Antimicrob Agents Chemother*. 2001;45:1151–61.
78. van Hout D, Verschuuren TD, Bruijning-Verhagen PCJ, Bosch T, Schürch AC, Willems RJJ, et al. Extended-spectrum beta-lactamase (ESBL)-producing and non-ESBL-producing *Escherichia coli* isolates causing bacteremia in the Netherlands (2014 - 2016) differ in clonal distribution, antimicrobial resistance gene and virulence gene content. *PLoS One*. 2020;15:e0227604.
79. Day MJ, Hopkins KL, Wareham DW, Toleman MA, Elviss N, Randall L, et al. Extended-spectrum β -lactamase-producing *Escherichia coli* in human-derived and foodchain-derived samples from England, Wales, and Scotland: an epidemiological surveillance and typing study. *Lancet Infect Dis*. 2019;19:1325–35.
80. Falgenhauer L, Imirzalioglu C, Oppong K, Akenten CW, Hogan B, Krumkamp R, et al. Detection and Characterization of ESBL-Producing *Escherichia coli* From Humans and Poultry in Ghana. *Front Microbiol*. 2018;9:3358.
81. Alegría Á, Arias-Temprano M, Fernández-Natal I, Rodríguez-Calleja JM, García-López M-L, Santos JA. Molecular Diversity of ESBL-Producing *Escherichia coli* from Foods of Animal Origin and Human Patients. *Int J Environ Res Public Health* [Internet]. 2020;17. Available from: <http://dx.doi.org/10.3390/ijerph17041312>
82. Kayastha K, Dhungel B, Karki S, Adhikari B, Banjara MR, Rijal KR, et al. Extended-Spectrum β -Lactamase-Producing *Escherichia coli* and *Klebsiella* Species in Pediatric Patients Visiting International Friendship Children's Hospital, Kathmandu, Nepal. *Infect Dis*. 2020;13:1178633720909798.
83. Leverstein-van Hall MA, Dierikx CM, Cohen Stuart J, Voets GM, van den Munckhof MP, van Essen-Zandbergen A, et al. Dutch patients, retail chicken meat and poultry share the same ESBL genes, plasmids and strains. *Clin Microbiol Infect*. 2011;17:873–80.
84. Pitout JDD, Laupland KB. Extended-spectrum beta-lactamase-producing Enterobacteriaceae: an emerging public-health concern. *Lancet Infect Dis*. 2008;8:159–66.
85. Shayan S, Bokaeian M. Detection of ESBL- and AmpC-producing *E. coli* isolates from urinary tract infections. *Adv Biomed Res*. 2015;4:220.
86. Peter-Getzlaff S, Polsfuss S, Poledica M, Hombach M, Giger J, Böttger EC, et al. Detection of AmpC beta-lactamase in *Escherichia coli*: comparison of three phenotypic confirmation assays and genetic analysis. *J Clin Microbiol*. 2011;49:2924–32.

87. Sepp E, Andreson R, Balode A, Bilozor A, Brauer A, Egorova S, et al. Phenotypic and Molecular Epidemiology of ESBL-, AmpC-, and Carbapenemase-Producing *Escherichia coli* in Northern and Eastern Europe. *Front Microbiol.* 2019;10:2465.
88. Evans BA, Hamouda A, Amyes SGB. The rise of carbapenem-resistant *Acinetobacter baumannii*. *Curr Pharm Des.* 2013;19:223–38.
89. Piperaki E-T, Tzouvelekis LS, Miriagou V, Daikos GL. Carbapenem-resistant *Acinetobacter baumannii*: in pursuit of an effective treatment. *Clin Microbiol Infect.* 2019;25:951–7.
90. New Treatment Options against Carbapenem-Resistant *Acinetobacter baumannii* Infections [Internet]. [cited 2021 Jun 24]. Available from: <https://journals.asm.org/doi/abs/10.1128/aac.01110-18>
91. Buehrle DJ, Shields RK, Clarke LG, Potoski BA, Clancy CJ, Nguyen MH. Carbapenem-Resistant *Pseudomonas aeruginosa* Bacteremia: Risk Factors for Mortality and Microbiologic Treatment Failure. *Antimicrob Agents Chemother* [Internet]. 2017;61. Available from: <http://dx.doi.org/10.1128/AAC.01243-16>
92. Meletis G, Exindari M, Vavatsi N, Sofianou D, Diza E. Mechanisms responsible for the emergence of carbapenem resistance in *Pseudomonas aeruginosa*. *Hippokratia.* 2012;16:303–7.
93. Markwart R, Willrich N, Haller S, Noll I, Koppe U, Werner G, et al. The rise in vancomycin-resistant *Enterococcus faecium* in Germany: data from the German Antimicrobial Resistance Surveillance (ARS). *Antimicrob Resist Infect Control.* 2019;8:147.
94. Kafil HS, Asgharzadeh M. Vancomycin-resistant *enterococcus faecium* and *enterococcus faecalis* isolated from education hospital of iran. *Maedica .* 2014;9:323–7.
95. Fridkin SK, Hageman JC, Morrison M, Sanza LT, Como-Sabetti K, Jernigan JA, et al. Methicillin-resistant *Staphylococcus aureus* disease in three communities. *N Engl J Med.* Mass Medical Soc; 2005;352:1436–44.
96. Boucher HW, Corey GR. Epidemiology of methicillin-resistant *Staphylococcus aureus*. *Clin Infect Dis.* academic.oup.com; 2008;46 Suppl 5:S344–9.
97. Brumfitt W, Hamilton-Miller J. Methicillin-resistant *Staphylococcus aureus*. *N Engl J Med.* Mass Medical Soc; 1989;320:1188–96.
98. Lund BC, Ernst EJ, Klepser ME. Strategies in the treatment of penicillin-resistant *Streptococcus pneumoniae*. *Am J Health Syst Pharm.* 1998;55:1987–94.
99. Jacobs MR. Drug-resistant *Streptococcus pneumoniae*: rational antibiotic choices. *Am J Med.* 1999;106:19S – 25S; discussion 48S – 52S.
100. Kariuki S, Gordon MA, Feasey N, Parry CM. Antimicrobial resistance and management of invasive *Salmonella* disease. *Vaccine.* 2015;33 Suppl 3:C21–9.
101. Cuypers WL, Jacobs J, Wong V, Klemm EJ, Deborggraeve S, Van Puyvelde S. Fluoroquinolone resistance in *Salmonella*: insights by whole-genome sequencing. *Microb Genom* [Internet]. 2018;4. Available from: <http://dx.doi.org/10.1099/mgen.0.000195>

102. Hao Chung The, Boinett C, Thanh DP, Jenkins C, Weill F-X, Howden BP, et al. Dissecting the molecular evolution of fluoroquinolone-resistant *Shigella sonnei*. *Nat Commun*. Nature Publishing Group; 2019;10:1–13.
103. Zhang W-X, Chen H-Y, Tu L-H, Xi M-F, Chen M, Zhang J. Fluoroquinolone Resistance Mechanisms in *Shigella* Isolates in Shanghai, China, Between 2010 and 2015. *Microb Drug Resist*. 2019;25:212–8.
104. Hennart M, Panunzi LG, Rodrigues C, Gaday Q, Baines SL, Barros-Pinkelning M, et al. Population genomics and antimicrobial resistance in *Corynebacterium diphtheriae*. *Genome Med*. 2020;12:107.
105. Marshall BM, Levy SB. Food animals and antimicrobials: impacts on human health. *Clin Microbiol Rev*. 2011;24:718–33.
106. Aarestrup FM. The livestock reservoir for antimicrobial resistance: a personal view on changing patterns of risks, effects of interventions and the way forward. *Philos Trans R Soc Lond B Biol Sci*. 2015;370:20140085.
107. Holman DB, Yang W, Alexander TW. Antibiotic treatment in feedlot cattle: a longitudinal study of the effect of oxytetracycline and tulathromycin on the fecal and nasopharyngeal microbiota. *Microbiome*. 2019;7:86.
108. Agga GE, Scott HM, Vinasco J, Nagaraja TG, Amachawadi RG, Bai J, et al. Effects of chlortetracycline and copper supplementation on the prevalence, distribution, and quantity of antimicrobial resistance genes in the fecal metagenome of weaned pigs. *Prev Vet Med*. 2015;119:179–89.
109. Hughes P, Heritage J, Others. Antibiotic growth-promoters in food animals. *FAO Animal Production and Health Paper*. FAO; 1997; 2004;129–52.
110. Aarestrup FM, Seyfarth AM, Emborg HD, Pedersen K, Hendriksen RS, Bager F. 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*. 2001;45:2054–9.
111. Alvarez J, Lopez G, Muellner P, de Frutos C, Ahlstrom C, Serrano T, et al. Identifying emerging trends in antimicrobial resistance using *Salmonella* surveillance data in poultry in Spain. *Transbound Emerg Dis*. 2020;67:250–62.
112. Gay E, Bour M, Cazeau G, Jarrige N, Martineau C, Madec J-Y, et al. Antimicrobial Usages and Antimicrobial Resistance in Commensal *Escherichia coli* From Veal Calves in France: Evolution During the Fattening Process. *Front Microbiol*. 2019;10:792.
113. Diaconu EL, Carfora V, Alba P, Di Matteo P, Stravino F, Buccella C, et al. Novel IncFII plasmid harbouring bla_{NDM-4} in a carbapenem-resistant *Escherichia coli* of pig origin, Italy. *J Antimicrob Chemother*. 2020;75:3475–9.
114. Irrgang A, Tausch SH, Pauly N, Grobbel M, Kaesbohrer A, Hammerl JA. First Detection of GES-5-Producing *Escherichia coli* from Livestock-An Increasing Diversity of Carbapenemases Recognized from German Pig Production. *Microorganisms* [Internet]. 2020;8. Available from: <http://dx.doi.org/10.3390/microorganisms8101593>

115. Morrison BJ, Rubin JE. Carbapenemase producing bacteria in the food supply escaping detection. *PLoS One*. 2015;10:e0126717.
116. Barza M. Potential mechanisms of increased disease in humans from antimicrobial resistance in food animals. *Clin Infect Dis*. academic.oup.com; 2002;34 Suppl 3:S123–5.
117. Rinsky JL, Nadimpalli M, Wing S, Hall D, Baron D, Price LB, et al. Livestock-associated methicillin and multidrug resistant *Staphylococcus aureus* is present among industrial, not antibiotic-free livestock operation workers in North Carolina. *PLoS One*. 2013;8:e67641.
118. Anders J, Bisha B. High-Throughput Detection and Characterization of Antimicrobial Resistant *Enterococcus* sp. Isolates from GI Tracts of European Starlings Visiting Concentrated Animal Feeding Operations. *Foods* [Internet]. 2020;9. Available from: <http://dx.doi.org/10.3390/foods9070890>
119. Islam S, Paul A, Talukder M, Roy K, Sobur A, Levy S, et al. Migratory birds travelling to Bangladesh are potential carriers of multi-drug resistant *Enterococcus* spp., *Salmonella* spp., and *Vibrio* spp. *Saudi J Biol Sci* [Internet]. 2021; Available from: <https://www.sciencedirect.com/science/article/pii/S1319562X21005258>
120. Islam MS, Sobur MA, Rahman S, Ballah FM, Levy S, Siddique MP, et al. Detection of blaTEM, blaCTX-M, blaCMY, and blaSHV Genes Among Extended-Spectrum Beta-Lactamase-Producing *Escherichia coli* Isolated from Migratory Birds Travelling to Bangladesh. *Microb Ecol* [Internet]. 2021; Available from: <http://dx.doi.org/10.1007/s00248-021-01803-x>
121. Plaza-Rodríguez C, Alt K, Grobbel M, Hammerl JA, Irrgang A, Szabo I, et al. Wildlife as Sentinels of Antimicrobial Resistance in Germany? *Front Vet Sci*. 2020;7:627821.
122. Atterby C, Börjesson S, Ny S, Järhult JD, Byfors S, Bonnedahl J. ESBL-producing *Escherichia coli* in Swedish gulls-A case of environmental pollution from humans? *PLoS One*. 2017;12:e0190380.
123. Dolejska M, Papagiannitsis CC. Plasmid-mediated resistance is going wild. *Plasmid*. 2018;99:99–111.
124. Hernández J, González-Acuña D. Anthropogenic antibiotic resistance genes mobilization to the polar regions. *Infect Ecol Epidemiol*. 2016;6:32112.
125. Dickinson AW, Power A, Hansen MG, Brandt KK, Piliposian G, Appleby P, et al. Heavy metal pollution and co-selection for antibiotic resistance: A microbial palaeontology approach. *Environ Int*. 2019;132:105117.
126. Baker-Austin C, Wright MS, Stepanauskas R, McArthur JV. Co-selection of antibiotic and metal resistance. *Trends Microbiol*. 2006;14:176–82.
127. Kraemer SA, Ramachandran A, Perron GG. Antibiotic Pollution in the Environment: From Microbial Ecology to Public Policy. *Microorganisms* [Internet]. 2019;7. Available from: <http://dx.doi.org/10.3390/microorganisms7060180>
128. Barancheshme F, Munir M. Strategies to Combat Antibiotic Resistance in the Wastewater Treatment Plants. *Front Microbiol*. 2017;8:2603.

129. Rodríguez-Molina D, Mang P, Schmitt H, Chifiriuc MC, Radon K, Wengenroth L. Do wastewater treatment plants increase antibiotic resistant bacteria or genes in the environment? Protocol for a systematic review. *Syst Rev*. 2019;8:304.
130. Alexander J, Hembach N, Schwartz T. Evaluation of antibiotic resistance dissemination by wastewater treatment plant effluents with different catchment areas in Germany. *Sci Rep*. 2020;10:8952.
131. Osińska A, Korzeniewska E, Harnisz M, Felis E, Bajkacz S, Jachimowicz P, et al. Small-scale wastewater treatment plants as a source of the dissemination of antibiotic resistance genes in the aquatic environment. *J Hazard Mater*. 2020;381:121221.
132. Bueno I, Verdugo C, Jimenez-Lopez O, Alvarez PP, Gonzalez-Rocha G, Lima CA, et al. Role of wastewater treatment plants on environmental abundance of Antimicrobial Resistance Genes in Chilean rivers. *Int J Hyg Environ Health*. 2020;223:56–64.
133. Allen HK, Donato J, Wang HH, Cloud-Hansen KA, Davies J, Handelsman J. Call of the wild: antibiotic resistance genes in natural environments. *Nat Rev Microbiol*. 2010;8:251–9.
134. Scott LC, Lee N, Aw TG. Antibiotic Resistance in Minimally Human-Impacted Environments. *Int J Environ Res Public Health* [Internet]. 2020;17. Available from: <http://dx.doi.org/10.3390/ijerph17113939>
135. Armalytė J, Skerniškytė J, Bakienė E, Krasauskas R, Šiugždinienė R, Kareivienė V, et al. Microbial Diversity and Antimicrobial Resistance Profile in Microbiota From Soils of Conventional and Organic Farming Systems. *Front Microbiol*. 2019;10:892.
136. Dantas G, Sommer MOA, Oluwasegun RD, Church GM. Bacteria subsisting on antibiotics. *Science*. 2008;320:100–3.
137. D'Costa VM, McGrann KM, Hughes DW, Wright GD. Sampling the antibiotic resistome. *Science*. 2006;311:374–7.
138. Wolters B, Jacquioud S, Sørensen SJ, Widyasari-Mehta A, Bech TB, Kreuzig R, et al. Bulk soil and maize rhizosphere resistance genes, mobile genetic elements and microbial communities are differently impacted by organic and inorganic fertilization. *FEMS Microbiol Ecol* [Internet]. 2018;94. Available from: <http://dx.doi.org/10.1093/femsec/fiy027>
139. Song M, Peng K, Jiang L, Zhang D, Song D, Chen G, et al. Alleviated Antibiotic-Resistant Genes in the Rhizosphere of Agricultural Soils with Low Antibiotic Concentration. *J Agric Food Chem*. American Chemical Society; 2020;68:2457–66.
140. Kittinger C, Kirschner A, Lipp M, Baumert R, Mascher F, Farnleitner AH, et al. Antibiotic Resistance of *Acinetobacter* spp. Isolates from the River Danube: Susceptibility Stays High. *Int J Environ Res Public Health* [Internet]. mdpi.com; 2017;15. Available from: <http://dx.doi.org/10.3390/ijerph15010052>
141. Paschoal RP, Campana EH, Corrêa LL, Montezzi LF, Barrueto LRL, da Silva IR, et al. Concentration and Variety of Carbapenemase Producers in Recreational Coastal Waters Showing Distinct Levels of Pollution. *Antimicrob Agents Chemother* [Internet]. 2017;61. Available from: <http://dx.doi.org/10.1128/AAC.01963-17>

142. Mahon BM, Brehony C, Cahill N, McGrath E, O'Connor L, Varley A, et al. Detection of OXA-48-like-producing Enterobacterales in Irish recreational water. *Sci Total Environ*. 2019;690:1–6.
143. Surette MD, Wright GD. Lessons from the Environmental Antibiotic Resistome. *Annu Rev Microbiol*. 2017;71:309–29.
144. Flach C-F, Johnning A, Nilsson I, Smalla K, Kristiansson E, Larsson DGJ. Isolation of novel IncA/C and IncN fluoroquinolone resistance plasmids from an antibiotic-polluted lake. *J Antimicrob Chemother*. 2015;70:2709–17.
145. Kristiansson E, Fick J, Janzon A, Grabic R, Rutgersson C, Weijdegård B, et al. Pyrosequencing of antibiotic-contaminated river sediments reveals high levels of resistance and gene transfer elements. *PLoS One*. 2011;6:e17038.
146. McCann CM, Christgen B, Roberts JA, Su J-Q, Arnold KE, Gray ND, et al. Understanding drivers of antibiotic resistance genes in High Arctic soil ecosystems. *Environ Int*. Elsevier; 2019;125:497–504.
147. Dancer SJ, Shears P, Platt DJ. Isolation and characterization of coliforms from glacial ice and water in Canada's High Arctic. *J Appl Microbiol*. 1997;82:597–609.
148. Nadeem SF, Gohar UF, Tahir SF, Mukhtar H, Pornpukdeewattana S, Nukthamna P, et al. Antimicrobial resistance: more than 70 years of war between humans and bacteria. *Crit Rev Microbiol*. 2020;46:578–99.
149. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J*. 2015;9:207–16.
150. Munck C, Albertsen M, Telke A, Ellabaan M, Nielsen PH, Sommer MOA. Limited dissemination of the wastewater treatment plant core resistome. *Nat Commun*. 2015;6:8452.
151. Ngoi ST, Chong CW, Ponnampalavanar SSLS, Tang SN, Idris N, Abdul Jabar K, et al. Genetic mechanisms and correlated risk factors of antimicrobial-resistant ESKAPEE pathogens isolated in a tertiary hospital in Malaysia. *Antimicrob Resist Infect Control*. 2021;10:70.
152. Van Den Broek IVF, Van Cleef B, Haenen A, Broens EM, Van der Wolf PJ, Van Den Broek MJM, et al. Methicillin-resistant *Staphylococcus aureus* in people living and working in pig farms. *Epidemiology & Infection*. Cambridge University Press; 2009;137:700–8.
153. Lewis HC, Mølbak K, Reese C, Aarestrup FM, Selchau M, Sørup M, et al. Pigs as source of methicillin-resistant *Staphylococcus aureus* CC398 infections in humans, Denmark. *Emerg Infect Dis*. Centers for Disease Control and Prevention (CDC); 2008;14:1383–9.
154. Loncaric I, Cabal Rosel A, Szostak MP, Licka T, Allerberger F, Ruppitsch W, et al. Broad-Spectrum Cephalosporin-Resistant *Klebsiella* spp. Isolated from Diseased Horses in Austria. *Animals (Basel)* [Internet]. 2020;10. Available from: <http://dx.doi.org/10.3390/ani10020332>
155. Mughini-Gras L, Dorado-García A, van Duijkeren E, van den Bunt G, Dierikx CM, Bonten MJM, et al. Attributable sources of community-acquired carriage of *Escherichia coli*

containing β -lactam antibiotic resistance genes: a population-based modelling study. *Lancet Planet Health*. 2019;3:e357–69.

156. Dahl LG, Joensen KG, Østerlund MT, Kiil K, Nielsen EM. Prediction of antimicrobial resistance in clinical *Campylobacter jejuni* isolates from whole-genome sequencing data. *Eur J Clin Microbiol Infect Dis*. 2021;40:673–82.

157. Mossong J, Mughini-Gras L, Penny C, Devaux A, Olinger C, Losch S, et al. Human *Campylobacteriosis* in Luxembourg, 2010–2013: A Case-Control Study Combined with Multilocus Sequence Typing for Source Attribution and Risk Factor Analysis. *Sci Rep*. 2016;6:20939.

158. Winokur PL, Brueggemann A, DeSalvo DL, Hoffmann L, Apley MD, Uhlenhopp EK, et al. Animal and human multidrug-resistant, cephalosporin-resistant salmonella isolates expressing a plasmid-mediated CMY-2 AmpC beta-lactamase. *Antimicrob Agents Chemother*. 2000;44:2777–83.

159. Genomic Investigation of the Emergence of Invasive Multidrug-Resistant *Salmonella enterica* Serovar Dublin in Humans and Animals in Canada [Internet]. [cited 2021 Jul 2]. Available from: <https://journals.asm.org/doi/abs/10.1128/aac.00108-19>

160. Zhang L, Fu Y, Xiong Z, Ma Y, Wei Y, Qu X, et al. Highly Prevalent Multidrug-Resistant *Salmonella* From Chicken and Pork Meat at Retail Markets in Guangdong, China. *Front Microbiol*. 2018;9:2104.

161. Dionisi AM, Lucarelli C, Benedetti I, Owczarek S, Luzzi I. Molecular characterisation of multidrug-resistant *Salmonella enterica* serotype Infantis from humans, animals and the environment in Italy. *Int J Antimicrob Agents*. 2011;38:384–9.

162. Forslund K, Sunagawa S, Coelho LP, Bork P. Metagenomic insights into the human gut resistome and the forces that shape it. *Bioessays*. 2014;36:316–29.

163. Busi SB, de Nies L, Habier J, Wampach L, Fritz JV, Heintz-Buschart A, et al. Persistence of birth mode-dependent effects on gut microbiome composition, immune system stimulation and antimicrobial resistance during the first year of life. *ISME Communications*. Nature Publishing Group; 2021;1:1–12.

164. Casaburi G, Duar RM, Brown H, Mitchell RD, Kazi S, Chew S, et al. Metagenomic insights of the infant microbiome community structure and function across multiple sites in the United States. *Sci Rep*. Nature Publishing Group; 2021;11:1–12.

165. Wampach L, Heintz-Buschart A, Fritz JV, Ramiro-Garcia J, Habier J, Herold M, et al. Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. *Nat Commun*. 2018;9:5091.

166. Gaeta NC, Bean E, Miles AM, de Carvalho DUOG, Alemán MAR, Carvalho JS, et al. A Cross-Sectional Study of Dairy Cattle Metagenomes Reveals Increased Antimicrobial Resistance in Animals Farmed in a Heavy Metal Contaminated Environment. *Front Microbiol*. 2020;11:590325.

167. Skarżyńska M, Leekitcharoenphon P, Hendriksen RS, Aarestrup FM, Wasyl D. A metagenomic glimpse into the gut of wild and domestic animals: Quantification of

antimicrobial resistance and more. *PLoS One*. 2020;15:e0242987.

168. Duarte ASR, Röder T, Van Gompel L, Petersen TN, Hansen RB, Hansen IM, et al. Metagenomics-Based Approach to Source-Attribution of Antimicrobial Resistance Determinants - Identification of Reservoir Resistome Signatures. *Front Microbiol*. 2020;11:601407.

169. Van Gompel L, Luiken REC, Hansen RB, Munk P, Bouwknecht M, Heres L, et al. Description and determinants of the faecal resistome and microbiome of farmers and slaughterhouse workers: A metagenome-wide cross-sectional study. *Environ Int*. 2020;143:105939.

170. Ma L, Xia Y, Li B, Yang Y, Li L-G, Tiedje JM, et al. Metagenomic Assembly Reveals Hosts of Antibiotic Resistance Genes and the Shared Resistome in Pig, Chicken, and Human Feces. *Environ Sci Technol*. ACS Publications; 2016;50:420–7.

171. Wang Y, Hu Y, Liu F, Cao J, Lv N, Zhu B, et al. Integrated metagenomic and metatranscriptomic profiling reveals differentially expressed resistomes in human, chicken, and pig gut microbiomes. *Environ Int*. 2020;138:105649.

172. Noyes NR, Yang X, Linke LM, Magnuson RJ, Cook SR, Zaheer R, et al. Characterization of the resistome in manure, soil and wastewater from dairy and beef production systems. *Sci Rep*. 2016;6:24645.

173. Sukhum KV, Vargas RC, Boolchandani M, D'Souza AW, Patel S, Kesaraju A, et al. Manure Microbial Communities and Resistance Profiles Reconfigure after Transition to Manure Pits and Differ from Those in Fertilized Field Soil. *MBio* [Internet]. 2021;12. Available from: <http://dx.doi.org/10.1128/mBio.00798-21>

174. Smith SD, Colgan P, Yang F, Rieke EL, Soupir ML, Moorman TB, et al. Investigating the dispersal of antibiotic resistance associated genes from manure application to soil and drainage waters in simulated agricultural farmland systems. *PLoS One*. 2019;14:e0222470.

175. Qian X, Gunturu S, Guo J, Chai B, Cole JR, Gu J, et al. Metagenomic analysis reveals the shared and distinct features of the soil resistome across tundra, temperate prairie, and tropical ecosystems. *Microbiome*. 2021;9:108.

176. Ju F, Li B, Ma L, Wang Y, Huang D, Zhang T. Antibiotic resistance genes and human bacterial pathogens: Co-occurrence, removal, and enrichment in municipal sewage sludge digesters. *Water Res*. 2016;91:1–10.

177. Hendriksen RS, Munk P, Njage P, van Bunnik B, McNally L, Lukjancenko O, et al. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nat Commun*. nature.com; 2019;10:1124.

178. Ma L, Li B, Jiang X-T, Wang Y-L, Xia Y, Li A-D, et al. Catalogue of antibiotic resistome and host-tracking in drinking water deciphered by a large scale survey. *Microbiome*. 2017;5:154.

179. Bai Y, Ruan X, Xie X, Yan Z. Antibiotic resistome profile based on metagenomics in raw surface drinking water source and the influence of environmental factor: A case study in Huaihe River Basin, China. *Environ Pollut*. 2019;248:438–47.

180. Pal C, Bengtsson-Palme J, Kristiansson E, Larsson DGJ. The structure and diversity of human, animal and environmental resistomes. *Microbiome*. 2016;4:54.
181. Li B, Yang Y, Ma L, Ju F, Guo F, Tiedje JM, et al. Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. *ISME J*. 2015;9:2490–502.
182. Beceiro A, Tomás M, Bou G. Antimicrobial resistance and virulence: a successful or deleterious association in the bacterial world? *Clin Microbiol Rev*. 2013;26:185–230.
183. Wu H-J, Wang AH-J, Jennings MP. Discovery of virulence factors of pathogenic bacteria. *Curr Opin Chem Biol*. 2008;12:93–101.
184. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, et al. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res*. 2005;33:D325–8.
185. Finlay BB, Falkow S. Common themes in microbial pathogenicity revisited. *Microbiol Mol Biol Rev*. 1997;61:136–69.
186. Chakraborty A, Ghosh S, Chowdhary G, Maulik U, Chakrabarti S. DBETH: a Database of Bacterial Exotoxins for Human. *Nucleic Acids Res*. 2012;40:D615–20.
187. Schiavo G, van der Goot FG. The bacterial toxin toolkit. *Nat Rev Mol Cell Biol*. 2001;2:530–7.
188. Martínez JL, Baquero F. Interactions among strategies associated with bacterial infection: pathogenicity, epidemicity, and antibiotic resistance. *Clin Microbiol Rev*. 2002;15:647–79.
189. Sommer MOA, Dantas G, Church GM. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science*. 2009;325:1128–31.
190. Burrus V, Waldor MK. Shaping bacterial genomes with integrative and conjugative elements. *Res Microbiol*. 2004;155:376–86.
191. Mediavilla JR, Patrawalla A, Chen L, Chavda KD, Mathema B, Vinnard C, et al. Colistin- and Carbapenem-Resistant *Escherichia coli* Harboring *mcr-1* and *bla*NDM-5, Causing a Complicated Urinary Tract Infection in a Patient from the United States. *MBio* [Internet]. 2016;7. Available from: <http://dx.doi.org/10.1128/mBio.01191-16>
192. Tsai Y-K, Fung C-P, Lin J-C, Chen J-H, Chang F-Y, Chen T-L, et al. *Klebsiella pneumoniae* outer membrane porins OmpK35 and OmpK36 play roles in both antimicrobial resistance and virulence. *Antimicrob Agents Chemother*. 2011;55:1485–93.
193. Barbosa TM, Levy SB. Differential expression of over 60 chromosomal genes in *Escherichia coli* by constitutive expression of MarA. *J Bacteriol*. 2000;182:3467–74.
194. Cabot G, Zamorano L, Moyà B, Juan C, Navas A, Blázquez J, et al. Evolution of *Pseudomonas aeruginosa* Antimicrobial Resistance and Fitness under Low and High Mutation Rates. *Antimicrob Agents Chemother*. 2016;60:1767–78.
195. Elie-Fadrosh EA, Ivanova NN, Woyke T, Kyrpides NC. Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nat Microbiol*. 2016;1:15032.

196. Alteio LV, Schulz F, Seshadri R, Varghese N, Rodriguez-Reillo W, Ryan E, et al. Complementary Metagenomic Approaches Improve Reconstruction of Microbial Diversity in a Forest Soil. *mSystems* [Internet]. 2020;5. Available from: <http://dx.doi.org/10.1128/mSystems.00768-19>
197. Kleinheinz KA, Joensen KG, Larsen MV. Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and *E. coli* virulence genes in bacteriophage and prophage nucleotide sequences. *Bacteriophage*. 2014;4:e27943.
198. Yin X, Jiang X-T, Chai B, Li L, Yang Y, Cole JR, et al. ARGs-OAP v2.0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. *Bioinformatics*. 2018;34:2263–70.
199. Gupta A, Kapil R, Dhakan DB, Sharma VK. MP3: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data. *PLoS One*. 2014;9:e93907.
200. Garg A, Gupta D. VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC Bioinformatics*. 2008;9:62.
201. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods*. 2018;15:962–8.
202. Köster J, Rahmann S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*. 2018;34:3600.
203. Anaconda INC. Conda [Internet]. [cited 2018]. Available from: <https://anaconda.com>
204. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 2010;11:119.
205. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res*. 2013;41:e121.
206. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol*. 2019;37:420–3.
207. Zhang R, Ou H-Y, Zhang C-T. DEG: a database of essential genes. *Nucleic Acids Res*. 2004;32:D271–2.
208. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006;22:1658–9.
209. Rowe WPM, Winn MD. Indexed variation graphs for efficient and accurate resistome profiling. *Bioinformatics*. 2018;34:3601–8.
210. Li W, Wooley JC, Godzik A. Probing metagenomics by rapid cluster analysis of very large datasets. *PLoS One*. 2008;3:e3375.
211. Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, et al. Integrated

multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat Microbiol.* 2016;2:16180.

212. Finn RD, Coghill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 2016;44:D279–85.

213. Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, et al. The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res.* 2005;33:D71–4.

214. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000;28:27–30.

215. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* 2004;32:D438–42.

216. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 2000;28:45–8.

217. Breiman L. Random Forests. *Mach Learn.* 2001;45:5–32.

218. Hastie T, Tibshirani R, Friedman J. Random Forests. *The Elements of Statistical Learning.* Springer; 2009. p. 567–603.

219. Chen Z, Zhao P, Li F, Leier A, Marquez-Lago TT, Wang Y, et al. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics.* 2018;34:2499–502.

220. McKinney W, Others. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference.* Austin, TX; 2010. p. 51–6.

221. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods.* 2020;17:261–72.

222. Pedregosa F. Scikit-learn: Machine Learning in Python. *J Mach Learn Res.* 2011;12:2825–30.

223. Wishart D, Arndt D, Pon A, Sajed T, Guo AC, Djoumbou Y, et al. T3DB: the toxic exposome database. *Nucleic Acids Res.* 2015;43:D928–34.

224. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, et al. The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother.* 2013;57:3348–57.

225. Liu B, Pop M. ARDB--Antibiotic Resistance Genes Database. *Nucleic Acids Res.* 2009;37:D443–7.

226. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 2019;47:D506–15.

227. Pierce NT, Irber L, Reiter T, Brooks P, Brown CT. Large-scale sequence comparisons with sourmash. *F1000Res*. 2019;8:1006.
228. Milani C, Ticinesi A, Gerritsen J, Nouvenne A, Lugli GA, Mancabelli L, et al. Gut microbiota composition and *Clostridium difficile* infection in hospitalized elderly individuals: a metagenomic study. *Sci Rep*. 2016;6:25945.
229. Tett A, Pasolli E, Farina S, Truong DT, Asnicar F, Zolfo M, et al. Unexplored diversity and strain-level structure of the skin microbiome associated with psoriasis. *NPJ Biofilms Microbiomes*. 2017;3:14.
230. Bedarf JR, Hildebrand F, Coelho LP, Sunagawa S, Bahram M, Goeser F, et al. Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naïve Parkinson's disease patients. *Genome Med*. 2017;9:39.
231. Narayanasamy S, Jarosz Y, Muller EEL, Heintz-Buschart A, Herold M, Kaysen A, et al. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol*. 2016;17:260.
232. Liao Y, Smyth GK, Shi W. featureCounts: An efficient general-purpose program for assigning sequence reads to genomic features [Internet]. *arXiv [q-bio.GN]*. 2013. Available from: <http://arxiv.org/abs/1305.3347>
233. Hu Y, Yang X, Qin J, Lu N, Cheng G, Wu N, et al. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nat Commun*. 2013;4:2151.
234. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15:550.
235. Trepod CM, Mott JE. Identification of the *Haemophilus influenzae* tolC gene by susceptibility profiles of insertionally inactivated efflux pump mutants. *Antimicrob Agents Chemother*. 2004;48:1416–8.
236. Chaudhuri D, Roy Chowdhury A, Biswas B, Chakravorty D. *Salmonella* Typhimurium Infection Leads to Colonization of the Mouse Brain and Is Not Completely Cured With Antibiotics. *Front Microbiol*. 2018;9:1632.
237. Shah D, Dang M-D, Hasbun R, Koo HL, Jiang Z-D, DuPont HL, et al. *Clostridium difficile* infection: update on emerging antibiotic treatment options and antibiotic resistance. *Expert Rev Anti Infect Ther*. 2010;8:555–64.
238. Mertsalmi TH, Pekkonen E, Scheperjans F. Antibiotic exposure and risk of Parkinson's disease in Finland: A nationwide case-control study. *Mov Disord*. 2020;35:431–42.
239. Boerma T, Ronsmans C, Melesse DY, Barros AJD, Barros FC, Juan L, et al. Global epidemiology of use of and disparities in caesarean sections. *Lancet*. 2018;392:1341–8.
240. Betrán AP, Ye J, Moller A-B, Zhang J, Gülmezoglu AM, Torloni MR. The Increasing Trend in Caesarean Section Rates: Global, Regional and National Estimates: 1990-2014. *PLoS One*. 2016;11:e0148343.
241. Gensollen T, Iyer SS, Kasper DL, Blumberg RS. How colonization by microbiota in early

life shapes the immune system. *Science*. 2016;352:539–44.

242. Wang S, Ryan CA, Boyaval P, Dempsey EM, Ross RP, Stanton C. Maternal Vertical Transmission Affecting Early-life Microbiota Development. *Trends Microbiol*. 2020;28:28–45.

243. Guittar J, Shade A, Litchman E. Trait-based community assembly and succession of the infant gut microbiome. *Nat Commun*. 2019;10:512.

244. Sandall J, Tribe RM, Avery L, Mola G, Visser GH, Homer CS, et al. Short-term and long-term effects of caesarean section on the health of women and children. *Lancet*. 2018;392:1349–57.

245. Korpela K, Salonen A, Hickman B, Kunz C, Sprenger N, Kukkonen K, et al. Fucosylated oligosaccharides in mother's milk alleviate the effects of caesarean birth on infant gut microbiota. *Sci Rep. nature.com*; 2018;8:13757.

246. Asnicar F, Manara S, Zolfo M, Truong DT, Scholz M, Armanini F, et al. Studying vertical microbiome transmission from mothers to infants by strain-level metagenomic profiling. *mSystems* [Internet]. American Society for Microbiology; 2017;2. Available from: <https://journals.asm.org/doi/10.1128/mSystems.00164-16>

247. Bouhanick B, Ehlinger V, Delpierre C, Chamontin B, Lang T, Kelly-Irving M. Mode of delivery at birth and the metabolic syndrome in midlife: the role of the birth environment in a prospective birth cohort study. *BMJ Open. British Medical Journal Publishing Group*; 2014;4:e005031.

248. Magne F, Puchi Silva A, Carvajal B, Gotteland M. The Elevated Rate of Cesarean Section and Its Contribution to Non-Communicable Chronic Diseases in Latin America: The Growing Involvement of the Microbiota. *Front Pediatr*. 2017;5:192.

249. Loo EXL, Sim JZT, Loy SL, Goh A, Chan YH, Tan KH, et al. Associations between caesarean delivery and allergic outcomes: Results from the GUSTO study. *Ann Allergy Asthma Immunol*. 2017;118:636–8.

250. Tamburini S, Shen N, Wu HC, Clemente JC. The microbiome in early life: implications for health outcomes. *Nat Med*. 2016;22:713–22.

251. Shao Y, Forster SC, Tsaliki E, Vervier K, Strang A, Simpson N, et al. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature*. 2019;574:117–21.

252. Stearns JC, Simioni J, Gunn E, McDonald H, Holloway AC, Thabane L, et al. Intrapartum antibiotics for GBS prophylaxis alter colonization patterns in the early infant gut microbiome of low risk infants. *Sci Rep*. 2017;7:16527.

253. Noval Rivas M, Crother TR, Arditi M. The microbiome in asthma. *Curr Opin Pediatr*. 2016;28:764–71.

254. Martinez KA 2nd, Devlin JC, Lacher CR, Yin Y, Cai Y, Wang J, et al. Increased weight gain by C-section: Functional significance of the primordial microbiome. *Sci Adv*. 2017;3:eaao1874.

255. Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, et al. Dynamics

- and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host Microbe*. 2015;17:690–703.
256. de Muinck EJ, Trosvik P. Individuality and convergence of the infant gut microbiota during the first year of life. *Nat Commun*. 2018;9:2233.
257. Ferretti P, Pasolli E, Tett A, Asnicar F, Gorfer V, Fedi S, et al. Mother-to-Infant Microbial Transmission from Different Body Sites Shapes the Developing Infant Gut Microbiome. *Cell Host Microbe*. 2018;24:133–45.e5.
258. Wopereis H, Oozeer R, Knipping K, Belzer C, Knol J. The first thousand days - intestinal microbiology of early life: establishing a symbiosis. *Pediatr Allergy Immunol*. 2014;25:428–38.
259. Baumann-Dudenhoeffer AM, D'Souza AW, Tarr PI, Warner BB, Dantas G. Infant diet and maternal gestational weight gain predict early metabolic maturation of gut microbiomes. *Nat Med*. 2018;24:1822–9.
260. Bokulich NA, Chung J, Battaglia T, Henderson N, Jay M, Li H, et al. Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci Transl Med*. 2016;8:343ra82.
261. Mueller NT, Bakacs E, Combellick J, Grigoryan Z, Dominguez-Bello MG. The infant microbiome development: mom matters. *Trends Mol Med*. 2015;21:109–17.
262. Stokholm J, Thorsen J, Chawes BL, Schjørring S, Krogfelt KA, Bønnelykke K, et al. Cesarean section changes neonatal gut colonization. *J Allergy Clin Immunol*. 2016;138:881–9.e2.
263. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, et al. Succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci U S A*. National Acad Sciences; 2011;108 Suppl 1:4578–85.
264. Heintz-Buschart A, Wilmes P. Human Gut Microbiome: Function Matters. *Trends Microbiol*. 2018;26:563–74.
265. Vatanen T, Kostic AD, d'Hennezel E, Siljander H, Franzosa EA, Yassour M, et al. Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell*. 2016;165:842–53.
266. Jennewein MF, Butler AL, Alter G. Neonate-omics: Charting the Unknown Immune Response in Early Life. *Cell*. 2018. p. 1051–3.
267. Romano-Keeler J, Weitkamp J-H. Maternal influences on fetal microbial colonization and immune development. *Pediatr Res*. 2015;77:189–95.
268. Spencer SJ, Martin S, Mouihate A, Pittman QJ. Early-life immune challenge: defining a critical window for effects on adult responses to immune challenge. *Neuropsychopharmacology*. 2006;31:1910–8.
269. Torow N, Hornef MW. The Neonatal Window of Opportunity: Setting the Stage for Life-Long Host-Microbial Interaction and Immune Homeostasis. *J Immunol*. 2017;198:557–63.
270. Gopalakrishna KP, Macadangdang BR, Rogers MB, Tometich JT, Firek BA, Baker R, et

al. Maternal IgA protects against the development of necrotizing enterocolitis in preterm infants. *Nat Med*. 2019;25:1110–5.

271. Perez-Muñoz ME, Arrieta M-C, Ramer-Tait AE, Walter J. A critical assessment of the “sterile womb” and “in utero colonization” hypotheses: implications for research on the pioneer infant microbiome. *Microbiome*. 2017;5:48.

272. Olin A, Henckel E, Chen Y, Lakshmikanth T, Pou C, Mikes J, et al. Stereotypic Immune System Development in Newborn Children. *Cell*. 2018;174:1277–92.e14.

273. Levan SR, Stamnes KA, Lin DL, Panzer AR, Fukui E, McCauley K, et al. Elevated faecal 12,13-diHOME concentration in neonates at high risk for asthma is produced by gut bacteria and impedes immune tolerance. *Nat Microbiol*. 2019;4:1851–61.

274. Ravi A, Avershina E, Foley SL, Ludvigsen J, Storrø O, Øien T, et al. The commensal infant gut meta-mobilome as a potential reservoir for persistent multidrug resistance integrons. *Sci Rep*. 2015;5:15317.

275. Wampach L, Heintz-Buschart A, Hogan A, Muller EEL, Narayanasamy S, Laczny CC, et al. Colonization and Succession within the Human Gut Microbiome by Archaea, Bacteria, and Microeukaryotes during the First Year of Life. *Front Microbiol*. 2017;8:738.

276. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. 2015;31:1674–6.

277. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*. 2019;7:e7359.

278. Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*. 2016;32:605–7.

279. Laczny CC, Sternal T, Plugaru V, Gawron P, Atashpendar A, Margossian HH, et al. VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome*. 2015;3:1.

280. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol*. 2018;3:836–43.

281. Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh H-J, Cuenca M, et al. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun*. [nature.com; 2019;10:1014](https://doi.org/10.1038/s41467-019-1014-1).

282. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. [genome.cshlp.org; 2015;25:1043–55](https://doi.org/10.1101/012168).

283. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*. Oxford Academic; 2019;36:1925–7.

284. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. academic.oup.com; 2014;30:2068–9.
285. Yoon B-J. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr Genomics*. 2009;10:402–15.
286. Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome*. microbiomejournal.biomedcentral ...; 2020;8:90.
287. Bolduc B, Jang HB, Doulier G, You Z-Q, Roux S, Sullivan MB. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ*. peerj.com; 2017;5:e3243.
288. Song W, Wemheuer B, Zhang S, Steensen K, Thomas T. MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and phylogenetic approaches. *Microbiome*. 2019;7:36.
289. Computing R, Others. R: A language and environment for statistical computing. Vienna: R Core Team [Internet]. 2013; Available from: <https://www.yumpu.com/en/document/view/6853895/r-a-language-and-environment-for-statistical-computing>
290. Blighe K. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling [Internet]. Github; [cited 2021 Sep 30]. Available from: <https://github.com/kevinblighe/EnhancedVolcano>
291. Friendly M. Corgrams. *Am Stat*. Taylor & Francis; 2002;56:316–24.
292. Foster ZSL, Sharpton TJ, Grünwald NJ. Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. *PLoS Comput Biol*. 2017;13:e1005404.
293. Miquel S, Martín R, Rossi O, Bermúdez-Humarán LG, Chatel JM, Sokol H, et al. *Faecalibacterium prausnitzii* and human intestinal health. *Curr Opin Microbiol*. 2013;16:255–61.
294. Salguero MV, Al-Obaide MAI, Singh R, Siepmann T, Vasylyeva TL. Dysbiosis of Gram-negative gut microbiota and the associated serum lipopolysaccharide exacerbates inflammation in type 2 diabetic patients with chronic kidney disease. *Exp Ther Med*. Spandidos Publications; 2019;18:3461–9.
295. Gasparrini AJ, Wang B, Sun X, Kennedy EA, Hernandez-Leyva A, Ndao IM, et al. Persistent metagenomic signatures of early-life hospitalization and antibiotic treatment in the infant gut microbiota and resistome. *Nature Microbiology*. Nature Publishing Group; 2019;4:2285–97.
296. Leeson N, Hsueh P-R. Antimicrobial resistance in the 21st century. *Future Microbiol*. 2015;10:297–8.
297. Paukner S, Riedl R. Pleuromutilins: Potent Drugs for Resistant Bugs-Mode of Action and Resistance. *Cold Spring Harb Perspect Med* [Internet]. 2017;7. Available from: <http://dx.doi.org/10.1101/cshperspect.a027110>

298. Wright PM, Seiple IB, Myers AG. The evolving role of chemical synthesis in antibacterial drug discovery. *Angew Chem Int Ed Engl*. Wiley Online Library; 2014;53:8840–69.
299. Martínez-Cano DJ, Reyes-Prieto M, Martínez-Romero E, Partida-Martínez LP, Latorre A, Moya A, et al. Evolution of small prokaryotic genomes. *Front Microbiol*. 2014;5:742.
300. Woodford N, Turton JF, Livermore DM. Multiresistant Gram-negative bacteria: the role of high-risk clones in the dissemination of antibiotic resistance. *FEMS Microbiol Rev*. academic.oup.com; 2011;35:736–55.
301. Gao NL, Chen J, Wang T, Lercher MJ, Chen W-H. Prokaryotic Genome Expansion Is Facilitated by Phages and Plasmids but Impaired by CRISPR. *Front Microbiol*. 2019;10:2254.
302. Reyman M, van Houten MA, van Baarle D, Bosch AATM, Man WH, Chu MLJN, et al. Impact of delivery mode-associated gut microbiota dynamics on health in the first year of life. *Nat Commun*. nature.com; 2019;10:4997.
303. Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N, et al. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc Natl Acad Sci U S A*. National Acad Sciences; 2010;107:11971–5.
304. Rinninella E, Raoul P, Cintoni M, Franceschi F, Miggiano GAD, Gasbarrini A, et al. What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases. *Microorganisms* [Internet]. 2019;7. Available from: <http://dx.doi.org/10.3390/microorganisms7010014>
305. Machiels K, Joossens M, Sabino J, De Preter V, Arijis I, Eeckhaut V, et al. A decrease of the butyrate-producing species *Roseburia hominis* and *Faecalibacterium prausnitzii* defines dysbiosis in patients with ulcerative colitis. *Gut*. gut.bmj.com; 2014;63:1275–83.
306. Sokol H, Pigneur B, Watterlot L, Lakhdari O, Bermúdez-Humarán LG, Gratadoux J-J, et al. *Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc Natl Acad Sci U S A*. National Acad Sciences; 2008;105:16731–6.
307. Rossi O, van Berkel LA, Chain F, Tanweer Khan M, Taverne N, Sokol H, et al. *Faecalibacterium prausnitzii* A2-165 has a high capacity to induce IL-10 in human and murine dendritic cells and modulates T cell responses. *Sci Rep*. nature.com; 2016;6:18507.
308. Jakobsson HE, Abrahamsson TR, Jenmalm MC, Harris K, Quince C, Jernberg C, et al. Decreased gut microbiota diversity, delayed *Bacteroidetes* colonisation and reduced Th1 responses in infants delivered by caesarean section. *Gut*. 2014;63:559–66.
309. Lloyd CM, Hawrylowicz CM. Regulatory T cells in asthma. *Immunity*. Elsevier; 2009;31:438–49.
310. Vuillermin PJ, Ponsonby A-L, Saffery R, Tang ML, Ellis JA, Sly P, et al. Microbial exposure, interferon gamma gene demethylation in naïve T-cells, and the risk of allergic disease. *Allergy*. Wiley; 2009;64:348–53.
311. Zhuang L, Chen H, Zhang S, Zhuang J, Li Q, Feng Z. Intestinal Microbiota in Early Life

and Its Implications on Childhood Health. *Genomics Proteomics Bioinformatics*. Elsevier; 2019;17:13–25.

312. Keag OE, Norman JE, Stock SJ. Long-term risks and benefits associated with cesarean delivery for mother, baby, and subsequent pregnancies: Systematic review and meta-analysis. *PLoS Med.* journals.plos.org; 2018;15:e1002494.

313. Coulthurst SJ, Barnard AML, Salmond GPC. Regulation and biosynthesis of carbapenem antibiotics in bacteria. *Nat Rev Microbiol.* nature.com; 2005;3:295–306.

314. Pierson LS 3rd, Pierson EA. Metabolism and function of phenazines in bacteria: impacts on the behavior of bacteria in the environment and biotechnological processes. *Appl Microbiol Biotechnol.* Springer; 2010;86:1659–70.

315. Vangay P, Ward T, Gerber JS, Knights D. Antibiotics, pediatric dysbiosis, and disease. *Cell Host Microbe.* Elsevier; 2015;17:553–64.

316. Stokholm J, Schjørring S, Pedersen L, Bischoff AL, Følsgaard N, Carson CG, et al. Prevalence and predictors of antibiotic administration during pregnancy and birth. *PLoS One.* journals.plos.org; 2013;8:e82932.

317. Gómez-Gómez C, Blanco-Picazo P, Brown-Jaque M, Quirós P, Rodríguez-Rubio L, Cerdà-Cuellar M, et al. Infectious phage particles packaging antibiotic resistance genes found in meat products and chicken feces. *Sci Rep.* nature.com; 2019;9:13281.

318. Brown-Jaque M, Calero-Cáceres W, Muniesa M. Transfer of antibiotic-resistance genes via phage-related mobile elements. *Plasmid.* Elsevier; 2015;79:1–7.

319. de la Cruz F, Davies J. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.* Elsevier; 2000;8:128–33.

320. Pärnänen K, Karkman A, Hultman J, Lyra C, Bengtsson-Palme J, Larsson DGJ, et al. Maternal gut and breast milk microbiota affect infant gut antibiotic resistome and mobile genetic elements. *Nat Commun.* nature.com; 2018;9:3891.

321. Wang M, Xiong W, Liu P, Xie X, Zeng J, Sun Y, et al. Metagenomic Insights Into the Contribution of Phages to Antibiotic Resistance in Water Samples Related to Swine Feedlot Wastewater Treatment. *Front Microbiol.* 2018;9:2474.

322. Bearson BL, Allen HK, Brunelle BW, Lee IS, Casjens SR, Stanton TB. The agricultural antibiotic carbadox induces phage-mediated gene transfer in *Salmonella*. *Front Microbiol.* 2014;5:52.

323. Torres-Barceló C. The disparate effects of bacteriophages on antibiotic-resistant bacteria. *Emerg Microbes Infect.* 2018;7:168.

324. Haaber J, Leisner JJ, Cohn MT, Catalan-Moreno A, Nielsen JB, Westh H, et al. Bacterial viruses enable their host to acquire antibiotic resistance genes from neighbouring cells. *Nat Commun.* nature.com; 2016;7:13333.

325. Hoffman SJ, Caleo GM, Daulaire N, Elbe S, Matsoso P, Mossialos E, et al. Strategies for achieving global collective action on antimicrobial resistance. *Bull World Health Organ. SciELO Public Health*; 2015;93:867–76.

326. Davies J, Davies D. Origins and Evolution of Antibiotic Resistance [Internet]. *Microbiology and Molecular Biology Reviews*. 2010. p. 417–33. Available from: <http://dx.doi.org/10.1128/mubr.00016-10>
327. Carpenter KL, Breckler FD, Gray BW. Role of Mechanical Bowel Preparation and Perioperative Antibiotics in Pediatric Pull-Through Procedures [Internet]. *Journal of Surgical Research*. 2019. p. 222–7. Available from: <http://dx.doi.org/10.1016/j.jss.2019.03.051>
328. Hawn MT, Itani KM, Gray SH, Vick CC. Association of timely administration of prophylactic antibiotics for major surgical procedures and surgical site infection. *Journal of the American [Internet]. Elsevier*; 2008; Available from: https://www.sciencedirect.com/science/article/pii/S1072751507019734?casa_token=jA9Q0XS-hioAAAAA:3k9KLhpPpA-T_vc-4f-x26lBrBP2XyC2V7uD32X4qrmu57BNR-4awAmO0ESu23VU1vCtQAv7lg
329. Adamu B, Abdu A, Abba AA, Borodo MM, Tleyjeh IM. Antibiotic prophylaxis for preventing post solid organ transplant tuberculosis. *Cochrane Database Syst Rev*. 2014;CD008597.
330. He Y, Yuan Q, Mathieu J, Stadler L, Senehi N, Sun R, et al. Antibiotic resistance genes from livestock waste: occurrence, dissemination, and treatment. *npj Clean Water*. Nature Publishing Group; 2020;3:1–11.
331. Saust LT, Monrad RN, Hansen MP, Arpi M, Bjerrum L. Quality assessment of diagnosis and antibiotic treatment of infectious diseases in primary care: a systematic review of quality indicators. *Scand J Prim Health Care*. Taylor & Francis; 2016;34:258.
332. Schleiss MR. Principles of Antibacterial Therapy [Internet]. *Nelson Textbook of Pediatrics*. 2011. p. 903–903.e23. Available from: <http://dx.doi.org/10.1016/b978-1-4377-0755-7.00173-1>
333. Zackular JP, Baxter NT, Iverson KD, Sadler WD. The gut microbiome modulates colon tumorigenesis. *MBio [Internet]. Am Soc Microbiol*; 2013; Available from: <https://mbio.asm.org/content/4/6/e00692-13.short>
334. Candon S, Perez-Arroyo A, Marquet C, Valette F, Foray A-P, Pelletier B, et al. Antibiotics in Early Life Alter the Gut Microbiome and Increase Disease Incidence in a Spontaneous Mouse Model of Autoimmune Insulin-Dependent Diabetes. *PLoS One*. Public Library of Science; 2015;10:e0125448.
335. Korte SW, Dorfmeier RA, Franklin CL, Ericsson AC. Acute and long-term effects of antibiotics commonly used in laboratory animal medicine on the fecal microbiota. *Vet Res*. 2020;51:116.
336. Franklin CL, Ericsson AC. Microbiota and reproducibility of rodent models. *Lab Anim [Internet]. nature.com*; 2017; Available from: <https://www.nature.com/articles/labani.1222.pdf?origin=ppub>
337. Cassini A, Högberg LD, Plachouras D, Quattrocchi A, Hoxha A, Simonsen GS, et al. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis. *Lancet Infect Dis*. Elsevier; 2019;19:56–66.

338. Baker S, Thomson N, Weill F-X, Holt KE. Genomic insights into the emergence and spread of antimicrobial-resistant bacterial pathogens. *Science*. 2018;360:733–8.
339. Chen J, Quiles-Puchalt N, Chiang YN, Bacigalupe R, Fillol-Salom A, Chee MSJ, et al. Genome hypermobility by lateral transduction. *Science*. 2018;362:207–12.
340. Hernando-Amado S, Coque TM, Baquero F, Martínez JL. Defining and combating antibiotic resistance from One Health and Global Health perspectives. *Nat Microbiol*. 2019;4:1432–42.
341. Park H, Yeo S, Arellano K, Kim HR, Holzapfel W. Role of the Gut Microbiota in Health and Disease [Internet]. *Probiotics and Prebiotics in Animal Health and Food Safety*. 2018. p. 35–62. Available from: http://dx.doi.org/10.1007/978-3-319-71950-4_2
342. Kennedy EA, King KY, Baldrige MT. Mouse Microbiota Models: Comparing Germ-Free Mice and Antibiotics Treatment as Tools for Modifying Gut Bacteria. *Front Physiol*. 2018;9:1534.
343. Raymond F, Ouameur AA, Déraspe M, Iqbal N, Gingras H, Dridi B, et al. The initial state of the human gut microbiome determines its reshaping by antibiotics. *ISME J*. 2016;10:707–20.
344. Rodrigues RR, Greer RL, Dong X, DSouza KN, Gurung M, Wu JY, et al. Antibiotic-Induced Alterations in Gut Microbiota Are Associated with Changes in Glucose Metabolism in Healthy Mice. *Front Microbiol*. 2017;8:2306.
345. Croswell A, Amir E, Tegatz P, Barman M, Salzman NH. Prolonged impact of antibiotics on intestinal microbial ecology and susceptibility to enteric *Salmonella* infection. *Infect Immun*. 2009;77:2741–53.
346. Bratzler DW, Dellinger EP, Olsen KM, Perl TM, Auwaerter PG, Bolon MK, et al. Clinical practice guidelines for antimicrobial prophylaxis in surgery. *Surg Infect*. 2013;14:73–156.
347. Ericsson AC, Akter S, Hanson MM, Busi SB. Differential susceptibility to colorectal cancer due to naturally occurring gut microbiota. *Oncotarget* [Internet]. *ncbi.nlm.nih.gov*; 2015; Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4741795/>
348. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*. *PeerJ Inc.*; 2015;3:e1319.
349. De Nies L, Lopes S, Heintz-Buschart A, Laczny CC. PathoFact: A pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. *BioRxiv* [Internet]. *biorxiv.org*; 2020; Available from: <https://www.biorxiv.org/content/10.1101/2020.03.24.006148v1.abstract>
350. Pereira MB, Wallroth M, Kristiansson E, Axelson-Fisk M. HattCI: Fast and Accurate attC site Identification Using Hidden Markov Models. *J Comput Biol*. *Mary Ann Liebert, Inc., publishers*; 2016;23:891–902.
351. Prism G. Graphpad software. San Diego, CA, USA. 1994;
352. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: an

advanced analysis and visualization platform for 'omics data. *PeerJ*. 2015;3:e1319.

353. Johnson TA, Stedtfeld RD, Wang Q, Cole JR, Hashsham SA, Looft T, et al. Clusters of Antibiotic Resistance Genes Enriched Together Stay Together in Swine Agriculture. *MBio*. 2016;7:e02214–5.

354. Pal C, Bengtsson-Palme J, Kristiansson E, Larsson DGJ. Co-occurrence of resistance genes to antibiotics, biocides and metals reveals novel insights into their co-selection potential. *BMC Genomics*. 2015;16:964.

355. Cury J, Jové T, Touchon M, Néron B, Rocha EP. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res*. 2016;44:4539–50.

356. Dosani S. Penicillin Man: Alexander Fleming and the Antibiotic Revolution. *BMJ*. British Medical Journal Publishing Group; 2004;330:50.

357. Howell JD, Macfarlane G, Sheehan JC. Alexander Fleming: The Man and the Myth [Internet]. *Technology and Culture*. 1986. p. 309. Available from: <http://dx.doi.org/10.2307/3105159>

358. Roberts SC, Zembower TR. Global increases in antibiotic consumption: a concerning trend for WHO targets. *Lancet Infect Dis* [Internet]. 2020; Available from: [http://dx.doi.org/10.1016/S1473-3099\(20\)30456-4](http://dx.doi.org/10.1016/S1473-3099(20)30456-4)

359. Klein EY, Van Boeckel TP, Martinez EM, Pant S, Gandra S, Levin SA, et al. Global increase and geographic convergence in antibiotic consumption between 2000 and 2015. *Proc Natl Acad Sci U S A*. 2018;115:E3463–70.

360. Organization WH, Others. WHO global strategy for containment of antimicrobial resistance [Internet]. World Health Organization; 2001. Available from: https://apps.who.int/iris/bitstream/handle/10665/66860/WHO_CDS_CSR_DRS_2001.2.pdf

361. Culp EJ, Waglechner N, Wang W, Fiebig-Comyn AA, Hsu Y-P, Koteva K, et al. Evolution-guided discovery of antibiotics that inhibit peptidoglycan remodelling. *Nature*. 2020;578:582–7.

362. Leekha S, Terrell CL, Edson RS. General principles of antimicrobial therapy. *Mayo Clin Proc*. 2011;86:156–67.

363. Lamb R, Ozsvári B, Lisanti CL, Tanowitz HB, Howell A, Martinez-Outschoorn UE, et al. Antibiotics that target mitochondria effectively eradicate cancer stem cells, across multiple tumor types: treating cancer like an infectious disease. *Oncotarget*. Impact Journals, LLC; 2015;6:4569.

364. Wilcox MH, Gerding DN, Poxton al IR, Kelly C, Nathan R, Birch T, et al. MODIFY I and MODIFY II Investigators. Bezlotoxumab for prevention of recurrent *Clostridium difficile* infection. *N Engl J Med*. 2017;376:305–17.

365. Staley C, Kaiser T, Beura LK, Hamilton MJ, Weingarden AR, Bobr A, et al. Stable engraftment of human microbiota into mice with a single oral gavage following antibiotic conditioning. *Microbiome*. 2017;5:87.

366. Hintze KJ, Cox JE, Rompato G, Benninghoff AD, Ward RE, Broadbent J, et al. Broad

scope method for creating humanized animal models for animal health and disease research through antibiotic treatment and human fecal transfer. *Gut Microbes*. 2014;5:183–91.

367. Allen J, David M, Veerman JL. Systematic review of the cost-effectiveness of preoperative antibiotic prophylaxis in reducing surgical-site infection [Internet]. *BJS Open*. 2018. p. 81–98. Available from: <http://dx.doi.org/10.1002/bjs5.45>

368. Crader MF, Varacallo M. Preoperative Antibiotic Prophylaxis. StatPearls. Treasure Island (FL): StatPearls Publishing; 2020.

369. Hansen CHF, Krych L, Nielsen DS, Vogensen FK, Hansen LH, Sørensen SJ, et al. Early life treatment with vancomycin propagates *Akkermansia muciniphila* and reduces diabetes incidence in the NOD mouse. *Diabetologia*. 2012;55:2285–94.

370. Ray P, Pandey U, Aich P. Comparative analysis of beneficial effects of vancomycin treatment on Th1- and Th2-biased mice and the role of gut microbiota. *J Appl Microbiol* [Internet]. 2020; Available from: <http://dx.doi.org/10.1111/jam.14853>

371. Basolo A, Hohenadel M, Ang QY, Piaggi P, Heinitz S, Walter M, et al. Effects of underfeeding and oral vancomycin on gut microbiome and nutrient absorption in humans. *Nat Med*. 2020;26:589–98.

372. van Passel MWJ, Kant R, Zoetendal EG, Plugge CM, Derrien M, Malfatti SA, et al. The genome of *Akkermansia muciniphila*, a dedicated intestinal mucin degrader, and its use in exploring intestinal metagenomes. *PLoS One*. 2011;6:e16876.

373. Palreja A, Mikkelsen KH, Forslund SK, Kashani A, Allin KH, Nielsen T, et al. Recovery of gut microbiota of healthy adults following antibiotic exposure. *Nat Microbiol*. 2018;3:1255–65.

374. Guo X, Li S, Zhang J, Wu F, Li X, Wu D, et al. Genome sequencing of 39 *Akkermansia muciniphila* isolates reveals its population structure, genomic and functional diversity, and global distribution in mammalian gut microbiotas. *BMC Genomics*. 2017;18:800.

375. Madsen JS, Sørensen SJ, Burmølle M. Bacterial social interactions and the emergence of community-intrinsic properties. *Curr Opin Microbiol*. 2018;42:104–9.

376. Camilli A, Bassler BL. Bacterial small-molecule signaling pathways. *Science*. 2006;311:1113–6.

377. Meredith HR, Srimani JK, Lee AJ, Lopatkin AJ, You L. Collective antibiotic tolerance: mechanisms, dynamics and intervention. *Nat Chem Biol*. 2015;11:182–8.

378. Hibbing ME, Fuqua C, Parsek MR, Peterson SB. Bacterial competition: surviving and thriving in the microbial jungle. *Nat Rev Microbiol*. 2010;8:15–25.

379. Xu L, Surathu A, Raplee I, Chockalingam A, Stewart S, Walker L, et al. The effect of antibiotics on the gut microbiome: a metagenomics analysis of microbial shift and gut antibiotic resistance in antibiotic treated mice. *BMC Genomics*. 2020;21:263.

380. Gillings MR. Integrins: past, present, and future. *Microbiol Mol Biol Rev*. 2014;78:257–77.

381. Bonilla AR, Muniz KP. Antibiotic resistance: causes and risk factors, mechanisms and

alternatives. Nova Science Publishers; 2009.

382. Mahoney AR, Safaei MM, Wuest WM, Furst AL. The silent pandemic: Emergent antibiotic resistances following the global response to SARS-CoV-2. *iScience*. 2021;24:102304.
383. Calero-Cáceres W, Melgarejo A, Colomer-Lluch M, Stoll C, Lucena F, Jofre J, et al. Sludge as a potential important source of antibiotic resistance genes in both the bacterial and bacteriophage fractions. *Environ Sci Technol*. 2014;48:7602–11.
384. Chen B, Yang Y, Liang X, Yu K, Zhang T, Li X. Metagenomic profiles of antibiotic resistance genes (ARGs) between human impacted estuary and deep ocean sediments. *Environ Sci Technol*. 2013;47:12753–60.
385. Fouz N, Pangesti KNA, Yasir M, Al-Malki AL, Azhar EI, Hill-Cawthorne GA, et al. The Contribution of Wastewater to the Transmission of Antimicrobial Resistance in the Environment: Implications of Mass Gathering Settings. *Trop Med Infect Dis* [Internet]. 2020;5. Available from: <http://dx.doi.org/10.3390/tropicalmed5010033>
386. Singer AC, Shaw H, Rhodes V, Hart A. Review of Antimicrobial Resistance in the Environment and Its Relevance to Environmental Regulators. *Front Microbiol*. 2016;7:1728.
387. Herold M, d'Hérouël AF, May P, Delogu F, Wienecke-Baldacchino A, Tapp J, et al. Genome Sequencing of SARS-CoV-2 Allows Monitoring of Variants of Concern through Wastewater. *Water*. Multidisciplinary Digital Publishing Institute; 2021;13:3018.
388. Tennstedt T, Szczepanowski R, Braun S, Pühler A, Schlüter A. Occurrence of integron-associated resistance gene cassettes located on antibiotic resistance plasmids isolated from a wastewater treatment plant. *FEMS Microbiol Ecol*. 2003;45:239–52.
389. Lood R, Ertürk G, Mattiasson B. Revisiting Antibiotic Resistance Spreading in Wastewater Treatment Plants - Bacteriophages as a Much Neglected Potential Transmission Vehicle. *Front Microbiol*. 2017;8:2298.
390. Strange JES, Leekitcharoenphon P, Møller FD, Aarestrup FM. Metagenomics analysis of bacteriophages and antimicrobial resistance from global urban sewage. *Sci Rep. Nature Publishing Group*; 2021;11:1–11.
391. Li Q, Chang W, Zhang H, Hu D, Wang X. The Role of Plasmids in the Multiple Antibiotic Resistance Transfer in ESBLs-Producing *Escherichia coli* Isolated From Wastewater Treatment Plants. *Front Microbiol*. 2019;10:633.
392. Che Y, Xia Y, Liu L, Li A-D, Yang Y, Zhang T. Mobile antibiotic resistome in wastewater treatment plants revealed by Nanopore metagenomic sequencing. *Microbiome*. 2019;7:44.
393. Reza A, Sutton JM, Rahman KM. Effectiveness of Efflux Pump Inhibitors as Biofilm Disruptors and Resistance Breakers in Gram-Negative (ESKAPEE) Bacteria. *Antibiotics (Basel)* [Internet]. 2019;8. Available from: <http://dx.doi.org/10.3390/antibiotics8040229>
394. Herold M, Martínez Arbas S, Narayanasamy S, Sheik AR, Kleine-Borgmann LAK, Lebrun LA, et al. Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance. *Nat Commun*. 2020;11:5281.

395. Roume H, Heintz-Buschart A, Muller EEL, Wilmes P. Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample. *Methods Enzymol.* 2013;531:219–36.
396. de Nies L, Lopes S, Busi SB, Galata V, Heintz-Buschart A, Laczny CC, et al. PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. *Microbiome.* 2021;9:49.
397. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics.* academic.oup.com; 2014;30:923–30.
398. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25:2078–9.
399. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* 2019;20:257.
400. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol.* 2012;30:918–20.
401. Vaudel M, Burkhardt JM, Zahedi RP, Oveland E, Berven FS, Sickmann A, et al. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat Biotechnol.* 2015;33:22–4.
402. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* 2021;49:D480–9.
403. Barsnes H, Vaudel M. SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. *J Proteome Res.* 2018;17:2552–5.
404. Langella O, Valot B, Balliau T, Blein-Nicolas M, Bonhomme L, Zivy M. X!TandemPipeline: A Tool to Manage Sequence Redundancy for Protein Inference and Phosphosite Identification. *J Proteome Res.* 2017;16:494–503.
405. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun.* 2014;5:5277.
406. Eng JK, Jahan TA, Hoopmann MR. Comet: an open-source MS/MS sequence database search tool. *Proteomics.* 2013;13:22–4.
407. Swanson HK, Lysy M, Power M, Stasko AD, Johnson JD, Reist JD. A new probabilistic method for quantifying n-dimensional ecological niches and niche overlap. *Ecology.* Wiley; 2015;96:318–24.
408. Team RC, Others. R: A language and environment for statistical computing. Vienna, Austria; 2013; Available from: <https://cran.microsoft.com/snapshot/2014-09-08/web/packages/dplR/vignettes/xdate-dplR.pdf>
409. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol.* 2015;11:e1004226.

410. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal, Complex Systems*. 2006;1695:1–9.
411. Martínez Arbas S, Narayanasamy S, Herold M, Lebrun LA, Hoopmann MR, Li S, et al. Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics. *Nat Microbiol*. 2021;6:123–35.
412. Delgado MA, Vincent PA, Farías RN, Salomón RA. Role of *Escherichia coli* functions as a microcin J25 efflux pump. *J Bacteriol*. 2005;187:3465–70.
413. Calusinska M, Goux X, Fossépré M, Muller EEL, Wilmes P, Delfosse P. A year of monitoring 20 mesophilic full-scale bioreactors reveals the existence of stable but different core microbiomes in bio-waste and wastewater anaerobic digestion systems. *Biotechnol Biofuels*. 2018;11:196.
414. Tacconelli E, Carrara E, Savoldi A, Harbarth S, Mendelson M, Monnet DL, et al. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect Dis*. 2018;18:318–27.
415. Wee BA, Muloi DM, van Bunnik BAD. Quantifying the transmission of antimicrobial resistance at the human and livestock interface with genomics. *Clin Microbiol Infect*. 2020;26:1612–6.
416. Slizovskiy IB, Mukherjee K, Dean CJ, Boucher C, Noyes NR. Mobilization of Antibiotic Resistance: Are Current Approaches for Colocalizing Resistomes and Mobilomes Useful? *Front Microbiol*. 2020;11:1376.
417. Westhaus S, Weber F-A, Schiwy S, Linnemann V, Brinkmann M, Widera M, et al. Detection of SARS-CoV-2 in raw and treated wastewater in Germany - Suitability for COVID-19 surveillance and potential transmission risks. *Sci Total Environ*. Elsevier BV; 2021;751:141750.
418. Kwak Y-K, Colque P, Byfors S, Giske CG, Möllby R, Kühn I. Surveillance of antimicrobial resistance among *Escherichia coli* in wastewater in Stockholm during 1 year: does it reflect the resistance trends in the society? *Int J Antimicrob Agents*. 2015;45:25–32.
419. Reinthaler FF, Galler H, Feierl G, Haas D, Leitner E, Mascher F, et al. Resistance patterns of *Escherichia coli* isolated from sewage sludge in comparison with those isolated from human patients in 2000 and 2009. *J Water Health*. 2013;11:13–20.
420. Szczepanowski R, Linke B, Krahn I, Gartemann K-H, Gützkow T, Eichler W, et al. Detection of 140 clinically relevant antibiotic-resistance genes in the plasmid metagenome of wastewater treatment plant bacteria showing reduced susceptibility to selected antibiotics. *Microbiology*. 2009;155:2306–19.
421. Parsley LC, Consuegra EJ, Kakirde KS, Land AM, Harper WF Jr, Liles MR. Identification of diverse antimicrobial resistance determinants carried on bacterial, plasmid, or viral metagenomes from an activated sludge microbial assemblage. *Appl Environ Microbiol*. 2010;76:3753–7.
422. Hiller CX, Hübner U, Fajnorova S, Schwartz T, Drewes JE. Antibiotic microbial resistance (AMR) removal efficiencies by conventional and advanced wastewater treatment

processes: A review. *Sci Total Environ.* 2019;685:596–608.

423. Proia L, Anzil A, Borrego C, Farrè M, Llorca M, Sanchis J, et al. Occurrence and persistence of carbapenemases genes in hospital and wastewater treatment plants and propagation in the receiving river. *J Hazard Mater.* 2018;358:33–43.

424. Rodriguez-Mozaz S, Chamorro S, Marti E, Huerta B, Gros M, Sànchez-Melsió A, et al. Occurrence of antibiotics and antibiotic resistance genes in hospital and urban wastewaters and their impact on the receiving river. *Water Res.* 2015;69:234–42.

425. Aarestrup FM, Woolhouse MEJ. Using sewage for surveillance of antimicrobial resistance. *Science.* 2020;367:630–2.

426. Pärnänen KMM, Narciso-da-Rocha C, Kneis D, Berendonk TU, Cacace D, Do TT, et al. Antibiotic resistance in European wastewater treatment plants mirrors the pattern of clinical antibiotic resistance prevalence. *Sci Adv.* 2019;5:eaau9124.

427. Su J-Q, An X-L, Li B, Chen Q-L, Gillings MR, Chen H, et al. Metagenomics of urban sewage identifies an extensively shared antibiotic resistome in China. *Microbiome.* 2017;5:84.

428. Frost I, Smith WPJ, Mitri S, Millan AS, Davit Y, Osborne JM, et al. Cooperation, competition and antibiotic resistance in bacterial colonies. *ISME J.* 2018;12:1582–93.

429. Turolla A, Cattaneo M, Marazzi F, Mezzanotte V, Antonelli M. Antibiotic resistant bacteria in urban sewage: Role of full-scale wastewater treatment plants on environmental spreading. *Chemosphere.* 2018;191:761–9.

430. Newton RJ, McClary JS. The flux and impact of wastewater infrastructure microorganisms on human and ecosystem health. *Curr Opin Biotechnol.* 2019;57:145–50.

431. Dubnau D, Grandi G, Grandi R, Gryczan TJ, Hahn J, Kozloff Y, et al. Regulation of Plasmid Specified MLS-Resistance in *Bacillus subtilis* by Conformational Alteration of RNA Structure. In: Levy SB, Clowes RC, Koenig EL, editors. *Molecular Biology, Pathogenicity, and Ecology of Bacterial Plasmids.* Boston, MA: Springer US; 1981. p. 157–67.

432. Galimand M, Courvalin P, Lambert T. Plasmid-mediated high-level resistance to aminoglycosides in Enterobacteriaceae due to 16S rRNA methylation. *Antimicrob Agents Chemother.* 2003;47:2565–71.

433. Han X, Du X-D, Southey L, Bulach DM, Seemann T, Yan X-X, et al. Functional Analysis of a Bacitracin Resistance Determinant Located on ICECp1, a Novel Tn916-Like Element from a Conjugative Plasmid in *Clostridium perfringens*. *Antimicrob Agents Chemother.* American Society for Microbiology Journals; 2015;59:6855–65.

434. Razavi M, Marathe NP, Gillings MR, Flach C-F, Kristiansson E, Joakim Larsson DG. Discovery of the fourth mobile sulfonamide resistance gene. *Microbiome.* 2017;5:160.

435. Vrancianu CO, Popa LI, Bleotu C, Chifiriuc MC. Targeting Plasmids to Limit Acquisition and Transmission of Antimicrobial Resistance. *Front Microbiol.* 2020;11:761.

436. Bouanchaud DH, Chabbert YA. The problems of drug-resistant pathogenic bacteria. Practical effectiveness of agents curing R factors and plasmids. *Ann N Y Acad Sci.* 1971;182:305–11.

437. Buckner MMC, Ciusa ML, Piddock LJV. Strategies to combat antimicrobial resistance: anti-plasmid and plasmid curing. *FEMS Microbiol Rev.* 2018;42:781–804.
438. Naimi S, Zirah S, Hammami R, Fernandez B, Rebuffat S, Fliss I. Fate and Biological Activity of the Antimicrobial Lasso Peptide Microcin J25 Under Gastrointestinal Tract Conditions. *Front Microbiol.* 2018;9:1764.
439. Ben Said L, Emond-Rheault J-G, Soltani S, Telhig S, Zirah S, Rebuffat S, et al. Phenomic and genomic approaches to studying the inhibition of multiresistant *Salmonella enterica* by microcin J25. *Environ Microbiol.* Wiley Online Library; 2020;22:2907–20.
440. Cameron DE, Collins JJ. Tunable protein degradation in bacteria. *Nat Biotechnol.* 2014;32:1276–81.
441. Balasegaram M. Learning from COVID-19 to Tackle Antibiotic Resistance. *ACS Infect Dis.* 2021;7:693–4.
442. Tyc O, Song C, Dickschat JS, Vos M, Garbeva P. The Ecological Role of Volatile and Soluble Secondary Metabolites Produced by Soil Bacteria. *Trends Microbiol.* 2017;25:280–92.
443. Chen R, Wong HL, Kindler GS, MacLeod FI, Benaud N, Ferrari BC, et al. Discovery of an Abundance of Biosynthetic Gene Clusters in Shark Bay Microbial Mats. *Front Microbiol.* 2020;11:1950.
444. Demain AL, Fang A. The Natural Functions of Secondary Metabolites. In: Fiechter A, editor. *History of Modern Biotechnology I.* Berlin, Heidelberg: Springer Berlin Heidelberg; 2000. p. 1–39.
445. Newman DJ, Cragg GM. Natural Products as Sources of New Drugs from 1981 to 2014. *J Nat Prod.* 2016;79:629–61.
446. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, et al. Minimum Information about a Biosynthetic Gene cluster. *Nat Chem Biol.* 2015;11:625–31.
447. Martinet L, Naômé A, Deflandre B, Maciejewska M, Tellatin D, Tenconi E, et al. A Single Biosynthetic Gene Cluster Is Responsible for the Production of Bagremycin Antibiotics and Ferroverdin Iron Chelators. *MBio* [Internet]. 2019;10. Available from: <http://dx.doi.org/10.1128/mBio.01230-19>
448. Martínez-Núñez MA, López VEL y. Nonribosomal peptides synthetases and their applications in industry. *Sustainable Chemical Processes.* SpringerOpen; 2016;4:1–8.
449. Ridley CP, Lee HY, Khosla C. Evolution of polyketide synthases in bacteria. *Proc Natl Acad Sci U S A.* 2008;105:4595–600.
450. Cundliffe E, Bate N, Butler A, Fish S, Gandeche A, Merson-Davies L. The tylosin-biosynthetic genes of *Streptomyces fradiae*. *Antonie Van Leeuwenhoek.* 2001;79:229–34.
451. Kwun MJ, Hong H-J. Genome Sequence of *Streptomyces toyocaensis* NRRL 15009, Producer of the Glycopeptide Antibiotic A47934. *Genome Announc* [Internet]. 2014;2. Available from: <http://dx.doi.org/10.1128/genomeA.00749-14>

452. Busi SB, Bourquin M, Fodelianakis S, Michoud G, Kohler TJ, Peter H, et al. Genomic and metabolic adaptations of biofilms to ecological windows of opportunities in glacier-fed streams [Internet]. *bioRxiv*. 2021 [cited 2021 Nov 4]. p. 2021.10.07.463499. Available from: <https://www.biorxiv.org/content/10.1101/2021.10.07.463499v1>
453. Battin TJ, Besemer K, Bengtsson MM, Romani AM, Packmann AI. The ecology and biogeochemistry of stream biofilms. *Nat Rev Microbiol*. 2016;14:251–63.
454. Battin TJ, Wille A, Sattler B, Psenner R. Phylogenetic and functional heterogeneity of sediment biofilms along environmental gradients in a glacial stream. *Appl Environ Microbiol*. 2001;67:799–807.
455. Gaynes R. The Discovery of Penicillin—New Insights After More Than 75 Years of Clinical Use. *Emerg Infect Dis*. Centers for Disease Control and Prevention; 2017;23:849.
456. Netzker T, Flak M, Krespach MK, Stroe MC, Weber J, Schroeckh V, et al. Microbial interactions trigger the production of antibiotics. *Curr Opin Microbiol*. 2018;45:117–23.
457. Busi SB, Pramateftaki P, Brandani J, Fodelianakis S, Peter H, Halder R, et al. Optimised biomolecular extraction for metagenomic analysis of microbial biofilms from high-mountain streams. *PeerJ*. 2020;8:e9973.
458. Eddy SR. Accelerated Profile HMM Searches. *PLoS Comput Biol*. 2011;7:e1002195.
459. Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res*. 2021;49:W29–35.
460. Hannigan GD, Prihoda D, Palicka A, Soukup J, Klempir O, Rampula L, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res*. 2019;47:e110.
461. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the tidyverse. *J Open Source Softw*. The Open Journal; 2019;4:1686.
462. Brunson J. ggalluvial: Layered Grammar for Alluvial Plots. *J Open Source Softw*. The Open Journal; 2020;5:2017.
463. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*. 2016;32:2847–9.
464. Krause KM, Serio AW, Kane TR, Connolly LE. Aminoglycosides: An Overview. *Cold Spring Harb Perspect Med* [Internet]. 2016;6. Available from: <http://dx.doi.org/10.1101/cshperspect.a027029>
465. Tahlan K, Jensen SE. Origins of the β -lactam rings in natural products. *J Antibiot* . 2013;66:401–10.
466. Borges-Walmsley MI, McKeegan KS, Walmsley AR. Structure and function of efflux pumps that confer resistance to drugs. *Biochem J*. 2003;376:313–38.
467. Tortorella E, Tedesco P, Palma Esposito F, January GG, Fani R, Jaspars M, et al. Antibiotics from Deep-Sea Microorganisms: Current Discoveries and Perspectives. *Mar Drugs*

[Internet]. 2018;16. Available from: <http://dx.doi.org/10.3390/md16100355>

468. Yuan K, Yu K, Yang R, Zhang Q, Yang Y, Chen E, et al. Metagenomic characterization of antibiotic resistance genes in Antarctic soils. *Ecotoxicol Environ Saf*. 2019;176:300–8.

469. Centurion VB, Delforno TP, Lacerda-Júnior GV, Duarte AWF, Silva LJ, Bellini GB, et al. Unveiling resistome profiles in the sediments of an Antarctic volcanic island. *Environ Pollut*. 2019;255:113240.

470. Van Goethem MW, Pierneef R, Bezuidt OKI, Van De Peer Y, Cowan DA, Makhalanyane TP. A reservoir of “historical” antibiotic resistance genes in remote pristine Antarctic soils. *Microbiome*. 2018;6:40.

471. Brown JR, Zhang J, Hodgson JE. A bacterial antibiotic resistance gene with eukaryotic origins. *Curr Biol*. 1998;8:R365–7.

472. Fairlamb AH, Gow NAR, Matthews KR, Waters AP. Drug resistance in eukaryotic microorganisms. *Nat Microbiol*. 2016;1:16092.

473. Silva A, Silva SA, Carpena M, Garcia-Oliveira P, Gullón P, Barroso MF, et al. Macroalgae as a Source of Valuable Antimicrobial Compounds: Extraction and Applications. *Antibiotics (Basel)* [Internet]. 2020;9. Available from: <http://dx.doi.org/10.3390/antibiotics9100642>

474. Martins RM, Nedel F, Guimarães VBS, da Silva AF, Colepicolo P, de Pereira CMP, et al. Macroalgae Extracts From Antarctica Have Antimicrobial and Anticancer Potential. *Front Microbiol*. frontiersin.org; 2018;9:412.

475. Karkman A, Pärnänen K, Larsson DGJ. Fecal pollution can explain antibiotic resistance gene abundances in anthropogenically impacted environments. *Nat Commun*. 2019;10:80.

476. Antelo V, Giménez M, Azziz G, Valdespino-Castillo P, Falcón LI, Ruberto LAM, et al. Metagenomic strategies identify diverse integron-integrase and antibiotic resistance genes in the Antarctic environment. *Microbiologyopen* [Internet]. Wiley; 2021;10. Available from: <https://onlinelibrary.wiley.com/doi/10.1002/mbo3.1219>

477. Hernández F, Calisto-Ulloa N, Gómez-Fuentes C, Gómez M, Ferrer J, González-Rocha G, et al. Occurrence of antibiotics and bacterial resistance in wastewater and sea water from the Antarctic. *J Hazard Mater*. 2019;363:447–56.

478. Waschulin V, Borsetto C, James R, Newsham KK, Donadio S, Corre C, et al. Biosynthetic potential of uncultured Antarctic soil bacteria revealed through long-read metagenomic sequencing. *ISME J* [Internet]. 2021; Available from: <http://dx.doi.org/10.1038/s41396-021-01052-3>

479. Liao L, Su S, Zhao B, Fan C, Zhang J, Li H, et al. Biosynthetic Potential of a Novel Antarctic Actinobacterium *Marisediminicola antarctica* ZS314T Revealed by Genomic Data Mining and Pigment Characterization. *Mar Drugs* [Internet]. 2019;17. Available from: <http://dx.doi.org/10.3390/md17070388>

480. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. *Nat Microbiol*. 2016;1:16048.

481. Tian R, Ning D, He Z, Zhang P, Spencer SJ, Gao S, et al. Small and mighty: adaptation of superphylum Patescibacteria to groundwater environment drives their genome simplicity. *Microbiome*. 2020;8:51.
482. Vigneron A, Cruaud P, Langlois V, Lovejoy C, Culley AI, Vincent WF. Ultra-small and abundant: Candidate phyla radiation bacteria are potential catalysts of carbon transformation in a thermokarst lake ecosystem. *Limnol Oceanogr Lett*. Wiley; 2020;5:212–20.
483. Maatouk M, Ibrahim A, Rolain J-M, Merhej V, Bittar F. Small and equipped: the rich repertoire of antibiotic resistance genes in Candidate Phyla Radiation genomes [Internet]. *bioRxiv*. 2021 [cited 2021 Sep 21]. p. 2021.07.02.450847. Available from: <https://www.biorxiv.org/content/10.1101/2021.07.02.450847v1.full>
484. Bottery MJ, Pitchford JW, Friman V-P. Ecology and evolution of antimicrobial resistance in bacterial communities. *ISME J*. 2021;15:939–48.
485. Bottery MJ, Passaris I, Dytham C, Wood AJ, van der Woude MW. Spatial Organization of Expanding Bacterial Colonies Is Affected by Contact-Dependent Growth Inhibition. *Curr Biol*. 2019;29:3622–34.e5.
486. Schluter J, Nadell CD, Bassler BL, Foster KR. Adhesion as a weapon in microbial competition. *ISME J*. 2015;9:139–49.
487. Stubbendieck RM, Straight PD. Multifaceted Interfaces of Bacterial Competition. *J Bacteriol*. 2016;198:2145–55.
488. Estrela S, Brown SP. Community interactions and spatial structure shape selection on antibiotic resistant lineages. *PLoS Comput Biol*. 2018;14:e1006179.
489. Zhou Z-C, Feng W-Q, Han Y, Zheng J, Chen T, Wei Y-Y, et al. Prevalence and transmission of antibiotic resistance and microbiota between humans and water environments. *Environ Int*. 2018;121:1155–61.
490. Allen HK, Moe LA, Rodbumrer J, Gaarder A, Handelsman J. Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil. *ISME J*. 2009;3:243–51.
491. Fonseca EL, Andrade BGN, Vicente ACP. The Resistome of Low-Impacted Marine Environments Is Composed by Distant Metallo- β -Lactamases Homologs. *Front Microbiol*. 2018;9:677.
492. Piotrowska M, Kowalska S, Popowska M. Diversity of β -lactam resistance genes in gram-negative rods isolated from a municipal wastewater treatment plant. *Ann Microbiol. BioMed Central*; 2019;69:591–601.
493. Majeed HJ, Riquelme MV, Davis BC, Gupta S, Angeles L, Aga DS, et al. Evaluation of Metagenomic-Enabled Antibiotic Resistance Surveillance at a Conventional Wastewater Treatment Plant. *Front Microbiol*. 2021;12:657954.
494. Botts RT, Apffel BA, Walters CJ, Davidson KE, Echols RS, Geiger MR, et al. Characterization of Four Multidrug Resistance Plasmids Captured from the Sediments of an Urban Coastal Wetland. *Front Microbiol*. 2017;8:1922.
495. Wang J, Stephan R, Power K, Yan Q, Hächler H, Fanning S. Nucleotide sequences of

- 16 transmissible plasmids identified in nine multidrug-resistant *Escherichia coli* isolates expressing an ESBL phenotype isolated from food-producing animals and healthy humans. *J Antimicrob Chemother.* 2014;69:2658–68.
496. Mathers AJ, Peirano G, Pitout JDD. The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant Enterobacteriaceae. *Clin Microbiol Rev.* 2015;28:565–91.
497. Vintov J, Aarestrup FM, Zinn CE, Olsen JE. Association between phage types and antimicrobial resistance among bovine *Staphylococcus aureus* from 10 countries. *Vet Microbiol.* Elsevier; 2003;95:133–47.
498. Perry JA, Wright GD. The antibiotic resistance “mobilome”: searching for the link between environment and clinic. *Front Microbiol.* 2013;4:138.
499. Hughes VM, Datta N. Conjugative plasmids in bacteria of the “pre-antibiotic” era. *Nature.* 1983;302:725–6.
500. Rao S, Linke L, Doster E, Hyatt D, Burgess BA, Magnuson R, et al. Genomic diversity of class I integrons from antimicrobial resistant strains of *Salmonella Typhimurium* isolated from livestock, poultry and humans. *PLoS One.* 2020;15:e0243477.
501. Amos GCA, Ploumakis S, Zhang L, Hawkey PM, Gaze WH, Wellington EMH. The widespread dissemination of integrons throughout bacterial communities in a riverine system. *ISME J.* 2018;12:681–91.
502. Zhang A-N, Gaston JM, Dai CL, Zhao S, Poyet M, Groussin M, et al. An omics-based framework for assessing the health risk of antimicrobial resistance genes. *Nat Commun.* 2021;12:4765.
503. Fernandes MR, Moura Q, Sartori L, Silva KC, Cunha MP, Esposito F, et al. Silent dissemination of colistin-resistant *Escherichia coli* in South America could contribute to the global spread of the *mcr-1* gene. *Euro Surveill [Internet].* 2016;21. Available from: <http://dx.doi.org/10.2807/1560-7917.ES.2016.21.17.30214>
504. Mather AE, Reid SWJ, Maskell DJ, Parkhill J, Fookes MC, Harris SR, et al. (2013) Distinguishable epidemics of multidrug-resistant *Salmonella Typhimurium* DT104 in different hosts. *Science* 341:1514–1517. <https://doi.org/10.1126/science.1240578>
505. Muloi D, Ward MJ, Pedersen AB, et al (2018) Are Food Animals Responsible for Transfer of Antimicrobial-Resistant *Escherichia coli* or Their Resistance Determinants to Human Populations? A Systematic Review. *Foodborne Pathog Dis* 15:467–474. <https://doi.org/10.1089/fpd.2017.2411>
506. Al-Shayeb B, Schoelmerich MC, West-Roberts J, et al (2021) Borgen are giant extrachromosomal elements with the potential to augment methane oxidation. *bioRxiv* 2021.07.10.451761

Appendix A.1
Reservoirs of antimicrobial resistance
in the context of One Health

Reservoirs of antimicrobial resistance in the context of One Health

Laura de Nies¹, Susheel Bhanu Busi¹ and Paul Wilmes^{1*}

¹Systems Ecology research group, Luxembourg Centre for Systems Biomedicine, Esch-sur-Alzette, Luxembourg

* Corresponding author: paul.wilmes@uni.lu

Abstract

The emergence and spread of antimicrobial resistance (AMR), as well as resistant bacteria, is a global threat to public health due to the inability to comprehensively treat bacterial infections. Through horizontal gene transfer, bacteria can acquire antimicrobial resistance genes (ARGs) via conjugation using plasmids or through transduction employing phages. Furthermore, although AMR primarily derives from the use and misuse of antibiotics in humans and animals, the environment also acts as an important conduit for the emergence and spread of AMR. Thereby, AMR has the potential to rapidly become a pandemic whereby it is no longer constrained by geographical or human-animal borders. In order to understand the dissemination of antibiotic resistance, it is therefore necessary to map the resistome (the collection of ARGs in a given environment or organism) within different microbial reservoirs, and to unravel the extent by which the transfer of genes to and from *de facto* and putative human pathogens occurs. The ability to sequence DNA indiscriminately from environmental, animal and clinical samples represents a paradigm shift in our ability to investigate, identify and monitor AMR. More specifically, metagenomics allows the investigation of AMR, including the prevalence and spread of ARGs, within distinct microbial populations without the need to isolate and grow the concerned microorganisms in pure culture. In general terms, there exists an urgent need to understand the exchange of ARGs within and between biomes in the context of microbiome-borne pathogenic potential as it relates to human infectious and chronic diseases.

Keywords: Antimicrobial resistance, Mobile Genetic Elements, One Health, Microbial Reservoirs, Metagenomics

Introduction

Throughout history, bacterial infections have been a major cause of human disease and mortality. The discovery, subsequent development, and medical use of antibiotics brought an end to this pre-antibiotic era by providing effective treatment against bacterial infections. However, the use of antibiotics has gone hand-in-hand with the emergence and spread of antimicrobial resistance (AMR). Although antibiotic resistance in itself is a prehistoric phenomenon [1], the over- and mis-use of antibiotics has led to a global and immense increase in AMR over the past decades. As a result, many bacteria have now acquired resistance against multiple antibiotics which has led to the emergence of multi-resistant microorganisms, i.e. “superbugs” [2]. This phenomenon, for instance, has led to an overgrowth of pathobionts, encoding antimicrobial resistance genes (ARGs), causing alterations to the microbiome both in chronic diseases as well as in infections [3, 4]. Consequently, this threatens human health through the spread of multidrug-resistant bacteria with an estimated number of deaths, although challenged by some, which may exceed ten million annually by 2050 [5, 6].

Mechanisms of antimicrobial resistance

On the one hand, antimicrobial agents for fighting bacterial infections can be characterized depending on the mechanisms of their activity, i.e. agents that i) inhibit cell wall synthesis, ii) depolarize the cell membrane, iii) inhibit protein synthesis, iv) inhibit nucleic acid synthesis, or v) inhibit metabolic pathways [7]. On the other hand, various counteractive mechanisms have evolved to confer resistance. These can be characterized into categories such as those I) limiting the uptake of and exposure to antibiotics, II) modifying antibiotic targets through for example mutations, III) directly inactivating antibiotic molecules, or IV) ensuring their immediate export through active efflux pumps [7]. In this context, limitations to the uptake of antibiotics are mostly classified as intrinsic resistance. Acquired resistance in turn mostly utilizes the modification of antibiotic targets, while the inactivation or efflux of antibiotics are both intrinsic and acquired resistance mechanisms [7].

Bacteria have a natural ability which limits the uptake of antimicrobial agents. Specifically, in Gram negative bacteria the structure and functions of the LPS provide an immediate barrier to antibiotics, thereby conferring an innate resistance [8]. Gram positive bacteria, on the other hand, lack LPS and resort to mechanisms such as enzymatic degradation of antibiotics or reducing the affinity and susceptibility of antibiotic target sites [9]. Additionally, other mechanisms to limit uptake of antibiotics may involve a decrease in the number of porin channels or mutations in the corresponding genes as well as the formation of protective biofilms [10, 11].

Specific examples of antimicrobial target modifications include alterations in the structure of antibiotic binding proteins or mutations therein to prevent antibiotics from binding to those proteins[10, 11]. Additionally, modifications of DNA gyrase and topoisomerase IV interfere with the antibiotics targeting the nucleic acid synthesis machinery [12], while further mutations in enzymes generate resistance to antibiotics inhibiting metabolic pathways. Alternatively, upregulation in the expression of these enzymes confers resistance through competitive inhibition [13]. Besides these mechanisms, inactivation of the antimicrobial drugs itself can occur, conferring resistance either through actual degradation or through the transfer of a chemical group to the drug. Lastly, bacteria possess various types of efflux pumps, such as the ABC, MATE, SMR, MFS and RND transporter families, which enable resistance via efflux of antimicrobial drugs [8].

With respect to these mechanisms, bacterial resistance can be classified as either natural or acquired resistance [14]. Natural resistance can be further subdivided into either ‘intrinsic’, which is constantly expressed in a bacterial species, or ‘induced’, in which resistance genes are only expressed upon exposure to antibiotics [12]. Acquired resistance can be defined as the acquisition of resistance-conferring genetic material through horizontal gene transfer (HGT), e.g. conjugation or transduction and alternatively via mutations in the chromosomal DNA after antibiotic exposure [15]. In most cases, ARGs are associated with conjugation events which are the most likely mechanisms for the dissemination of AMR compared to transduction [16]. Interestingly, the rate of transfer of ARGs via the individual mechanisms is a complex process involving several factors, not limited to the mode, species of interest, bacterial environment (*in vitro* or *in vivo*), and also the antibiotics [17]. Despite previous reports suggesting low rates of ARG transfer via conjugation [18], Leclerc *et al.* [17] reported that an estimated gene transfer rate cannot be generalized across all species and antibiotics due to the multitude of factors highlighted above.

Dissemination of antimicrobial resistance through horizontal gene transfer

Horizontal gene transfer (HGT) is key to the evolution and adaptation of bacteria, allowing for the rapid gain of beneficial traits including ARGs. [19]. Employing HGT, bacteria can acquire ARGs through either conjugation or transduction via mobile genetic elements (MGEs). In conjugation, plasmids carrying one or more resistance genes are transferred between microorganisms, while in transduction, bacteriophages encoding ARGs infect bacteria thereby transferring resistance [20]. The collective MGEs within a given microbiome in this context are defined as the ‘mobilome’.

With respect to HGT or ARGs, plasmids represent an optimal vehicle. Plasmids are composed of either circular or linear DNA distinct from bacterial chromosomal DNA, capable of autonomous replication [21]. Besides encoding for resistance to most, if not all, major classes of antibiotics, multiple genes conferring resistance to different antibiotic categories can be found on the same plasmid. This is especially evident in the case of multidrug-resistant *Klebsiella pneumoniae*, against which antibiotic combination therapies are ineffective [22][23]. Furthermore, plasmids encoding ARGs are not only found within pathogenic bacteria but can also be detected in commensals [24]. Generally speaking, the predisposition of a HGT event has been deduced to depend on ecological and phylogenetic factors [25]. However, as described by Porse *et al.* [25] and others [26], in addition to the phylogenetic relatedness of the donor and recipient species, the AMR mechanisms themselves also act as crucial determinants of gene functionality and fitness cost. The functional compatibility of an ARG in a new host is dependent on the interaction with the host physiology and metabolism. Consequently, resistance mechanisms, i.e. drug-modifying enzymes, with limited cellular interactions are more likely to be functionally compatible and integrate easily into host physiology [25]. These observations suggest that depending on the ARG and the plasmid, they can be shared between closely and distantly related taxonomic clades, thereby contributing to widespread and rapid propagation of AMR [25, 27]. Alongside plasmids, integrons, often overlooked, can also play a significant role in AMR dissemination and prevalence [28]. Integrons, widely distributed and carried by plasmids, can acquire, exchange and express genes embedded within gene cassettes [29] further promoting their spread within and between microbial communities [30]. Generally, two distinct groups of integrons have been described, namely chromosomal and mobile integrons. Chromosomal integrons are encoded by many bacterial species and are also referred to as “super-integrons” due to their large size and ability to carry up to 2000 gene cassettes [31]. Mobile integrons, on the other hand, are located on MGEs such as plasmids or phages and have been associated with AMR and the dissemination of resistance among bacterial populations [28]. Collectively, integrons are efficient tools for bacterial adaptation and play a significant role in the spread of AMR in conjunction with plasmids.

Besides plasmids, (bacterio-)phages contribute to the horizontal gene transfer of ARGs via transduction. Transducing phages mediating AMR can be either virulent or temperate [32]. Upon infection, temperate phages integrate their DNA into the host chromosome in which the prophage subsequently becomes dormant. When induced by stress factors such as DNA damage [33] and/or environmental cues [34], the phage will be excised from the chromosome, inducing phage particle formation and lysis of the host cell [35]. In contrast, lytic phages immediately induce the formation of phage particles resulting in lysis of the host cell [36]. Additionally, transduction can be further separated into two types:

generalized or specialized. In generalized transduction, the genetic material is transferred to another bacterial cell where it is further integrated through homologous recombination. Specialized transduction, on the other hand, results in the packaging of bacteria DNA into phages at a higher frequency compared to generalized transduction. This lateral transfer of ARGs through phage-mediated transduction could be an important contributing factor in the global spread of AMR [37].

Methods for detecting antimicrobial resistance

Traditionally, culture-based methods, such as antimicrobial susceptibility testing (AST), have been, and still are, used in clinical settings to investigate AMR and resistant bacteria [38]. For phenotypic testing, bacterial isolates are cultured from samples using either non-selective or selective growth media. Subsequently, the susceptibility of the isolates to antibiotics is tested to identify AMR. These solid media techniques use Kirby-Bauer disc diffusion or gradient diffusion strips to measure the zone of inhibition, and thereby provide a proxy for the level of resistance [38, 39]. Although these methods provide crucial information regarding AMR, they are only suitable for bacteria which are readily culturable using standard cultivation methods. However, microbial communities such as those inhabiting the human gut are composed of significant proportions of, at present, difficult to culture or outright unculturable taxa. In this context, the ability to sequence DNA from samples (clinical and environmental) using high-throughput sequencing methodologies have improved our ability to investigate and identify AMR. Sequencing-based metagenomics, which involves the study of the total genetic material (e.g. DNA or RNA) recoverable directly from samples, allows for the genomic analysis of all organisms within a microbial ecosystem without previous identification [40]. This enables the investigation of the resistome, i.e. AMR, including the mechanisms and spread of ARGs, without the immediate need to isolate microorganisms.

Different bioinformatic workflows have been developed to investigate the presence of AMR and MGEs within metagenomes. These include both read- and *de novo* assembly-based methods which have been extensively discussed by Boolchandani *et al.* [38]. While read-based methods allow for identification of low-abundance AMR genes [38], *de novo* assembly strategies enable the genetic contextualization of AMR surveillance, such as their presence on MGEs [41]. Some of the AMR prediction tools including DeepARG [42], RGI [43], Resfinder [44], ARG-ANNOT [45], and NCBI-AMRFinder [46] can be used for ARG identification, albeit through use of their associated databases. For example, while DeepARG, RGI and NCBI-AMRFinder use the recently updated CARD database [43], other tools provide custom versions leading to discrepancies in identified ARGs. Nonetheless, none of the above tools provide information with

respect to contextualization of ARGs on MGEs which represent critical elements for AMR transmission. Alternatively, many tools have been developed for the independent prediction of MGEs alone, such as plasflow [47], MOB-suite [48] and gplas [49] for plasmid identification. For the prediction of phages in general, the following tools exist: DeepVirFinder [50], VirSorter [51], MARVEL [52] and PPR-Meta [53]. Each of these tools are specialised to allow identification of a single or limited set of MGEs. While some tools like DeepVirFinder are based on machine-learning methodologies, others are restricted to databases populated with previously identified MGEs. The former allows for discovery of putatively novel MGEs, while the latter methods allow for precision and confidence in the identified MGEs. In a One Health context, bridging together human, animal and environmental health [54], it is crucial to study both the prevalence and spread of AMR simultaneously. Such methods to systematically assess AMR within and between biomes have long remained elusive [55]. However, to precisely address this gap in methodologies, PathoFact [56] which genomically contextualizes ARGs, including their localization on MGEs, is applicable. Meanwhile, tools such as MOCAT2 [57] and HUMANn3 [58] also enable AMR gene identification, however, do not provide any information with respect to MGE contextualization. In a One Health setting, by combining effective study designs with computational analyses methods, it is thereby now possible to trace the origins and dissemination of AMR from one reservoir to another using metagenomic sequencing coupled to *de novo* reconstruction of genomic fragments.

Microbial reservoirs of antimicrobial resistance

Natural microbial communities, or microbiomes, represent multi-species assemblages which interact in a contiguous environment [59]. Current evidence suggests that the structure of human and animal microbiomes are shaped by several factors, including exposure to microorganisms through contacts with exogenous sources (e.g. parents, animals, environment), specific host-microbe interactions linked in particular to host immune responses, and the outcome of competitive, cooperative and/or predatory (phage) interactions [60]. Although in recent years an increase in AMR has primarily been pinned on the use and misuse of antibiotics in humans and in animals, there is strong evidence suggesting that AMR dissemination is fueled by other factors with the environment being an important conduit [61]. However, AMR in itself is an ancient phenomenon [1] that has largely evolved in response to natural antibiotics produced by microorganisms themselves to provide a competitive advantage. As a result of these microorganism interactions bacteria have developed resistance strategies against these natural products to mitigate competition [62]. Additionally, to avoid suicide, antibiotic-producing microorganisms themselves often contain at least one gene conferring resistance against the potentially harmful secondary metabolites that the microorganism produces [63, 64]. Leveraging these naturally available

compounds produced by bacteria, anthropogenic efforts have led to antibiotic production which are either natural products of microorganisms, semi-synthetically produced from natural products, and/or chemically synthesized based naturally-available products [65]. Therefore, the use of antibiotics, both natural and (semi-)synthetic, has created unparalleled conditions for the spread of AMR through various reservoirs.

Human

It has long been recognized that the microbiome affects human health through its influence on gut maturation, immune responses, digestion of food, and pathogen resistance [66]. A majority of the microorganisms constituting the human microbiome are commensals contributing to both essential functions and physiological development. However, commensal and bacteria from the immediate and built environments can also be key distributors of AMR to the microbial community with the potential to spread to pathogenic bacteria [67, 68]. Recent evidence suggests that ARGs in environmental bacteria can be taken up by human-associated and pathogenic bacteria [69], thereby posing a considerable threat to human health. Schmidt *et al.* demonstrated that the gut microbiota strains found in patients across five countries, indicated an endogenous transmission, whereby strains found in the oral cavity were transmitted to the gut [70]. Interestingly, the oral cavity has been reported to be a microbial reservoir contributing to the resistome [71] and it is plausible that this in turn is linked to the environment itself [72, 73] including sanitary conditions [74]. While sanitary conditions such as open defecation, access to clean water have been discussed extensively [74], ARGs were recently discovered to be transmitted via air [75, 76] in conjunction with a report by Gilbert *et al.* where ARGs were found in airborne bacteria found in a hospital setting [77].

During the recent decades, research has predominantly focused on AMR prevalence within clinically-relevant bacteria. For example, extended spectrum beta-lactamases (ESBL)-producing and carbapenem resistant *K. pneumoniae* isolates have been characterized as early as 2001 by Yigit *et al.* [78]. Similarly, several studies have reported on the mechanisms of ESBL- [79–85] and plasmid-mediated AmpC-producing *Escherichia coli* [86, 87] rendering the bacteria resistant to third-generation cephalosporins. From a surveillance perspective, Sepp *et al.* screened 10,780 clinical strains using whole genome sequencing to investigate the prevalence of ESBL-, AmpC-, and Carbapenemase-producing *E. coli* across northern and eastern Europe [88]. Despite a low prevalence of ESBL-, AmpC-, Carbapenemase-producing *E. coli* strains, they identified inter-country differences in the distribution and prevalence of resistance genes [88]. Other studies have included research on carbapenem-resistant *Acinetobacter baumannii* [89–91] and *Pseudomonas aeruginosa* [30, 92, 93], vancomycin-resistant *Enterococcus faecium* [94, 95], methicillin-resistant *Staphylococcus aureus* [96–98], penicillin-

resistant *Streptococcus pneumoniae* [99, 100] as well as fluoroquinolone resistant *Salmonella* [101, 102] and *Shigella* species [103, 104]. More recently, less known human pathogens such as *Corynebacterium diphtheriae* isolates have been reported to carry penicillin, macrolide and multidrug resistance [105].

Livestock, poultry and other animals

In livestock and poultry, especially in food production, antibiotics are used as metaphylactics and prophylactics, for disease control and treatment, as well as for growth augmentation. On the one hand, metaphylactics involve the treatment of all animals belonging to the same flock or pen where a clinically sick animal is identified. This is a mitigation strategy which allows for treatment prior to observable clinical signs of disease, for example, by water-based medication [106] [107], simultaneously shortening the overall treatment period. Holman *et al.* investigated the effect of metaphylactic antibiotic usage of the common veterinary antibiotics (oxytetracycline and tulathromycin) on the bovine fecal and nasopharyngeal microbiomes. In addition to shifts in the microbial composition after the first five days of treatment, they found an increase in the relative abundance of several antibiotic resistance genes in both microbiomes at either day 12 or 34 after treatment [108]. Prophylactics, on the other hand, are used to either eradicate a specific pathogen or treat healthy animals as a preventive measure especially during periods of disease susceptibility, e.g. early weaning of piglets [107]. Despite the utility of such treatments including low-dose antibiotics, Agga *et al.* demonstrated that prophylactic treatment limited shipping fever in weanling pigs [109], they may however over protracted periods of use result in a selective pressure yielding resistant bacteria. Consequently, in many countries the use of antibiotics as prophylactic or for pathogen eradication in livestock is prohibited [107].

Apart from their use in infectious disease management, antibiotics are also used as growth promoters, whereby industrialized animal production includes antibiotics as feed supplements [110]. The low concentrations of antibiotics, similar to the levels used in prophylactics, additionally raises the possibility of emergent resistant bacteria due to longer-term selective pressure. In this context, in a five-year longitudinal study, Aarestrup *et al.* investigated the use of growth promotion in pigs and broilers. They found a concomitant increase in AMR in *Enterococcus* spp. isolated from the animals. Moreover, the mitigation of AMR was associated with the banning of antibiotics as growth promoters over the years [111], strongly suggesting the need for measures to reduce the emergence of resistant bacteria.

Even though the emergence of resistant pathogens is a critical consideration, of more immediate concern is the spread of ARGs from the animal microbiome to human microbiota through the acquisition of AMR gene complements. Such spread

can occur via multiple routes, one of which is the direct transmission through food products, i.e. meat and eggs, especially through confined animal feeding operations (CAFOs). Multiple studies have reported food animals as a source of AMR. Examples include multidrug-resistant *Salmonella* from poultry [112], cephalosporin resistant *E. coli* from veal calves [113] and carbapenem resistant *E. coli* from pigs [114, 115] to name a few. In a study by Morrison and Rubin a number of carbapenem resistant bacteria including *Pseudomonas*, *Stenotrophomonas* and *Myroides* species were identified in a variety of seafood products [116]. This phenomenon reiterates the argument that non-pathogenic bacteria, regularly excluded from surveillance programs, may indeed serve as a reservoir for AMR along the food supply chain [116, 117]. Furthermore, resistant bacteria may also be spread from animals to humans through direct contact such as in the agricultural sector [60]. For example, in a study by Rinsky *et al.* livestock-associated multidrug-resistant *S. aureus* was identified in workers at an industrial livestock operation but was not detected in workers at an antibiotic-free livestock operation [118]. These reports collectively underline the need for a more comprehensive analysis and monitoring of livestock reservoirs of AMR.

Interestingly, CAFOs have also been reported to be AMR reservoirs and a source of resistant organisms in migratory birds [119]. Similarly, other studies following the migratory patterns of birds found multi-drug resistant bacteria (*Enterococcus* spp., *Salmonella* spp. and *Vibrio* spp.) in bird fecal material [120]. Other findings simultaneously highlight the role of migratory birds in disseminating extended-spectrum β -lactamase (ESBL)-producing *E. coli* to Bangladesh [121]. Given the propensity for these birds to come in contact with humans in populated countries like Bangladesh, it is likely that these ARGs may in turn influence human health or likely disseminate within the human population.

In general terms, the role of human-influenced environments in sustaining and disseminating AMR is largely unexplored. A comprehensive study by Plasa-Rodriguez *et al.* found AMR associated with several bacterial species in wild boar, roe deer, wild ducks and geese [122]. Atterby *et al.* [123] previously reported the possibility of human-mediated environmental pollution as a source of AMR in wild gulls. Simultaneously, other reports have indicated that clinically relevant AMR bacteria have been found in synanthropic birds partially mediated via human-influenced habitats such as landfills or areas with intensive agriculture [124]. Such anthropogenic influences have spread even to the polar regions, whereby antibiotic-resistant *E. coli* were found in penguin feces, while ESBL-type resistant genes were observed in bacteria such as *E. coli* and *K. pneumoniae* isolated from both seawater and Arctic birds [125].

Environment

The environment is a critical factor for the prediction of emergent and resistant pathogens by understanding the presence, origins and mechanisms of dissemination of AMR. Polluted environments (e.g. with heavy-metals, biocides) further contribute to the evolution and spread of AMR through co-selection. For instance, through cross-resistance, a single genetic mutation may mediate resistance to both metals and antibiotics, or through co-segregation where both metal- and antibiotic resistance genes are localized on the same MGE [126, 127]. The risk of a specific environment being contaminated with AMR is often based on the interaction between the different environments. Built environments in particular, e.g. hospitals and extended care facilities, where bacteria are exposed to high and repeated doses of antibiotics, are hotspots of AMR. Hospitals in specific are of high interest to study both the evolution and dissemination of AMR through the prevalence of hospital-acquired infections of resistant bacteria. Resistant pathogens may enter the hospital environment via infected patients or acquire resistance through in-hospital evolution. In both cases resistant pathogens may spread epidemically between patients or the ARG itself can be transmitted through HGT into other genetic backgrounds [128]. Furthermore, sewage from both the hospital and the general population are ultimately transported to wastewater treatment plants (WWTP).

Urban WWTPs therefore provide a vast reservoir of antimicrobial resistance [129] and are considered to be AMR hotspots with respect to resistant bacteria and ARGs [130]. Moreover, the extensive dissemination of ARGs between various bacterial species through HGT may facilitate the transfer of ARGs to pathogenic bacteria. For example, Alexander *et al.* identified facultative pathogenic bacteria such as *E. coli*, *K. pneumoniae*, *P. aeruginosa*, and Enterococci with 12 clinically relevant ARGs within 23 different WWTPs [131]. Additionally, since WWTPs generally do not have the necessary measures to remove either ARGs or resistant bacteria, these are then released into the receiving water bodies, promoting their dissemination into and through the aquatic environment [131]. This is in line with a study by Osinska *et al.* where a significant increase in ARGs (e.g. *bla*_{TEM}, *tetA*, *sulI1*) was identified downstream of the WWTP when compared to the upstream river water [132]. A similar study by Bueno *et al.* reported a significant increase in 17 ARGs contributing to aminoglycoside-, beta-lactam-, diaminopyrimidine-, fluoroquinolone-, sulfonamide-, tetracycline- and multidrug-resistance, in the receiving water of three different WWTPs [133], thereby highlighting the overall role and impact of the built-environment in AMR dissemination.

The contamination of natural environments with antibiotics originating from built environments as well as agricultural sources, results in selective pressure promoting both the evolution and the spread of ARGs. Additionally, many antibiotics are naturally produced by fungal and bacterial strains and consequently have been used by microorganisms as a competitive mechanism [134, 135]. Due to their high complexity and multi-faceted interactions of microorganism, soil microbiomes are considered a hotbed for the evolution and development of AMR [136]. Multiple bacteria identified in soils encode genes that either degrade or inactivate antibiotics. For instance, Dantas *et al.* isolated hundreds of soil-dwelling bacteria capable of utilizing antibiotics as a carbon source and found up to 17 antibiotics, including those of synthetic origin, supporting the growth of clonal soil bacteria [137]. Furthermore, bacteria isolated from forest, urban and agricultural soils have been found to have highly varied resistomes, even in some cases harbouring novel mutation sites conferring resistance [138]. Linked to agricultural soils, the plant rhizosphere is of further interest due to the transmission of ARGs from soil to plants via the rhizosphere microbial community. Wolters *et al.* investigated the effect of various organic soil fertilizers such as manure and found increased relative abundances of sulfonamide and tetracycline resistance in the maize rhizosphere [139]. Similarly, Song *et al.* investigated the abundance of 35 antibiotic resistance genes in the rhizosphere of 10 plant species and identified a positive association between ARGs and MGEs [140].

Similar to soils, aquatic environments also represent known reservoirs of AMR. Aquatic habitats harbour resistant microorganisms such as carbapenem-resistant *Acinetobacter* spp. in rivers [141], carbapenem-resistant *Pseudomonas* in coastal waters [142], and carbapenem-resistant Enterobacteriales in seawater [143]. Environments are further affected greatly when in proximity to anthropogenic activity such as pharmaceutical industries. Consequently, these environments are abundant with ARGs and multidrug-resistant bacteria which have been associated with a high impact on human health [144]. For instance, Flach *et al.* found that antibiotic-polluted lakes harbored considerably higher proportions of ciprofloxacin- and sulfamethoxazole-resistant bacteria as well as several novel multi-resistance plasmids compared to non-polluted lakes [145]. Additionally, Kristiansson *et al.* identified a similar phenomenon in river sediments exposed to antibiotic pharmaceutical wastewater and reported high levels of ciprofloxacin-resistance as well as corresponding mobile quinolone resistance genes [146].

AMR, on the other hand, does not exclusively exist in human-impacted environments. As several studies have revealed, vast reservoirs of AMR are also found in environments pre-dating the antibiotic era [135]. These include glacier lakes, remote lakes [90] and oceans [103, 104]. Polar regions in particular, as one of the least human-impacted environments to date, are of interest for the study of AMR. Arctic soil isolates have previously revealed the presence of multidrug efflux

pumps [147], while in a study by Dancer *et al.*, bacterial isolates from arctic glacial ice and water were found to carry resistance to antibiotics such as cefazolin, cefamandole and ampicillin [148]. The melting of glaciers and icecaps due to climate change, therefore, may give important insights into, potentially, prehistoric mechanisms of AMR. On the other hand, this may also lead to the remobilization of ARGs, which we have not seen since before the dawn of human evolution.

Metagenomic approaches in assessing antimicrobial resistance: a One Health perspective

Resistant bacteria residing within human, animal and environmental reservoirs may spread from one to the other, at both local and global levels (Figure 1). This phenomenon has the potential to rapidly trigger a pandemic where AMR is no longer constrained by either geographic or human-animal borders [149]. It is therefore necessary to understand the dissemination of antibiotic resistance by characterizing the resistome within various environments and to unravel how they act as a reservoir for bacterial pathogens in the context of overall pandemic preparedness. A One Health perspective integrating research on AMR as well as resistant microorganisms, circulating in humans, animals and the environment is therefore crucial to enhance our understanding of the complex epidemiology of antimicrobial resistance [149].

In recent years, many studies, of which some have been discussed in the previous sections, have used different techniques to sample the resistomes of soils, wastewater, as well as human and animal microbiota [16]. Recent metagenomic studies

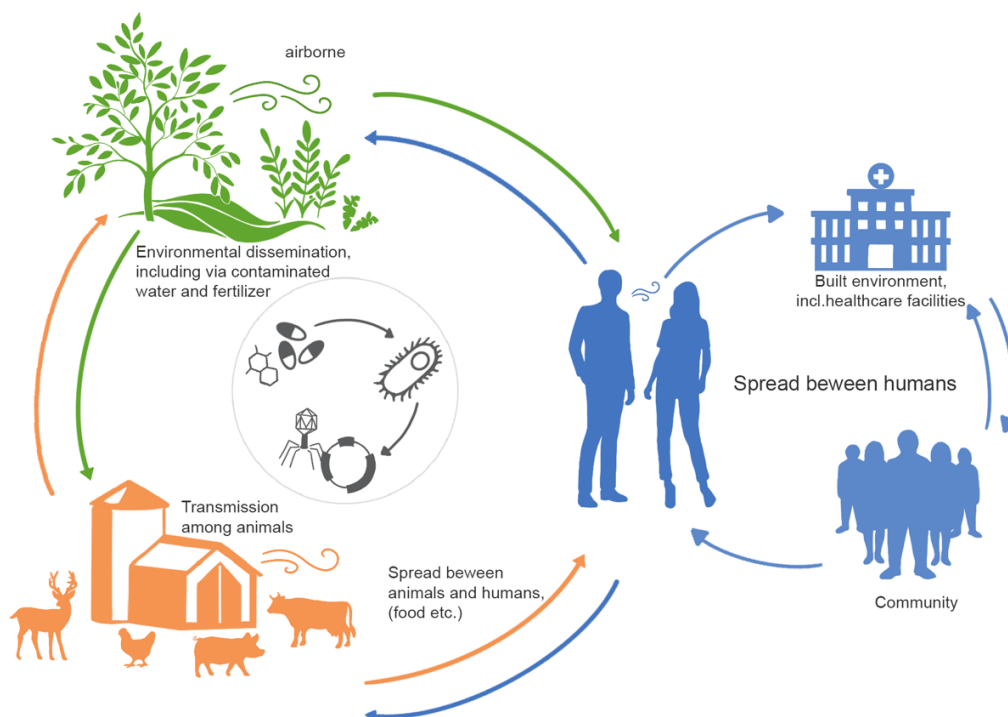


Figure 1: AMR dissemination in One Health.

MGE-mediated (i.e. phage, plasmids and integrons) dissemination of AMR across different biomes.

by both Gibson *et al.* [114] and Munck *et al.* [115] suggest that ARGs predominantly cluster by microbial reservoir, implying that the resistomes in soils and WWTPs differ significantly from those found in human pathogens. Gibson *et al.* found that resistance against β -lactams and tetracyclines differed mostly between ecosystems [150] while Munck *et al.* highlighted that only a few genes within the WWTP core resistome were found in other environments [151]. Nonetheless, part of these resistomes may still be shared and the importance of continued exploration of the resistome in such environments should be stressed [16]. While shared resistome elements between various microbial reservoirs are of interest to understand the dissemination of AMR, resistome differences between ecosystems are equally, if not more important. They represent a pool of potential novel resistance mechanisms and thereby a likely threat to public health.

As described within this review, several published studies have investigated AMR in humans, animals or the wider environment. However, many of these focus specifically on the ESKAPEE pathogens (*Enterococcus geacium*, *S. aureus*, *K. pneumoniae*, *A. baumannii*, *P. aeruginosa*, *Enterobacter* spp., and *E. coli*), which have been classified by the World Health Organisation for their high to critical drug-resistance. They are also of particular interest due to their increased resistance to last-resort drugs [152]. There presently exists no lack in reports of the resistance mechanisms encoded by the ESKAPEE pathogens in different microbial reservoirs. Methicillin-resistant *S. aureus* (MRSA) for instance, has been

reported by van den Broek *et al.* [153] and Lewis *et al.* [154] to be both human- and animal-associated with a high risk for zoonotic transmission. Similarly, Ruiz-Roldan *et al.* reported the presence of resistant *P. aeruginosa* in animals in addition to humans. On the other hand the drug resistant strains of ESKAPEE pathogens belonging to the Enterobacteriales order (i.e. *E. coli* and *P. aeruginosa*) have been extensively described in all microbial reservoirs [78–80, 143, 155, 156]. Recent research has been extended to focus on other pathogens posing a threat to human health such as resistant *Campylobacter jejuni* where infections have been reported in both humans, animals and the environment [46, 157, 158][46, 157]. Similarly, other reports include multidrug-resistant *Salmonella* which have been identified in human [159, 160], animal [159, 161] and environmental reservoirs [162].

While the above studies are focused on specific pathogens or resistance categories, research utilizing sequence-based metagenomics provides a comprehensive perspective on all ARGs within different microbial reservoirs. For instance, Forslund *et al.* provide extensive insights into the human gut resistomes of 832 individuals spanning 10 geographical areas. They reported significant differences in gut resistance potential between countries resulting from differences in antibiotic usage as well as direct links to medical and food production activities [163]. Other metagenomic studies have focused on the development of the resistome early on in life with several studies reporting a diversity of ARGs within the infant gut [164, 165]. During the first days of life the bacteria colonizing the infant gut originate primarily from the mother's birth canal, the living environment and handling by other individuals. Birth mode affects colonization since vaginally born infants are colonized firstly by fecal and vaginal bacteria from the mother, while infants born via cesarean section are initially exposed to bacteria originating from both the hospital environment and healthcare workers [66, 166]. Therefore, infants born by cesarean section may also have a higher chance of acquiring hospital-mediated AMR and thereby resistant bacteria [164]. Other metagenomic studies have focused on the animal resistome, especially food production animals, such as dairy cattle, revealing an increase in AMR linked to heavy metal-contaminated environments [167]. Furthermore, a study by Skarzynska *et al.* leveraged metagenomic data to study AMR in the gut of both wild (boars, foxes and rodents) and domestic (chicken, turkey and pig) animals. Importantly, they identified increased AMR abundance in farm animals compared to wildlife [168]. Furthermore, the lowest AMR abundance in this study was observed in wild rodents due to their limited exposure to antimicrobials. In this context, further evidence was found linking ARGs conferring resistance to important antimicrobials such as quinolones and cephalosporins to wild foxes [168]. Alongside human and animal studies, metagenomic studies on the environmental resistome focus on characterizing AMR either in WWTPs or the natural environment or built environment (e.g. healthcare facilities). However, few are specifically tailored

towards understanding the role of the environmental ecosystems as microbial reservoirs of AMR, especially in a One Health setting.

The few metagenomic studies that are focused on multiple microbial reservoirs still largely target only one side of the One Health triad, e.g. human-animal [169–172], animal-environment [173–176] or environment-human [150, 151, 177–180]. Nonetheless, some studies have pursued a complete One Health AMR approach [181, 182]. Li *et al.* investigated wide-spectrum profiles of ARGs and their co-occurrence patterns in 50 samples from 10 microbial reservoirs, spanning human, environment and animal habitats. They found that samples could be clustered into four groups according to AMR abundance, with samples derived from livestock and wastewater demonstrating the highest abundance followed by humans, and with the lowest abundance found in sediments, soil, river and drinking water, in that particular order. A widespread occurrence of vancomycin resistance genes was identified in all environments except from river sediments and drinking-water [182]. Another study by Pal *et al.* investigated AMR, MGEs and bacterial taxonomic compositions of 864 human, 145 animal and 369 environmental metagenomes. Both human and animal microbial communities demonstrated a limited taxonomic diversity, a low abundance/diversity of biocide and metal resistance genes and MGEs, yet a high abundance in ARGs. Additionally, a number of ARGs corresponding to aminoglycoside, macrolide, beta-lactam and tetracycline resistance was found to be widespread and present in almost all of the investigated environments [181]. Collectively, these studies report the cross-domain similarities and likely transmission of AMR in a One Health setting, potentially highlighting the need for more in-depth characterization of AMR transmission mechanisms.

Understanding One Health reservoirs and future perspectives

Antimicrobial resistance is an ever-present challenge, not necessarily due to the use of antibiotics alone, but also due to the role of mobile genetic elements. One major challenge still faced by most One Health studies is attributing the directionality of ARGs between various metagenomes. While some of the discussed studies strive to do so, it is, with rare exceptions, impossible to accurately attribute directionality of transmission due to limitations of the current methods. To infer directionality in microbial populations, studies have focused on identifying similarities of bacterial and/or plasmid sequences along with the ARGs they encode. However, these overlapping patterns do not take into account co-colonization from a shared source, nor do they allow for interpretation of directionality [183]. In contrast, Mather *et al.* quantified the relative contributions of animal- and human- derived multidrug resistant *Salmonella* isolates using the phylogenetic association of the bacterium and its antimicrobial resistance genes over the course of an epidemic. They subsequently

determined that there was only a limited transmission in either direction, while the bacterium and its resistance genes were largely independently maintained within animal and human populations [184]. Collectively, to accurately reconstruct patterns of transmission, especially directionality of said transmission, one needs to further combine both (meta)genomic data analysis, including phylogenetic analysis, with epidemiological approaches [183]. What appears evident is that every AMR reservoir may affect another. This is further compounded by the recent discovery of giant extrachromosomal elements such as “borgs” in *Methanoperedens* archaea, which may be capable of augmenting microbial activity by encoding putative resistance genes and also via HGT [185]. Thereby, understanding the interactions/mechanisms and role of each component contributing to the spread of AMR is a critical step in monitoring this ultimate challenge to human health and wellbeing. Therefore, recognizing the One Health reservoirs of antimicrobial resistance is an important first-step towards this goal. Several methods exist both *in vitro* and *in silico* to identify the potential resistance genes and categories found in commensal microorganisms alongside well-characterized pathogens. However, future endeavors including molecular validation of identified AMR with the help of meta-omics will be required. Furthermore, combined methods incorporating the identity of ARGs, modes of transmission and integration into the individual reservoirs, alongside crossover mechanisms may be needed for comprehensive characterization of AMR dissemination mechanisms.

Strategies designed to mitigate AMR and its dissemination should simultaneously focus on improving awareness and understanding of AMR through education and training across all affected groups including farmers, veterinarians, physicians, CAFOs and also the general public. It is prudent to complement these strategies with increased surveillance and research to identify AMR at the regional, national and international levels [186]. Furthermore, mitigation, education and surveillance will need to be accompanied by the respective sanitation, hygiene and infection prevention measures. Prevention of infection and further AMR dissemination could be instituted in healthcare settings [187], at the farm level [188] and also in animal disease control programs [189]. Given the possibility of interactions between humans, livestock, animals, and the environment, future studies on human health and disease will benefit from the consequential incorporation of One Health reservoirs into all aspects of studies regarding antimicrobial resistance.

424 Declarations

425 Funding

426 P.W. acknowledges the European Research Council (ERC-CoG 863664). L.dN. and P.W. were supported by the
427 Luxembourg National Research Fund PRIDE17/11823097. S.B.B. was supported by the Synergia grant (CRSII5_180241)
428 through the Swiss National Science Foundation.

429 Conflicts of interest

430 The authors do not have any conflicts of interest relating to this work.

431 Availability of data and material

432 Not applicable

433 Code availability

434 Not applicable

435 Authors' contributions

436 LdN designed and created the overview figure. LdN, SBB, PW conceptualized the review and contributed to the
437 writing.

438 Acknowledgements

439 We are grateful for the feedback and input by Dr. Deepthi Budagavi.

440 Ethics approval

441 Not applicable

442 Consent to participate

443 Not applicable

444 Consent for publication

445 The authors consent to publication.

446 References

- 447 1. D'Costa VM, King CE, Kalan L, et al (2011) Antibiotic resistance is ancient. *Nature* 477:457–461.
448 <https://doi.org/10.1038/nature10388>
- 449 2. Wright GD (2007) The antibiotic resistome: the nexus of chemical and genetic diversity. *Nat Rev Microbiol* 5:175–
450 186. <https://doi.org/10.1038/nrmicro1614>
- 451 3. Zhang X-S, Li J, Krautkramer KA, et al (2018) Antibiotic-induced acceleration of type 1 diabetes alters maturation
452 of innate intestinal immunity. *Elife* 7:e37816. <https://doi.org/10.7554/eLife.37816>
- 453 4. Roubaud-Baudron C, Ruiz VE, Swan AM Jr, et al (2019) Long-Term Effects of Early-Life Antibiotic Exposure on
454 Resistance to Subsequent Bacterial Infection. *MBio* 10.: <https://doi.org/10.1128/mBio.02820-19>
- 455 5. O'Neill J (2014) Antimicrobial resistance: tackling a crisis for the health and wealth of nations. Review on
456 antimicrobial resistance
- 457 6. Brogan DM, Mossialos E (2016) A critical analysis of the review on antimicrobial resistance report and the
458 infectious disease financing facility. *Global Health* 12:8. <https://doi.org/10.1186/s12992-016-0147-y>
- 459 7. Reygaert WC (2018) An overview of the antimicrobial resistance mechanisms of bacteria. *AIMS Microbiol* 4:482–
460 501. <https://doi.org/10.3934/microbiol.2018.3.482>
- 461 8. Blair JMA, Richmond GE, Piddock LJV (2014) Multidrug efflux pumps in Gram-negative bacteria and their role in
462 antibiotic resistance. *Future Microbiol* 9:1165–1177. <https://doi.org/10.2217/fmb.14.66>
- 463 9. Karaman R, Jubeh B, Breijyeh Z (2020) Resistance of Gram-Positive Bacteria to Current Antibacterial Agents and
464 Overcoming Approaches. *Molecules* 25.: <https://doi.org/10.3390/molecules25122888>
- 465 10. Kumar A, Schweizer HP (2005) Bacterial resistance to antibiotics: active efflux and reduced uptake. *Adv Drug*
466 *Deliv Rev* 57:1486–1513. <https://doi.org/10.1016/j.addr.2005.04.004>

- 467 11. Mah T-F (2012) Biofilm-specific antibiotic resistance. *Future Microbiol* 7:1061–1072.
468 <https://doi.org/10.2217/fmb.12.76>
- 469 12. Redgrave LS, Sutton SB, Webber MA, Piddock LJV (2014) Fluoroquinolone resistance: mechanisms, impact on
470 bacteria, and role in evolutionary success. *Trends Microbiol* 22:438–445. <https://doi.org/10.1016/j.tim.2014.04.007>
- 471 13. Huovinen P, Sundström L, Swedberg G, Sköld O (1995) Trimethoprim and sulfonamide resistance. *Antimicrob*
472 *Agents Chemother* 39:279–289. <https://doi.org/10.1128/aac.39.2.279>
- 473 14. Martinez JL (2014) General principles of antibiotic resistance in bacteria. *Drug Discov Today Technol* 11:33–39.
474 <https://doi.org/10.1016/j.ddtec.2014.02.001>
- 475 15. Davies J, Davies D (2010) Origins and evolution of antibiotic resistance. *Microbiol Mol Biol Rev* 74:417–433.
476 <https://doi.org/10.1128/MMBR.00016-10>
- 477 16. von Wintersdorff CJH, Penders J, van Niekerk JM, et al (2016) Dissemination of Antimicrobial Resistance in
478 Microbial Ecosystems through Horizontal Gene Transfer. *Front Microbiol* 7:173.
479 <https://doi.org/10.3389/fmicb.2016.00173>
- 480 17. Leclerc QJ, Lindsay JA, Knight GM (2019) Mathematical modelling to study the horizontal transfer of
481 antimicrobial resistance genes in bacteria: current state of the field and recommendations. *J R Soc Interface*
482 16:20190260. <https://doi.org/10.1098/rsif.2019.0260>
- 483 18. Lopatkin AJ, Huang S, Smith RP, et al (2016) Antibiotics as a selective driver for conjugation dynamics. *Nat*
484 *Microbiol* 1:16044. <https://doi.org/10.1038/nmicrobiol.2016.44>
- 485 19. Bello-López JM, Cabrero-Martínez OA, Ibáñez-Cervantes G, et al (2019) Horizontal Gene Transfer and Its
486 Association with Antibiotic Resistance in the Genus *Aeromonas* spp. *Microorganisms* 7.:
487 <https://doi.org/10.3390/microorganisms7090363>
- 488 20. MacLean RC, San Millan A (2019) The evolution of antibiotic resistance. *Science* 365:1082–1083.
489 <https://doi.org/10.1126/science.aax3879>
- 490 21. Carattoli A (2013) Plasmids and the spread of resistance. *Int J Med Microbiol* 303:298–304.
491 <https://doi.org/10.1016/j.ijmm.2013.02.001>

22. Bassetti M, Righi E, Carnelutti A, et al (2018) Multidrug-resistant *Klebsiella pneumoniae*: challenges for treatment, prevention and infection control. *Expert Rev Anti Infect Ther* 16:749–761.
<https://doi.org/10.1080/14787210.2018.1522249>
23. Huang T-W, Chen T-L, Chen Y-T, et al (2013) Copy Number Change of the NDM-1 sequence in a multidrug-resistant *Klebsiella pneumoniae* clinical isolate. *PLoS One* 8:e62774. <https://doi.org/10.1371/journal.pone.0062774>
24. Salinas L, Cárdenas P, Johnson TJ, et al (2019) Diverse Commensal *Escherichia coli* Clones and Plasmids Disseminate Antimicrobial Resistance Genes in Domestic Animals and Children in a Semirural Community in Ecuador. *mSphere* 4.: <https://doi.org/10.1128/mSphere.00316-19>
25. Porse A, Schou TS, Munck C, et al (2018) Biochemical mechanisms determine the functional compatibility of heterologous genes. *Nat Commun* 9:522. <https://doi.org/10.1038/s41467-018-02944-3>
26. Soucy SM, Huang J, Gogarten JP (2015) Horizontal gene transfer: building the web of life. *Nat Rev Genet* 16:472–482. <https://doi.org/10.1038/nrg3962>
27. Partridge SR, Kwong SM, Firth N, Jensen SO (2018) Mobile Genetic Elements Associated with Antimicrobial Resistance. *Clin Microbiol Rev* 31.: <https://doi.org/10.1128/CMR.00088-17>
28. Stalder T, Barraud O, Casellas M, et al (2012) Integron involvement in environmental spread of antibiotic resistance. *Front Microbiol* 3:119. <https://doi.org/10.3389/fmicb.2012.00119>
29. Buongiorno Pereira M, Österlund T, Eriksson KM, et al (2020) A comprehensive survey of integron-associated genes present in metagenomes. *BMC Genomics* 21:495. <https://doi.org/10.1186/s12864-020-06830-5>
30. Liapis E, Bour M, Triponney P, et al (2019) Identification of diverse integron and Plasmid structures carrying a novel carbapenemase among *Pseudomonas* species. *Front Microbiol* 10:404.
<https://doi.org/10.3389/fmicb.2019.00404>
31. Rowe-Magnus DA, Guerout A-M, Ploncard P, et al (2001) The evolutionary history of chromosomal super-integrations provides an ancestry for multiresistant integrons. *Proc Natl Acad Sci U S A* 98:652–657.
<https://doi.org/10.1073/pnas.98.2.652>
32. Colavecchio A, Cadieux B, Lo A, Goodridge LD (2017) Bacteriophages Contribute to the Spread of Antibiotic

- Resistance Genes among Foodborne Pathogens of the Enterobacteriaceae Family - A Review. *Front Microbiol* 8:1108. <https://doi.org/10.3389/fmicb.2017.01108>
33. Oh J-H, Alexander LM, Pan M, et al (2019) Dietary Fructose and Microbiota-Derived Short-Chain Fatty Acids Promote Bacteriophage Production in the Gut Symbiont *Lactobacillus reuteri*. *Cell Host Microbe* 25:273–284.e6. <https://doi.org/10.1016/j.chom.2018.11.016>
34. McDaniel L, Paul JH (2005) Effect of nutrient addition and environmental factors on prophage induction in natural populations of marine *synechococcus* species. *Appl Environ Microbiol* 71:842–850. <https://doi.org/10.1128/AEM.71.2.842-850.2005>
35. Knowles B, Silveira CB, Bailey BA, et al (2016) Lytic to temperate switching of viral communities. *Nature* 531:466–470. <https://doi.org/10.1038/nature17193>
36. Dion MB, Oechslin F, Moineau S (2020) Phage diversity, genomics and phylogeny. *Nat Rev Microbiol* 18:125–138. <https://doi.org/10.1038/s41579-019-0311-5>
37. Chiang YN, Penadés JR, Chen J (2019) Genetic transduction by phages and chromosomal islands: The new and noncanonical. *PLoS Pathog* 15:e1007878. <https://doi.org/10.1371/journal.ppat.1007878>
38. Boolchandani M, D’Souza AW, Dantas G (2019) Sequencing-based methods and resources to study antimicrobial resistance. *Nat Rev Genet* 20:356–370. <https://doi.org/10.1038/s41576-019-0108-4>
39. Jorgensen JH, Ferraro MJ (2009) Antimicrobial susceptibility testing: a review of general principles and contemporary practices. *Clin Infect Dis* 49:1749–1755. <https://doi.org/10.1086/647952>
40. Lepage P, Leclerc MC, Joossens M, et al (2013) A metagenomic insight into our gut’s microbiome. *Gut* 62:146–158. <https://doi.org/10.1136/gutjnl-2011-301805>
41. Hendriksen RS, Bortolaia V, Tate H, et al (2019) Using Genomics to Track Global Antimicrobial Resistance. *Front Public Health* 7:242. <https://doi.org/10.3389/fpubh.2019.00242>
42. Arango-Argoty G, Garner E, Pruden A, et al (2018) DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* 6:23. <https://doi.org/10.1186/s40168-018-0401-z>
43. Alcock BP, Raphenya AR, Lau TTY, et al (2020) CARD 2020: antibiotic resistome surveillance with the

- comprehensive antibiotic resistance database. *Nucleic Acids Res* 48:D517–D525.
<https://doi.org/10.1093/nar/gkz935>
44. Zankari E, Hasman H, Cosentino S, et al (2012) Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother* 67:2640–2644. <https://doi.org/10.1093/jac/dks261>
 45. Gupta SK, Padmanabhan BR, Diene SM, et al (2014) ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob Agents Chemother* 58:212–220.
<https://doi.org/10.1128/AAC.01310-13>
 46. Feldgarden M, Brover V, Haft DH, et al (2019) Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrob Agents Chemother* 63.: <https://doi.org/10.1128/AAC.00483-19>
 47. Krawczyk PS, Lipinski L, Dziembowski A (2018) PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res* 46:e35. <https://doi.org/10.1093/nar/gkx1321>
 48. Robertson J, Nash JHE (2018) MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom* 4.: <https://doi.org/10.1099/mgen.0.000206>
 49. Arredondo-Alonso S, Bootsma M, Hein Y, et al (2020) gplas: a comprehensive tool for plasmid analysis using short-read graphs. *Bioinformatics* 36:3874–3876. <https://doi.org/10.1093/bioinformatics/btaa233>
 50. Ren J, Song K, Deng C, et al (2018) Identifying viruses from metagenomic data by deep learning. *arXiv [q-bio.GN]*
 51. Roux S, Enault F, Hurwitz BL, Sullivan MB (2015) VirSorter: mining viral signal from microbial genomic data. *PeerJ* 3:e985. <https://doi.org/10.7717/peerj.985>
 52. Amgarten D, Braga LPP, da Silva AM, Setubal JC (2018) MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. *Front Genet* 9:304. <https://doi.org/10.3389/fgene.2018.00304>
 53. Fang Z, Tan J, Wu S, et al (2019) PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *Gigascience* 8.: <https://doi.org/10.1093/gigascience/giz066>
 54. Atlas, Ronald M. (2013) One Health: Its Origins and Future. In: Mackenzie JS, Jeggo M, Daszak P, Richt JA (eds)

One Health: The Human-Animal-Environment Interfaces in Emerging Infectious Diseases: The Concept and Examples of a One Health Approach. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 1–13

55. Kim D-W, Cha C-J (2021) Antibiotic resistome from the One-Health perspective: understanding and controlling antimicrobial resistance transmission. *Exp Mol Med* 53:301–309. <https://doi.org/10.1038/s12276-021-00569-z>

56. de Nies L, Lopes S, Busi SB, et al (2021) PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. *Microbiome* 9:49. <https://doi.org/10.1186/s40168-020-00993-9>

57. Kultima JR, Coelho LP, Forslund K, et al (2016) MOCAT2: a metagenomic assembly, annotation and profiling framework. *Bioinformatics* 32:2520–2523. <https://doi.org/10.1093/bioinformatics/btw183>

58. Beghini F, McIver LJ, Blanco-Míguez A, et al (2021) Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *Elife* 10.: <https://doi.org/10.7554/eLife.65088>

59. Berg G, Rybakova D, Fischer D, et al (2020) Microbiome definition re-visited: old concepts and new challenges. *Microbiome* 8:103. <https://doi.org/10.1186/s40168-020-00875-0>

60. Trinh P, Zaneveld JR, Safranek S, Rabinowitz PM (2018) One Health Relationships Between Human, Animal, and Environmental Microbiomes: A Mini-Review. *Front Public Health* 6:235. <https://doi.org/10.3389/fpubh.2018.00235>

61. Graham DW, Bergeron G, Bourassa MW, et al (2019) Complexities in understanding antimicrobial resistance across domesticated animal, human, and environmental systems. *Ann N Y Acad Sci* 1441:17–30. <https://doi.org/10.1111/nyas.14036>

62. Granato ET, Meiller-Legrand TA, Foster KR (2019) The Evolution and Ecology of Bacterial Warfare. *Curr Biol* 29:R521–R537. <https://doi.org/10.1016/j.cub.2019.04.024>

63. Cundliffe E, Demain AL (2010) Avoidance of suicide in antibiotic-producing microbes. *J Ind Microbiol Biotechnol* 37:643–672. <https://doi.org/10.1007/s10295-010-0721-x>

64. Tran PN, Yen M-R, Chiang C-Y, et al (2019) Detecting and prioritizing biosynthetic gene clusters for bioactive compounds in bacteria and fungi. *Appl Microbiol Biotechnol* 103:3277–3287. <https://doi.org/10.1007/s00253-019->

65. Demain AL (2009) Antibiotics: natural products essential to human health. *Med Res Rev* 29:821–842.
<https://doi.org/10.1002/med.20154>
66. Penders J, Stobberingh EE, Savelkoul PHM, Wolffs PFG (2013) The human microbiome as a reservoir of antimicrobial resistance. *Front Microbiol* 4:87. <https://doi.org/10.3389/fmicb.2013.00087>
67. Brinkac L, Voorhies A, Gomez A, Nelson KE (2017) The Threat of Antimicrobial Resistance on the Human Microbiome. *Microb Ecol* 74:1001–1008. <https://doi.org/10.1007/s00248-017-0985-z>
68. Blake DP, Hillman K, Fenlon DR, Low JC (2003) Transfer of antibiotic resistance between commensal and pathogenic members of the Enterobacteriaceae under ileal conditions. *J Appl Microbiol* 95:428–436.
<https://doi.org/10.1046/j.1365-2672.2003.01988.x>
69. Stanton IC, Bethel A, Leonard AFC, et al (2020) What is the research evidence for antibiotic resistance exposure and transmission to humans from the environment? A systematic map protocol. *Environ Evid* 9:12.
<https://doi.org/10.1186/s13750-020-00197-6>
70. Schmidt TS, Hayward MR, Coelho LP, et al (2019) Extensive transmission of microbes along the gastrointestinal tract. *Elife* 8.: <https://doi.org/10.7554/eLife.42693>
71. Carr VR, Witherden EA, Lee S, et al (2020) Abundance and diversity of resistomes differ between healthy human oral cavities and gut. *Nat Commun* 11:693. <https://doi.org/10.1038/s41467-020-14422-w>
72. Ben Y, Fu C, Hu M, et al (2019) Human health risk assessment of antibiotic resistance associated with antibiotic residues in the environment: A review. *Environ Res* 169:483–493. <https://doi.org/10.1016/j.envres.2018.11.040>
73. Ben Maamar S, Hu J, Hartmann EM (2019) Implications of indoor microbial ecology and evolution on antibiotic resistance. *J Expo Sci Environ Epidemiol* 30:1–15. <https://doi.org/10.1038/s41370-019-0171-0>
74. Graham DW, Giesen MJ, Bunce JT (2018) Strategic Approach for Prioritising Local and Regional Sanitation Interventions for Reducing Global Antibiotic Resistance. *Water* 11:27. <https://doi.org/10.3390/w11010027>
75. Li J, Cao J, Zhu Y-G, et al (2018) Global Survey of Antibiotic Resistance Genes in Air. *Environ Sci Technol* 52:10975–10984. <https://doi.org/10.1021/acs.est.8b02204>

76. Dueker ME, O'Mullan GD, Martínez JM, et al (2017) Onshore Wind Speed Modulates Microbial Aerosols along an Urban Waterfront. *Atmosphere* 8:215. <https://doi.org/10.3390/atmos8110215>
77. Gilbert Y, Veillette M, Duchaine C (2010) Airborne bacteria and antibiotic resistance genes in hospital rooms. *Aerobiologia* 26:185–194. <https://doi.org/10.1007/s10453-010-9155-1>
78. Yigit H, Queenan AM, Anderson GJ, et al (2001) Novel carbapenem-hydrolyzing beta-lactamase, KPC-1, from a carbapenem-resistant strain of *Klebsiella pneumoniae*. *Antimicrob Agents Chemother* 45:1151–1161. <https://doi.org/10.1128/AAC.45.4.1151-1161.2001>
79. van Hout D, Verschuuren TD, Bruijning-Verhagen PCJ, et al (2020) Extended-spectrum beta-lactamase (ESBL)-producing and non-ESBL-producing *Escherichia coli* isolates causing bacteremia in the Netherlands (2014 - 2016) differ in clonal distribution, antimicrobial resistance gene and virulence gene content. *PLoS One* 15:e0227604. <https://doi.org/10.1371/journal.pone.0227604>
80. Day MJ, Hopkins KL, Wareham DW, et al (2019) Extended-spectrum β -lactamase-producing *Escherichia coli* in human-derived and foodchain-derived samples from England, Wales, and Scotland: an epidemiological surveillance and typing study. *Lancet Infect Dis* 19:1325–1335. [https://doi.org/10.1016/S1473-3099\(19\)30273-7](https://doi.org/10.1016/S1473-3099(19)30273-7)
81. Falgenhauer L, Imirzalioglu C, Oppong K, et al (2018) Detection and Characterization of ESBL-Producing *Escherichia coli* From Humans and Poultry in Ghana. *Front Microbiol* 9:3358. <https://doi.org/10.3389/fmicb.2018.03358>
82. Alegría Á, Arias-Temprano M, Fernández-Natal I, et al (2020) Molecular Diversity of ESBL-Producing *Escherichia coli* from Foods of Animal Origin and Human Patients. *Int J Environ Res Public Health* 17.: <https://doi.org/10.3390/ijerph17041312>
83. Kayastha K, Dhungel B, Karki S, et al (2020) Extended-Spectrum β -Lactamase-Producing *Escherichia coli* and *Klebsiella* Species in Pediatric Patients Visiting International Friendship Children's Hospital, Kathmandu, Nepal. *Infect Dis* 13:1178633720909798. <https://doi.org/10.1177/1178633720909798>
84. Leverstein-van Hall MA, Dierikx CM, Cohen Stuart J, et al (2011) Dutch patients, retail chicken meat and poultry share the same ESBL genes, plasmids and strains. *Clin Microbiol Infect* 17:873–880. <https://doi.org/10.1111/j.1469-0691.2011.03497.x>

- 643 85. Pitout JDD, Laupland KB (2008) Extended-spectrum beta-lactamase-producing Enterobacteriaceae: an emerging
644 public-health concern. *Lancet Infect Dis* 8:159–166. [https://doi.org/10.1016/S1473-3099\(08\)70041-0](https://doi.org/10.1016/S1473-3099(08)70041-0)
- 645 86. Shayan S, Bokaeian M (2015) Detection of ESBL- and AmpC-producing *E. coli* isolates from urinary tract
646 infections. *Adv Biomed Res* 4:220. <https://doi.org/10.4103/2277-9175.166643>
- 647 87. Peter-Getzlaff S, Polsfuss S, Poledica M, et al (2011) Detection of AmpC beta-lactamase in *Escherichia coli*:
648 comparison of three phenotypic confirmation assays and genetic analysis. *J Clin Microbiol* 49:2924–2932.
649 <https://doi.org/10.1128/JCM.00091-11>
- 650 88. Sepp E, Andreson R, Balode A, et al (2019) Phenotypic and Molecular Epidemiology of ESBL-, AmpC-, and
651 Carbapenemase-Producing *Escherichia coli* in Northern and Eastern Europe. *Front Microbiol* 10:2465.
652 <https://doi.org/10.3389/fmicb.2019.02465>
- 653 89. Evans BA, Hamouda A, Amyes SGB (2013) The rise of carbapenem-resistant *Acinetobacter baumannii*. *Curr*
654 *Pharm Des* 19:223–238
- 655 90. Piperaki E-T, Tzouveleakis LS, Miriagou V, Daikos GL (2019) Carbapenem-resistant *Acinetobacter baumannii*: in
656 pursuit of an effective treatment. *Clin Microbiol Infect* 25:951–957. <https://doi.org/10.1016/j.cmi.2019.03.014>
- 657 91. New Treatment Options against Carbapenem-Resistant *Acinetobacter baumannii* Infections.
658 <https://journals.asm.org/doi/abs/10.1128/aac.01110-18>. Accessed 24 Jun 2021
- 659 92. Buehrle DJ, Shields RK, Clarke LG, et al (2017) Carbapenem-Resistant *Pseudomonas aeruginosa* Bacteremia: Risk
660 Factors for Mortality and Microbiologic Treatment Failure. *Antimicrob Agents Chemother* 61.:
661 <https://doi.org/10.1128/AAC.01243-16>
- 662 93. Meletis G, Exindari M, Vavatsi N, et al (2012) Mechanisms responsible for the emergence of carbapenem
663 resistance in *Pseudomonas aeruginosa*. *Hippokratia* 16:303–307
- 664 94. Markwart R, Willrich N, Haller S, et al (2019) The rise in vancomycin-resistant *Enterococcus faecium* in Germany:
665 data from the German Antimicrobial Resistance Surveillance (ARS). *Antimicrob Resist Infect Control* 8:147.
666 <https://doi.org/10.1186/s13756-019-0594-3>
- 667 95. Kafil HS, Asgharzadeh M (2014) Vancomycin-resistant enterococcus faecium and enterococcus faecalis isolated

from education hospital of iran. *Maedica* 9:323–327

96. Fridkin SK, Hageman JC, Morrison M, et al (2005) Methicillin-resistant *Staphylococcus aureus* disease in three communities. *N Engl J Med* 352:1436–1444. <https://doi.org/10.1056/NEJMoa043252>
97. Boucher HW, Corey GR (2008) Epidemiology of methicillin-resistant *Staphylococcus aureus*. *Clin Infect Dis* 46 Suppl 5:S344–9. <https://doi.org/10.1086/533590>
98. Brumfitt W, Hamilton-Miller J (1989) Methicillin-resistant *Staphylococcus aureus*. *N Engl J Med* 320:1188–1196. <https://doi.org/10.1056/NEJM198905043201806>
99. Lund BC, Ernst EJ, Klepser ME (1998) Strategies in the treatment of penicillin-resistant *Streptococcus pneumoniae*. *Am J Health Syst Pharm* 55:1987–1994. <https://doi.org/10.1093/ajhp/55.19.1987>
100. Jacobs MR (1999) Drug-resistant *Streptococcus pneumoniae*: rational antibiotic choices. *Am J Med* 106:19S–25S; discussion 48S–52S. [https://doi.org/10.1016/s0002-9343\(98\)00351-9](https://doi.org/10.1016/s0002-9343(98)00351-9)
101. Kariuki S, Gordon MA, Feasey N, Parry CM (2015) Antimicrobial resistance and management of invasive *Salmonella* disease. *Vaccine* 33 Suppl 3:C21–9. <https://doi.org/10.1016/j.vaccine.2015.03.102>
102. Cuypers WL, Jacobs J, Wong V, et al (2018) Fluoroquinolone resistance in *Salmonella*: insights by whole-genome sequencing. *Microb Genom* 4.: <https://doi.org/10.1099/mgen.0.000195>
103. Hao Chung The, Boinett C, Thanh DP, et al (2019) Dissecting the molecular evolution of fluoroquinolone-resistant *Shigella sonnei*. *Nat Commun* 10:1–13. <https://doi.org/10.1038/s41467-019-12823-0>
104. Zhang W-X, Chen H-Y, Tu L-H, et al (2019) Fluoroquinolone Resistance Mechanisms in *Shigella* Isolates in Shanghai, China, Between 2010 and 2015. *Microb Drug Resist* 25:212–218. <https://doi.org/10.1089/mdr.2018.0113>
105. Hennart M, Panunzi LG, Rodrigues C, et al (2020) Population genomics and antimicrobial resistance in *Corynebacterium diphtheriae*. *Genome Med* 12:107. <https://doi.org/10.1186/s13073-020-00805-7>
106. Marshall BM, Levy SB (2011) Food animals and antimicrobials: impacts on human health. *Clin Microbiol Rev* 24:718–733. <https://doi.org/10.1128/CMR.00002-11>

107. Aarestrup FM (2015) The livestock reservoir for antimicrobial resistance: a personal view on changing patterns of risks, effects of interventions and the way forward. *Philos Trans R Soc Lond B Biol Sci* 370:20140085. <https://doi.org/10.1098/rstb.2014.0085>
108. Holman DB, Yang W, Alexander TW (2019) Antibiotic treatment in feedlot cattle: a longitudinal study of the effect of oxytetracycline and tulathromycin on the fecal and nasopharyngeal microbiota. *Microbiome* 7:86. <https://doi.org/10.1186/s40168-019-0696-4>
109. Agga GE, Scott HM, Vinasco J, et al (2015) Effects of chlortetracycline and copper supplementation on the prevalence, distribution, and quantity of antimicrobial resistance genes in the fecal metagenome of weaned pigs. *Prev Vet Med* 119:179–189. <https://doi.org/10.1016/j.prevetmed.2015.02.008>
110. Hughes P, Heritage J, Others (2004) Antibiotic growth-promoters in food animals. *FAO Animal Production and Health Paper* 129–152
111. Aarestrup FM, Seyfarth AM, Emborg HD, et al (2001) 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. <https://doi.org/10.1128/AAC.45.7.2054-2059.2001>
112. Alvarez J, Lopez G, Muellner P, et al (2020) Identifying emerging trends in antimicrobial resistance using *Salmonella* surveillance data in poultry in Spain. *Transbound Emerg Dis* 67:250–262. <https://doi.org/10.1111/tbed.13346>
113. Gay E, Bour M, Cazeau G, et al (2019) Antimicrobial Usages and Antimicrobial Resistance in Commensal *Escherichia coli* From Veal Calves in France: Evolution During the Fattening Process. *Front Microbiol* 10:792. <https://doi.org/10.3389/fmicb.2019.00792>
114. Diaconu EL, Carfora V, Alba P, et al (2020) Novel IncFII plasmid harbouring blaNDM-4 in a carbapenem-resistant *Escherichia coli* of pig origin, Italy. *J Antimicrob Chemother* 75:3475–3479. <https://doi.org/10.1093/jac/dkaa374>
115. Irrgang A, Tausch SH, Pauly N, et al (2020) First Detection of GES-5-Producing *Escherichia coli* from Livestock-An Increasing Diversity of Carbapenemases Recognized from German Pig Production. *Microorganisms* 8.: <https://doi.org/10.3390/microorganisms8101593>

116. Morrison BJ, Rubin JE (2015) Carbapenemase producing bacteria in the food supply escaping detection. *PLoS One* 10:e0126717. <https://doi.org/10.1371/journal.pone.0126717>
117. Barza M (2002) Potential mechanisms of increased disease in humans from antimicrobial resistance in food animals. *Clin Infect Dis* 34 Suppl 3:S123–5. <https://doi.org/10.1086/340249>
118. Rinsky JL, Nadimpalli M, Wing S, et al (2013) Livestock-associated methicillin and multidrug resistant *Staphylococcus aureus* is present among industrial, not antibiotic-free livestock operation workers in North Carolina. *PLoS One* 8:e67641. <https://doi.org/10.1371/journal.pone.0067641>
119. Anders J, Bisha B (2020) High-Throughput Detection and Characterization of Antimicrobial Resistant *Enterococcus* sp. Isolates from GI Tracts of European Starlings Visiting Concentrated Animal Feeding Operations. *Foods* 9.: <https://doi.org/10.3390/foods9070890>
120. Islam S, Paul A, Talukder M, et al (2021) Migratory birds travelling to Bangladesh are potential carriers of multi-drug resistant *Enterococcus* spp., *Salmonella* spp., and *Vibrio* spp. *Saudi J Biol Sci*. <https://doi.org/10.1016/j.sjbs.2021.06.053>
121. Islam MS, Sobur MA, Rahman S, et al (2021) Detection of blaTEM, blaCTX-M, blaCMY, and blaSHV Genes Among Extended-Spectrum Beta-Lactamase-Producing *Escherichia coli* Isolated from Migratory Birds Travelling to Bangladesh. *Microb Ecol*. <https://doi.org/10.1007/s00248-021-01803-x>
122. Plaza-Rodríguez C, Alt K, Grobbel M, et al (2020) Wildlife as Sentinels of Antimicrobial Resistance in Germany? *Front Vet Sci* 7:627821. <https://doi.org/10.3389/fvets.2020.627821>
123. Atterby C, Börjesson S, Ny S, et al (2017) ESBL-producing *Escherichia coli* in Swedish gulls-A case of environmental pollution from humans? *PLoS One* 12:e0190380. <https://doi.org/10.1371/journal.pone.0190380>
124. Dolejska M, Papagiannitsis CC (2018) Plasmid-mediated resistance is going wild. *Plasmid* 99:99–111. <https://doi.org/10.1016/j.plasmid.2018.09.010>
125. Hernández J, González-Acuña D (2016) Anthropogenic antibiotic resistance genes mobilization to the polar regions. *Infect Ecol Epidemiol* 6:32112. <https://doi.org/10.3402/iee.v6.32112>
126. Dickinson AW, Power A, Hansen MG, et al (2019) Heavy metal pollution and co-selection for antibiotic

- resistance: A microbial palaeontology approach. *Environ Int* 132:105117.
<https://doi.org/10.1016/j.envint.2019.105117>
127. Baker-Austin C, Wright MS, Stepanauskas R, McArthur JV (2006) Co-selection of antibiotic and metal resistance. *Trends Microbiol* 14:176–182. <https://doi.org/10.1016/j.tim.2006.02.006>
128. Kraemer SA, Ramachandran A, Perron GG (2019) Antibiotic Pollution in the Environment: From Microbial Ecology to Public Policy. *Microorganisms* 7.: <https://doi.org/10.3390/microorganisms7060180>
129. Barancheshme F, Munir M (2017) Strategies to Combat Antibiotic Resistance in the Wastewater Treatment Plants. *Front Microbiol* 8:2603. <https://doi.org/10.3389/fmicb.2017.02603>
130. Rodríguez-Molina D, Mang P, Schmitt H, et al (2019) Do wastewater treatment plants increase antibiotic resistant bacteria or genes in the environment? Protocol for a systematic review. *Syst Rev* 8:304.
<https://doi.org/10.1186/s13643-019-1236-9>
131. Alexander J, Hembach N, Schwartz T (2020) Evaluation of antibiotic resistance dissemination by wastewater treatment plant effluents with different catchment areas in Germany. *Sci Rep* 10:8952.
<https://doi.org/10.1038/s41598-020-65635-4>
132. Osińska A, Korzeniewska E, Harnisz M, et al (2020) Small-scale wastewater treatment plants as a source of the dissemination of antibiotic resistance genes in the aquatic environment. *J Hazard Mater* 381:121221.
<https://doi.org/10.1016/j.jhazmat.2019.121221>
133. Bueno I, Verdugo C, Jimenez-Lopez O, et al (2020) Role of wastewater treatment plants on environmental abundance of Antimicrobial Resistance Genes in Chilean rivers. *Int J Hyg Environ Health* 223:56–64.
<https://doi.org/10.1016/j.ijheh.2019.10.006>
134. Allen HK, Donato J, Wang HH, et al (2010) Call of the wild: antibiotic resistance genes in natural environments. *Nat Rev Microbiol* 8:251–259. <https://doi.org/10.1038/nrmicro2312>
135. Scott LC, Lee N, Aw TG (2020) Antibiotic Resistance in Minimally Human-Impacted Environments. *Int J Environ Res Public Health* 17.: <https://doi.org/10.3390/ijerph17113939>
136. Armalytė J, Skerniškytė J, Bakienė E, et al (2019) Microbial Diversity and Antimicrobial Resistance Profile in

768 Microbiota From Soils of Conventional and Organic Farming Systems. *Front Microbiol* 10:892.
769 <https://doi.org/10.3389/fmicb.2019.00892>

770 137. Dantas G, Sommer MOA, Oluwasegun RD, Church GM (2008) Bacteria subsisting on antibiotics. *Science*
771 320:100–103. <https://doi.org/10.1126/science.1155157>

772 138. D’Costa VM, McGrann KM, Hughes DW, Wright GD (2006) Sampling the antibiotic resistome. *Science*
773 311:374–377. <https://doi.org/10.1126/science.1120800>

774 139. Wolters B, Jacquiod S, Sørensen SJ, et al (2018) Bulk soil and maize rhizosphere resistance genes, mobile
775 genetic elements and microbial communities are differently impacted by organic and inorganic fertilization. *FEMS*
776 *Microbiol Ecol* 94.: <https://doi.org/10.1093/femsec/fiy027>

777 140. Song M, Peng K, Jiang L, et al (2020) Alleviated Antibiotic-Resistant Genes in the Rhizosphere of Agricultural
778 Soils with Low Antibiotic Concentration. *J Agric Food Chem* 68:2457–2466.
779 <https://doi.org/10.1021/acs.jafc.9b06634>

780 141. Kittinger C, Kirschner A, Lipp M, et al (2017) Antibiotic Resistance of *Acinetobacter* spp. Isolates from the
781 River Danube: Susceptibility Stays High. *Int J Environ Res Public Health* 15.:
782 <https://doi.org/10.3390/ijerph15010052>

783 142. Paschoal RP, Campana EH, Corrêa LL, et al (2017) Concentration and Variety of Carbapenemase Producers in
784 Recreational Coastal Waters Showing Distinct Levels of Pollution. *Antimicrob Agents Chemother* 61.:
785 <https://doi.org/10.1128/AAC.01963-17>

786 143. Mahon BM, Brehony C, Cahill N, et al (2019) Detection of OXA-48-like-producing Enterobacterales in Irish
787 recreational water. *Sci Total Environ* 690:1–6. <https://doi.org/10.1016/j.scitotenv.2019.06.480>

788 144. Surette MD, Wright GD (2017) Lessons from the Environmental Antibiotic Resistome. *Annu Rev Microbiol*
789 71:309–329. <https://doi.org/10.1146/annurev-micro-090816-093420>

790 145. Flach C-F, Johnning A, Nilsson I, et al (2015) Isolation of novel IncA/C and IncN fluoroquinolone resistance
791 plasmids from an antibiotic-polluted lake. *J Antimicrob Chemother* 70:2709–2717.
792 <https://doi.org/10.1093/jac/dkv167>

146. Kristiansson E, Fick J, Janzon A, et al (2011) Pyrosequencing of antibiotic-contaminated river sediments reveals high levels of resistance and gene transfer elements. *PLoS One* 6:e17038. <https://doi.org/10.1371/journal.pone.0017038>
147. McCann CM, Christgen B, Roberts JA, et al (2019) Understanding drivers of antibiotic resistance genes in High Arctic soil ecosystems. *Environ Int* 125:497–504. <https://doi.org/10.1016/j.envint.2019.01.034>
148. Dancer SJ, Shears P, Platt DJ (1997) Isolation and characterization of coliforms from glacial ice and water in Canada's High Arctic. *J Appl Microbiol* 82:597–609. <https://doi.org/10.1111/j.1365-2672.1997.tb03590.x>
149. Nadeem SF, Gohar UF, Tahir SF, et al (2020) Antimicrobial resistance: more than 70 years of war between humans and bacteria. *Crit Rev Microbiol* 46:578–599. <https://doi.org/10.1080/1040841X.2020.1813687>
150. Gibson MK, Forsberg KJ, Dantas G (2015) Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J* 9:207–216. <https://doi.org/10.1038/ismej.2014.106>
151. Munck C, Albertsen M, Telke A, et al (2015) Limited dissemination of the wastewater treatment plant core resistome. *Nat Commun* 6:8452. <https://doi.org/10.1038/ncomms9452>
152. Ngoi ST, Chong CW, Ponnampalavanar SSLS, et al (2021) Genetic mechanisms and correlated risk factors of antimicrobial-resistant ESKAPEE pathogens isolated in a tertiary hospital in Malaysia. *Antimicrob Resist Infect Control* 10:70. <https://doi.org/10.1186/s13756-021-00936-5>
153. Van Den Broek IVF, Van Cleef B, Haenen A, et al (2009) Methicillin-resistant *Staphylococcus aureus* in people living and working in pig farms. *Epidemiology & Infection* 137:700–708
154. Lewis HC, Mølbak K, Reese C, et al (2008) Pigs as source of methicillin-resistant *Staphylococcus aureus* CC398 infections in humans, Denmark. *Emerg Infect Dis* 14:1383–1389. <https://doi.org/10.3201/eid1409.071576>
155. Loncaric I, Cabal Rosel A, Szostak MP, et al (2020) Broad-Spectrum Cephalosporin-Resistant *Klebsiella* spp. Isolated from Diseased Horses in Austria. *Animals (Basel)* 10.: <https://doi.org/10.3390/ani10020332>
156. Mughini-Gras L, Dorado-García A, van Duijkeren E, et al (2019) Attributable sources of community-acquired carriage of *Escherichia coli* containing β -lactam antibiotic resistance genes: a population-based modelling study. *Lancet Planet Health* 3:e357–e369. [https://doi.org/10.1016/S2542-5196\(19\)30130-5](https://doi.org/10.1016/S2542-5196(19)30130-5)

157. Dahl LG, Joensen KG, Østerlund MT, et al (2021) Prediction of antimicrobial resistance in clinical *Campylobacter jejuni* isolates from whole-genome sequencing data. *Eur J Clin Microbiol Infect Dis* 40:673–682. <https://doi.org/10.1007/s10096-020-04043-y>
158. Mossong J, Mughini-Gras L, Penny C, et al (2016) Human *Campylobacteriosis* in Luxembourg, 2010–2013: A Case-Control Study Combined with Multilocus Sequence Typing for Source Attribution and Risk Factor Analysis. *Sci Rep* 6:20939. <https://doi.org/10.1038/srep20939>
159. Winokur PL, Brueggemann A, DeSalvo DL, et al (2000) Animal and human multidrug-resistant, cephalosporin-resistant salmonella isolates expressing a plasmid-mediated CMY-2 AmpC beta-lactamase. *Antimicrob Agents Chemother* 44:2777–2783. <https://doi.org/10.1128/AAC.44.10.2777-2783.2000>
160. Genomic Investigation of the Emergence of Invasive Multidrug-Resistant *Salmonella enterica* Serovar Dublin in Humans and Animals in Canada. <https://journals.asm.org/doi/abs/10.1128/aac.00108-19>. Accessed 2 Jul 2021
161. Zhang L, Fu Y, Xiong Z, et al (2018) Highly Prevalent Multidrug-Resistant *Salmonella* From Chicken and Pork Meat at Retail Markets in Guangdong, China. *Front Microbiol* 9:2104. <https://doi.org/10.3389/fmicb.2018.02104>
162. Dionisi AM, Lucarelli C, Benedetti I, et al (2011) Molecular characterisation of multidrug-resistant *Salmonella enterica* serotype Infantis from humans, animals and the environment in Italy. *Int J Antimicrob Agents* 38:384–389. <https://doi.org/10.1016/j.ijantimicag.2011.07.001>
163. Forslund K, Sunagawa S, Coelho LP, Bork P (2014) Metagenomic insights into the human gut resistome and the forces that shape it. *Bioessays* 36:316–329. <https://doi.org/10.1002/bies.201300143>
164. Busi SB, de Nies L, Habier J, et al (2021) Persistence of birth mode-dependent effects on gut microbiome composition, immune system stimulation and antimicrobial resistance during the first year of life. *ISME Communications* 1:1–12. <https://doi.org/10.1038/s43705-021-00003-5>
165. Casaburi G, Duar RM, Brown H, et al (2021) Metagenomic insights of the infant microbiome community structure and function across multiple sites in the United States. *Sci Rep* 11:1–12. <https://doi.org/10.1038/s41598-020-80583-9>
166. Wampach L, Heintz-Buschart A, Fritz JV, et al (2018) Birth mode is associated with earliest strain-conferred

- 844 gut microbiome functions and immunostimulatory potential. *Nat Commun* 9:5091. [https://doi.org/10.1038/s41467-](https://doi.org/10.1038/s41467-018-07631-x)
845 018-07631-x
- 846 167. Gaeta NC, Bean E, Miles AM, et al (2020) A Cross-Sectional Study of Dairy Cattle Metagenomes Reveals
847 Increased Antimicrobial Resistance in Animals Farmed in a Heavy Metal Contaminated Environment. *Front*
848 *Microbiol* 11:590325. <https://doi.org/10.3389/fmicb.2020.590325>
- 849 168. Skarżyńska M, Leekitcharoenphon P, Hendriksen RS, et al (2020) A metagenomic glimpse into the gut of wild
850 and domestic animals: Quantification of antimicrobial resistance and more. *PLoS One* 15:e0242987.
851 <https://doi.org/10.1371/journal.pone.0242987>
- 852 169. Duarte ASR, Röder T, Van Gompel L, et al (2020) Metagenomics-Based Approach to Source-Attribution of
853 Antimicrobial Resistance Determinants - Identification of Reservoir Resistome Signatures. *Front Microbiol*
854 11:601407. <https://doi.org/10.3389/fmicb.2020.601407>
- 855 170. Van Gompel L, Luiken REC, Hansen RB, et al (2020) Description and determinants of the faecal resistome and
856 microbiome of farmers and slaughterhouse workers: A metagenome-wide cross-sectional study. *Environ Int*
857 143:105939. <https://doi.org/10.1016/j.envint.2020.105939>
- 858 171. Ma L, Xia Y, Li B, et al (2016) Metagenomic Assembly Reveals Hosts of Antibiotic Resistance Genes and the
859 Shared Resistome in Pig, Chicken, and Human Feces. *Environ Sci Technol* 50:420–427.
860 <https://doi.org/10.1021/acs.est.5b03522>
- 861 172. Wang Y, Hu Y, Liu F, et al (2020) Integrated metagenomic and metatranscriptomic profiling reveals
862 differentially expressed resistomes in human, chicken, and pig gut microbiomes. *Environ Int* 138:105649.
863 <https://doi.org/10.1016/j.envint.2020.105649>
- 864 173. Noyes NR, Yang X, Linke LM, et al (2016) Characterization of the resistome in manure, soil and wastewater
865 from dairy and beef production systems. *Sci Rep* 6:24645. <https://doi.org/10.1038/srep24645>
- 866 174. Sukhum KV, Vargas RC, Boolchandani M, et al (2021) Manure Microbial Communities and Resistance
867 Profiles Reconfigure after Transition to Manure Pits and Differ from Those in Fertilized Field Soil. *MBio* 12.:
868 <https://doi.org/10.1128/mBio.00798-21>
- 869 175. Smith SD, Colgan P, Yang F, et al (2019) Investigating the dispersal of antibiotic resistance associated genes

from manure application to soil and drainage waters in simulated agricultural farmland systems. PLoS One
14:e0222470. <https://doi.org/10.1371/journal.pone.0222470>

176. Qian X, Gunturu S, Guo J, et al (2021) Metagenomic analysis reveals the shared and distinct features of the soil
resistome across tundra, temperate prairie, and tropical ecosystems. Microbiome 9:108.
<https://doi.org/10.1186/s40168-021-01047-4>

177. Ju F, Li B, Ma L, et al (2016) Antibiotic resistance genes and human bacterial pathogens: Co-occurrence,
removal, and enrichment in municipal sewage sludge digesters. Water Res 91:1–10.
<https://doi.org/10.1016/j.watres.2015.11.071>

178. Hendriksen RS, Munk P, Njage P, et al (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>

179. Ma L, Li B, Jiang X-T, et al (2017) Catalogue of antibiotic resistome and host-tracking in drinking water
deciphered by a large scale survey. Microbiome 5:154. <https://doi.org/10.1186/s40168-017-0369-0>

180. Bai Y, Ruan X, Xie X, Yan Z (2019) Antibiotic resistome profile based on metagenomics in raw surface
drinking water source and the influence of environmental factor: A case study in Huaihe River Basin, China.
Environ Pollut 248:438–447. <https://doi.org/10.1016/j.envpol.2019.02.057>

181. Pal C, Bengtsson-Palme J, Kristiansson E, Larsson DGJ (2016) The structure and diversity of human, animal
and environmental resistomes. Microbiome 4:54. <https://doi.org/10.1186/s40168-016-0199-5>

182. Li B, Yang Y, Ma L, et al (2015) Metagenomic and network analysis reveal wide distribution and co-
occurrence of environmental antibiotic resistance genes. ISME J 9:2490–2502.
<https://doi.org/10.1038/ismej.2015.59>

183. Muloi D, Ward MJ, Pedersen AB, et al (2018) Are Food Animals Responsible for Transfer of Antimicrobial-
Resistant Escherichia coli or Their Resistance Determinants to Human Populations? A Systematic Review.
Foodborne Pathog Dis 15:467–474. <https://doi.org/10.1089/fpd.2017.2411>

184. Mather AE, Reid SWJ, Maskell DJ, et al (2013) Distinguishable epidemics of multidrug-resistant Salmonella
Typhimurium DT104 in different hosts. Science 341:1514–1517. <https://doi.org/10.1126/science.1240578>

- 895 185. Al-Shayeb B, Schoelmerich MC, West-Roberts J, et al (2021) Borgs are giant extrachromosomal elements with
896 the potential to augment methane oxidation. bioRxiv 2021.07.10.451761
- 897 186. McEwen SA, Collignon PJ (2018) Antimicrobial Resistance: a One Health Perspective. Microbiol Spectr 6.:
898 <https://doi.org/10.1128/microbiolspec.ARBA-0009-2017>
- 899 187. Collignon P (2013) The importance of a One Health approach to preventing the development and spread of
900 antibiotic resistance. Curr Top Microbiol Immunol 366:19–36. https://doi.org/10.1007/82_2012_224
- 901 188. Aarestrup FM, Wegener HC, Collignon P (2008) Resistance in bacteria of the food chain: epidemiology and
902 control strategies. Expert Rev Anti Infect Ther 6:733–750. <https://doi.org/10.1586/14787210.6.5.733>
- 903 189. Organization WH, Others (2004) Second Joint FAO/OIE/WHO Expert Workshop on Non-Human
904 Antimicrobial Usage and Antimicrobial Resistance: Management options: 15-18 March 2004, Oslo, Norway.
905 World Health Organization

906

Appendix A.2


PathoFact: a pipeline for the prediction of
virulence factors and antimicrobial
resistance genes in metagenomic data

SOFTWARE ARTICLE

Open Access



PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data

Laura de Nies¹, Sara Lopes¹, Susheel Bhanu Busi¹, Valentina Galata¹, Anna Heintz-Buschart^{1,2,3}, Cedric Christian Laczny¹, Patrick May⁴ and Paul Wilmes^{1*} 

Abstract

Background: Pathogenic microorganisms cause disease by invading, colonizing, and damaging their host. Virulence factors including bacterial toxins contribute to pathogenicity. Additionally, antimicrobial resistance genes allow pathogens to evade otherwise curative treatments. To understand causal relationships between microbiome compositions, functioning, and disease, it is essential to identify virulence factors and antimicrobial resistance genes in situ. At present, there is a clear lack of computational approaches to simultaneously identify these factors in metagenomic datasets.

Results: Here, we present PathoFact, a tool for the contextualized prediction of virulence factors, bacterial toxins, and antimicrobial resistance genes with high accuracy (0.921, 0.832 and 0.979, respectively) and specificity (0.957, 0.989 and 0.994). We evaluate the performance of PathoFact on simulated metagenomic datasets and perform a comparison to two other general workflows for the analysis of metagenomic data. PathoFact outperforms all existing workflows in predicting virulence factors and toxin genes. It performs comparably to one pipeline regarding the prediction of antimicrobial resistance while outperforming the others. We further demonstrate the performance of PathoFact on three publicly available case-control metagenomic datasets representing an actual infection as well as chronic diseases in which either pathogenic potential or bacterial toxins are hypothesized to play a role. In each case, we identify virulence factors and AMR genes which differentiated between the case and control groups, thereby revealing novel gene associations with the studied diseases.

Conclusion: PathoFact is an easy-to-use, modular, and reproducible pipeline for the identification of virulence factors, bacterial toxins, and antimicrobial resistance genes in metagenomic data. Additionally, our tool combines the prediction of these pathogenicity factors with the identification of mobile genetic elements. This provides further depth to the analysis by considering the genomic context of the pertinent genes. Furthermore, PathoFact's modules for virulence factors, toxins, and antimicrobial resistance genes can be applied independently, thereby making it a flexible and versatile tool. PathoFact, its models, and databases are freely available at <https://pathofact.lcsb.uni.lu>.

Keywords: Virulence factors, Bacterial toxins, Antimicrobial resistance, Mobile genetic elements, Metagenomics, Microbiome, Bioinformatics

* Correspondence: paul.wilmes@uni.lu

¹Systems Ecology Research Group, Luxembourg Centre for Systems Biomedicine, Esch-sur-Alzette, Luxembourg

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Most of the microorganisms constituting the human microbiome are commensals [1]. They contribute essential functions to the human host and contribute to its physiological development. In contrast, pathogenic microorganisms including bacteria, viruses, fungi, and protozoa cause disease by invading, colonizing, and damaging the host. Virulence factors, including bacterial toxins among others, contribute to this pathogenicity by enhancing not only the infectivity of pathogenic bacteria but also by exacerbating antimicrobial resistance which in turn restricts treatment options [1].

Virulence factors enable pathogenic microorganisms to colonize host niches ultimately resulting in tissue damage as well as local and systemic inflammation. These factors are important for pathogens to establish an infection and span a wide range, thus contributing both directly and indirectly to disease processes [2]. These virulence traits include cell-surface structures, secretion machineries, siderophores, regulators, etc. [3, 4]. However, of all virulence factors employed by pathogens, bacterial toxins often have a crucial function in the pathogenesis of infectious diseases [5]. Different types of bacterial toxins have evolved over time to counteract human defenses. These bacterial toxins can be coarsely categorized into two groups: the cell-associated endotoxins and the extracellular diffusible exotoxins. Exotoxins are typically polypeptides and proteins that act to stimulate a variety of host responses either through direct action with cell receptors or via enzymatic modulation [5, 6].

Partly through the utilization of these virulence factors, and toxins in particular, pathogenic microorganisms have been a major cause of infectious diseases including in the context of viral co-infections [1]. The development and medical use of antibiotics has limited the development and spread of these pathogens by providing an effective treatment for bacterial infections. However, the over- and mis-use of antibiotics has resulted in a global increase in antimicrobial resistance (AMR) which now threatens human health through the emergence and spread of multidrug resistant bacteria [1, 7]. As a result, many pathogenic bacteria have now acquired resistance against the main classes of antibiotics which has led to a dramatic rise in untreatable infections, resulting in the emergence of so-called “superbugs” [8]. Consequently, AMR is an urgent and growing threat to public health with an estimated number of deaths exceeding ten million annually by 2050 [9, 10].

The acquisition of antimicrobial resistance genes (ARGs) is not restricted to a single strain or species of bacteria. While commensal bacteria provide a source of ARGs, antimicrobial resistance can be transferred to pathogenic species through horizontal gene transfer, e.g., conjugation or transduction [11–13]. Therefore, to

understand the emergence and spread of ARGs, it is necessary to monitor microbial communities in situ. Metagenomic sequencing, in this context, represents a pertinent technique for in situ studies as it provides less biased views of the genomic complements of individual microbial populations compared to amplicon-based methods [14, 15].

Pathogenic microorganisms have modified and adapted their virulence to host defense systems over millions of years. Similarly, AMR is thought to have evolved over extensive periods of time in bacteria, indicating that it is an ancient phenomenon [16]. However, with an increase in selective pressure through the use of antibiotics, an excessive increase in the spread and evolution of AMR has been observed in the last 50 years. Yet, despite differences in evolutionary paths, virulence factors and AMR share common characteristics. Most importantly, virulence factors and AMR are necessary for pathogenic bacteria to adapt to, and survive in, competitive microbial environments [7]. Additionally, both virulence and resistance mechanisms are frequently transferred between bacteria by horizontal gene transfer [13]. Furthermore, both processes make use of similar systems (i.e., cell wall alterations, efflux pumps, two-component systems and porins) that activate or repress the expression of various genes [17–19]. Therefore, although AMR in itself is not a virulence factor, in environments with selective antibiotic pressure, opportunistic pathogens are able to colonize through acquisition or presence of AMR [1].

Considering the burden of bacterial infections in which virulence factors and ARGs play crucial roles, it is important to be able to identify these in microbial communities. The advent of high-throughput DNA sequencing provides a powerful means to profile the full complement of DNA derived from genomic extracts obtained from a wide range of environments [20]. However, currently there is a lack of automated pipelines to simultaneously identify these different factors in metagenomic datasets. Various tools exist for the prediction of ARGs themselves, such as DeepARG [20], RGI [21], ResFinder [22], and ARGsOAP [23], with a very few prediction tools for virulence factors existing, such as MP3 [24] and VirulentPred [25]. Most of the latter tools are based on outdated databases of virulence factors which have since been expanded greatly. Moreover, there is a lack of recent bioinformatics tools for the prediction of bacterial toxin genes in particular. Furthermore, although various AMR prediction tools exist, these primarily focus on the prediction of genes without considering their location, i.e., these tools do not differentiate between localization on mobile genetic elements (MGEs) or on bacterial genomes. Since MGEs are the main mechanism by which ARGs are transmitted, it is crucial to identify the relationship between ARGs and MGEs. Outside of these prediction tools, it is common

practice to use standard homology search algorithms against specific databases. However, such practices require several intermediate steps which may vary from lab to lab. Additionally, using these methods is restrictive in the sense that only a single database can be searched at a time.

Here, we present PathoFact, a pipeline for the simultaneous prediction of virulence factors, bacterial toxins in particular, and ARGs. Our tool furthermore contextualizes these with respect to their localization on MGEs. Moreover, PathoFact aggregates the information obtained via different prediction tools and databases into a single output, allowing both novices and experts in bioinformatics alike to parse information as needed. PathoFact thus provides a unified perspective on pathogenic mechanisms. We provide evaluation results on our tool's sensitivity, specificity, and accuracy, and demonstrate PathoFact's versatility using both a simulated metagenomic dataset and public case-control metagenomic datasets for Parkinson's disease, psoriasis, and *Clostridioides difficile* infection. Using the simulated metagenomic dataset, we further perform a comparison of PathoFact to other metagenomic characterization workflows, namely MOCAT2 [26] and HUMANN3 [27].

Implementation

PathoFact architecture

PathoFact is a command-line tool for UNIX-based systems that integrates three distinct workflows for the prediction of (i) virulence factors, (ii) bacterial toxins, and (iii) antimicrobial resistance genes from metagenomic data (Fig. 1a). Each workflow can be applied individually or in combination with the other workflows. Our tool is written in Python (version 3.6) and uses the Snakemake (version 5.5.4) workflow management software [28]. This implementation offers several advantages, including workflow assembly, parallelism, and the ability to resume processing following an interruption. Each step of the pipeline is implemented as a rule in the Snakemake framework specifying the input needed and the output files generated. We use conda (version 4.7) environments wherever possible thus reducing the need for explicit installation of software dependencies. Moreover, the use of conda environments makes it possible to incorporate prediction tools dependent on older Python versions incompatible with version 5.5 of Snakemake. As such, Python, Snakemake, and (mini)conda (version 4.7) [29] installations are required. PathoFact is open-source and freely available at <https://pathofact.lcsb.uni.lu>.

The input to the PathoFact pipeline consists of an assembly FASTA file containing nucleotide sequences of the contigs. PathoFact subsequently predicts the ORFs using Prodigal (version 2.6.3) for the prediction of virulence factors, toxins, and antimicrobial resistance genes.

The MGEs are predicted from the initial assembly file, and a mapping file is generated by PathoFact which aggregates all the results. PathoFact aggregates the information obtained from the different sub modules into both module-specific reports as well as a complete final report. The reports describe all virulence factors, bacterial toxins, and antimicrobial resistance genes identified from the input as well as their assigned confidence level (virulence factors/bacterial toxins), their resistance mechanisms (AMR), and their corresponding localization on MGEs.

Workflow for the prediction of virulence factors

For the prediction of virulence factors, we created a prediction tool consisting of two parts: (i) a database consisting of virulence factor HMM profiles (HMMER3 v3.2.1) [30] and (ii) a random forest model. Hits against the virulence factor HMM database are then combined with the classification of the random forest model to result in the final prediction (Fig. 1b). The development of the tool was inspired by the MP3 software tool for the prediction of virulence factors which has not received an update since 2014 and was thus outdated [24]. In addition, PathoFact combines these annotations with the prediction of signal peptides by SignalP (v5.0) [31] to distinguish between secreted and non-secreted virulence factors.

Dataset for the prediction of virulence factors

A dataset, consisting of both a positive and negative subset, was constructed for the training of the virulence factor prediction tool. The positive subset consisted of known virulence factor sequences retrieved from the Virulence Factors Database (8945 sequences) (VFDB) [3]. All sequences were obtained from the VFDB core dataset containing (translated) gene sequences associated with experimentally verified virulence factors. The negative subset of the training set consisted of protein sequences that were retrieved from the Database of Essential Genes (DEG) (7995 sequences) [32] and which were known not to be virulence factors. For both subsets, all sequences were clustered with CD-HIT [33], and sequences with a 90% sequence identity were collapsed to prevent redundancy within the subsets. This 90% cut-off is routinely used to reduce redundancy in similar protein datasets, improving efficiency without foregoing specificity given the large metagenomic database sizes [34, 35]. The resulting training set was used for (i) the implementation of the HMM profiles and (ii) the training of the random forest model.

Construction of the virulence factor HMM database

For the construction of the virulence HMM database, HMM profiles were annotated for the training set using

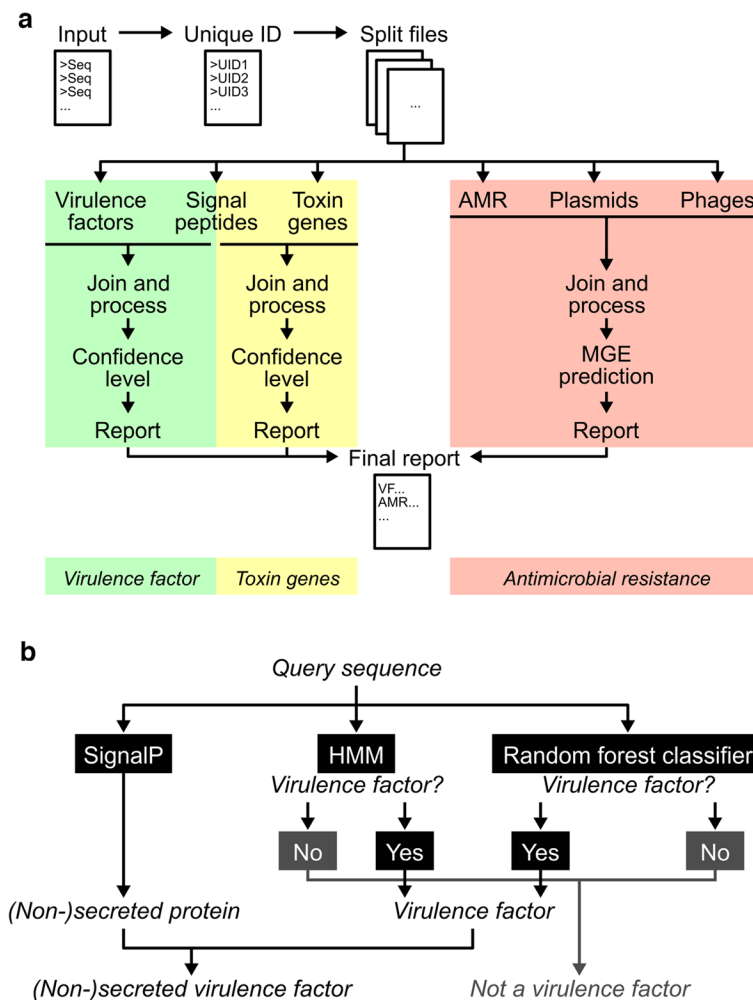


Fig. 1 The PathoFact pipeline. **a** Framework of the PathoFact pipeline. The pipeline consists of three different modules related to (i) virulence factors, incl. (ii) bacterial toxins, and (iii) antimicrobial resistance genes. SignalP is incorporated for the prediction of secreted toxins and virulence factors. All modules can either be run independently or jointly. **b** Classification framework for the prediction of virulence factors. The prediction of virulence factors depends on two different aspects: (i) a HMM domain database, (ii) a random forest classifier. Sequences predicted positive from both are classified as virulence factors. The incorporation of SignalP in the framework allows integration of information regarding the likely secretion of the virulence factors

HMMER3 (version 3.2.1) against multiple pre-compiled and in-house annotation databases [36]: PFAM-A [37], TIGR [38], KEGG [39], MetaCyc [40], and Swissprot [41]. The best hit in each HMM set was assigned to each gene in the training set if the HMM score was higher than the binary logarithm of the number of target genes, in accordance with the recommendations in the HMMer manual. HMM profiles were subsequently retrieved and the databases were concatenated to form the virulence HMM database. Binary compressed data files were constructed with the *hmmcompress* (HMMER3 v3.2.1) [30]. For the prediction of virulence factors by the virulence HMM database, identified HMM profiles are separated by those matching to the positive or negative subset of

the training set, as well as HMM profiles ambiguous for both positive and negative subset.

Machine learning model for the prediction of virulence factors

In addition to the virulence HMM database, we created a random forest model [42]. A random forest model operates from decision trees and output classification of the individual trees while correcting for overfitting of the training set. While overfitting, in which models perform highly on the training set but poorly on the test set, is a common problem in machine learning, a random forest model corrects for overfitting by continuously creating trees on random subsets. This does not

mean that random forest classifiers are not capable of overfitting. However, they are less sensitive to variance, and effects of overfitting are therefore rarely observed [43]. For training of the random forest model, the following five features of the sequences were selected and implemented: amino acid composition (AAC), dipeptide composition (DPC), composition (CTDC), transition (CTDT), and distribution (CTDD) [44]. A feature matrix was built with rows corresponding to the sequence composition of the features. The random forest model was implemented using pandas (v 0.25.0) [45], Numpy (v 1.17.0) [46], and scikit-learn (v0.21.3) [47] and consisted of 1600 trees with a maximum depth of 340.

Workflow for the prediction of toxin genes

For the prediction of toxin genes, a workflow consisting of a toxin HMM database combined with SignalP version 5.0 [31] was developed. The toxin HMM database consists of bacterial toxin domains to identify toxin-related domains in the query sequences. Using the *hmmsearch* function of the HMMER3 (v3.2.1) program [30], the input query sequences are searched against the collection of profiles present in the toxin HMM database. In addition, analyses are combined with SignalP [31] to differentiate between secreted and non-secreted toxins.

Construction of the toxin HMM database

For the toxin HMM database, an HMM model based on a training set of known toxins was developed and implemented. The training set was compiled from the Toxin and Toxin Target Database (T3DB) [48] and the training set derived from the DBETH prediction tool [5]. Protein sequences from within the training set with a similarity greater than 90% were clustered and collapsed with CD-HIT-2D to reduce redundancy [33]. The corresponding toxin HMM profiles were identified from the same five HMM databases as used for the virulence factors (see above). The datasets were extended with HMM profiles already annotated as bacterial toxin domains in the PFAM, TIGR, KEGG, MetaCyc, and Swissprot databases. Finally, in order to have a short description of all HMM profiles present in the toxin HMM database, a toxin library was created. This lists (i) all HMM profiles, (ii) their names, (iii) their alternative names, and (iv) the original database from which the HMM profile was derived.

Workflow for the prediction of antimicrobial resistance genes

For the prediction of ARGs, the workflow is separated into two parts: (i) the prediction of ARGs and (ii) the prediction of MGEs. For the prediction of ARGs, the tools DeepARG (v1.0.1) [20] and RGI (v5.1.0) [21] are

used. DeepARG uses a deep learning approach that improves classification accuracy while at the same time reducing false negatives. It offers a powerful approach for metagenomic profiling of ARGs as it expands on the available databases for ARGs by combining the widely used CARD [49], ARDB [50], and UNIPROT [51] databases. Additionally, RGI [21] is included which is able to identify mutation-driven AMR within genes, allowing for a strain-resolved profiling of AMR genes.

MGEs: plasmids and phages

The prediction of MGEs is split into two parts focusing on the prediction of (i) plasmids and (ii) phages. For the prediction of plasmids, PlasFlow (v1.1) [52] is used, while for the prediction of phages VirSorter (v1.0.6) [53] and DeepVirFinder (v1.0) [54] were incorporated. All three tools were selected because of their performance compared to other, similar tools [52–54]. The predictions of these different tools are merged with the prediction of ARGs to provide localization information of the resistance genes to either MGEs or genomes. Considering the different predictions of MGEs, the final classification includes plasmid, phage, genome, unclassified, and ambiguous when localization predictions contradict each other, for example predicted to be both phage and plasmid.

Evaluation of the PathoFact pipeline

To evaluate the performance of PathoFact, validations were conducted for the prediction of toxins, for virulence factors, and for ARGs. The prediction quality was evaluated by sensitivity, specificity, and accuracy criteria as defined below.

$$\begin{aligned} \text{Sensitivity} &= \frac{tp}{tp + fn} & \text{Specificity} &= \frac{tn}{tn + fp} \\ &= \frac{tp}{tp + fn} & \text{Accuracy} &= \frac{tp + tn}{tp + fn + tn + fp} \end{aligned}$$

where **tp** represents true positives (i.e., virulence factors (incl. bacterial toxins) or AMR gene is predicted correctly), **tn** (i.e., a gene is correctly predicted not to be a virulence factor, toxin genes, or AMR gene), **fp** false positive (i.e., a gene incorrectly identified as a virulence factor, toxin genes or AMR gene), and **fn** false negatives (i.e., a virulence factor, toxin genes or AMR gene is incorrectly identified as non-pathogenic). We evaluated the sequence similarities between the training and validation (test set) datasets after removing the sequences from the validation set with 90% identity to the training set sequences using sourmash [55] (Additional File 1: Figure S1).

Validation of virulence factors

A validation dataset was constructed to assess the performance of the prediction of virulence factors. Analogous to the training set, the validation set consisted of a positive subset of 2639 sequences (VFDB database) and a negative subset of 2628 (DEG database) sequences. Importantly, the sequences in the validation dataset were removed from the training set to avoid overfitting. The test set for virulence predictions was used to run both the standalone MP3 (v1.0) tool and our newly generated tool for prediction of virulence factors. For MP3, the standard advised parameters were used: set on metagenomic protein fragments, a minimum length of 90 bases and a threshold value of 0.2 for the svm module [24].

Validation of toxin genes

For the validation of toxin genes, a validation dataset containing both positive and negative subsets was constructed. The positive subset was constructed from sequences in the EMBL-EBI database annotated as bacterial toxins. The results were limited to protein sequences described in the UniProtDB. Further filtering of the protein sequences removed sequences with uncertain predictions (i.e., hypothetical, probable). To limit redundancy within the dataset, sequences were clustered in terms of similarity by using a 90% sequence identity cutoff. Furthermore, to limit redundancy between the validation and the training set, sequences with a similarity of greater than 90% were discarded. The remaining 202 positive sequences were combined with 202 random-selected sequences from the negative dataset, consisting of housekeeping genes representing the validation dataset.

Validation of AMR prediction

For the prediction of AMR genes, both the DeepARG and RGI prediction tools were used. DeepARG has proven to be more accurate than most AMR prediction tools with a great reduction in false negatives [20], while RGI is capable to annotate SNPs contributing to AMR. For further validation, before inclusion in the pipeline, the prediction tools were tested using the NCBI's resistance gene database (5265 sequences) [56]. This positive subset was combined with a negative subset (consisting of sequences retrieved from the Database of Essential Genes) of equal size. For DeepARG default settings were applied, while parameters for model were set to **LS** and type was set to **prot**. Similar to DeepARG, default settings of RGI were applied while input-type was set to **protein**.

Data analysis and data availability of publicly available datasets

Metagenomic sequences for the publicly case-control metagenomic datasets were obtained from the European

Bioinformatics Institute-Sequence Read Archive database, with accession numbers PRJNA297269 (Milani et al. [57]), PRJNA281366 (Tett et al. [58]), and ERP019674 (Bedarf et al. [59]). Information on the analyzed samples per study can be found in Additional File 1: Table S1. Metagenomic reads were processed and assembled using IMP (v2) [60]. The resulting FASTA files containing the assembled contigs and genes were used as input for PathoFact. For analyses of the predictions, FeatureCounts (v1.6.4) [61] was used to extract the number of reads per functional category. Thereafter, the relative abundance of the toxin genes was calculated using the Rnum_Gi method described by Hu et al [62]. Additionally, the DESeq2 (v1.24) [63] package was used to analyze the differential abundance of virulence factors, toxins, and AMR genes.

Data analysis and data availability of a simulated dataset

To evaluate the performance of PathoFact compared to other metagenome characterization workflows, a high-complexity stimulated dataset consisting of 5 time series samples with 596 genomes and 478 circular elements was obtained from CAMI [64]. As with the case-control metagenomic dataset reads were processed and assembled using IMP (v2), after which the dataset was run through PathoFact. In addition, both MOCAT2 and HUMAnN3 were run on the stimulated metagenomic dataset using default settings of both workflows. Further data analysis was performed as described for the case-control datasets.

Results and discussion

Benchmarking

The PathoFact pipeline has an in-built multi-threading option to improve computational efficiency. In fact, certain tools, e.g., DeepVirFinder, are memory intensive and may require additional resources. Table 1 corresponds to the runtime of a metagenomic dataset (363, 933 metagenomic sequences) with differing numbers of threads. A minimum usage of 8 threads, in this case corresponding to 28 GB/thread, is advised for running the pipeline. Additionally, for the installation of PathoFact, an initial storage of 6.3 GB is required.

Validation of the PathoFact pipeline

For the prediction of virulence factors, the prediction tool consists of two parts: a virulence factor HMM

Table 1 PathoFact runtimes with different threads/computational resources

Threads	Memory	Running time
8	224 GB	25 h 19 min
16	448 GB	15 h 58 min

database and a random forest classifier. The random forest classifier's out-of-bag (OOB) error value reported an accuracy of 0.822. To improve performance for virulence prediction, the random forest model was combined with the HMM database which resulted in an overall sensitivity of 0.886, specificity of 0.957, and an accuracy of 0.921 (Table 2). Additionally, we compared our tool to the MP3 tool for the prediction of virulence factors (Additional File 1: Table S2). PathoFact scored overall higher than MP3 which scored 0.125, 0.992, and 0.558, respectively. In addition to the prediction of virulence factors, for the prediction of bacterial toxins, an overall sensitivity of 0.777, specificity of 0.989, and accuracy of 0.832 were obtained. Finally, for the prediction of ARGs, the sensitivity, specificity, and accuracy of both DeepARG and RGI were determined at 0.720, 0.996, 0.858 and 0.920, 0.997, 0.958, respectively. A combined approach merging the use of both tools resulted in the highest scores with an overall sensitivity of 0.963, specificity of 0.994, and accuracy of 0.979 for the prediction of AMR genes.

Performance evaluation using a simulated dataset

To further evaluate the performance of PathoFact and compare it to other existing tools, the PathoFact pipeline was run on a simulated metagenome comprised of high-quality annotated genomes, i.e., the CAMI high complexity toy test dataset. Both MOCAT2 [26] and HUMAnN3 [27] were run on the original reads of the simulated CAMI datasets, while the same read datasets were processed and assembled with IMP followed by execution of PathoFact. Subsequently, annotations resulting from the different workflows were compared to evaluate the performance of PathoFact (Fig. 2a). PathoFact demonstrated increased numbers of predictions compared to both MOCAT2 and HUMAnN3 regarding virulence and toxin predictions (<0.05 , ANOVA) while performing similarly regarding AMR prediction compared to MOCAT2. Furthermore, and importantly, no additional curation or data-wrangling is needed for PathoFact compared to the other workflows tested above.

Additionally, we aimed to further characterize the performance of the metagenomic workflows against annotations of the CAMI high complexity toy test dataset. To achieve this, we annotated the underlying genomic data using the NCBI database of resistance genes [56], as well

as a BLAST search of the original 450 genomes against known virulence factors and toxin genes [3, 5]. The resulting annotations were compared to the prediction reports of PathoFact, MOCAT2, and HUMAnN3. PathoFact identifies a similar number of virulence factors and toxin genes in the annotated genomes compared to the original annotations, while MOCAT2 and HUMAnN3 identified a significantly lower number (Fig. 2b). Regarding antimicrobial resistance, PathoFact was able to identify many more gene variants compared to MOCAT2 and HUMAnN3 (Fig. 2c).

Performance of PathoFact on metagenomic datasets

Virulence factors and toxins may contribute to dysbiosis of the microbiome and favor a pro-inflammatory environment [65]. In addition, particular pathogenic bacteria may adapt to, and survive in, the presence of antimicrobials through acquisition or expression of AMR. Thereby, virulence factors, toxins, and AMR may all contribute to the pathogenic potential of the microbiome, which in turn may have an effect on the onset and development of disease and infection. The performance of PathoFact was demonstrated using three publicly available case-control metagenomic datasets which were chosen considering the following criteria: representing an actual infection or a chronic disease in which either pathogenic potential or toxins are believed to play a role. The Milani et al.'s [57] study represents actual infections with *Clostridioides difficile* (CDI) in the human gut microbiome of five patients along with five healthy controls. Furthermore, skin metagenomes of five psoriasis patients along with five healthy controls from Tett et al. [58] were chosen to represent a chronic disease in which a pathogenic potential is believed to have a function. Additionally, from Bedarf et al. [59], the metagenomes of fecal microbiomes derived from 10 early stage Parkinson's disease (PD) patients, as well as 10 age-matched controls, was obtained to represent a chronic disease in which bacterial toxins are believed to be involved [59].

Prediction of virulence factors and bacterial toxins

The predictions from PathoFact resulted in the identification of virulence factors in all three case-control metagenomic datasets. Furthermore, predicted virulence factors were characterized as secreted and non-secreted through the incorporation of SignalP in the pipeline. No statistically significantly (P value <0.05 , Wilcoxon rank

Table 2 Validation of the PathoFact pipeline

	Toxin prediction	Virulence factor prediction	AMR prediction
Sensitivity	0.777	0.886	0.963
Specificity	0.989	0.957	0.994
Accuracy	0.832	0.921	0.979

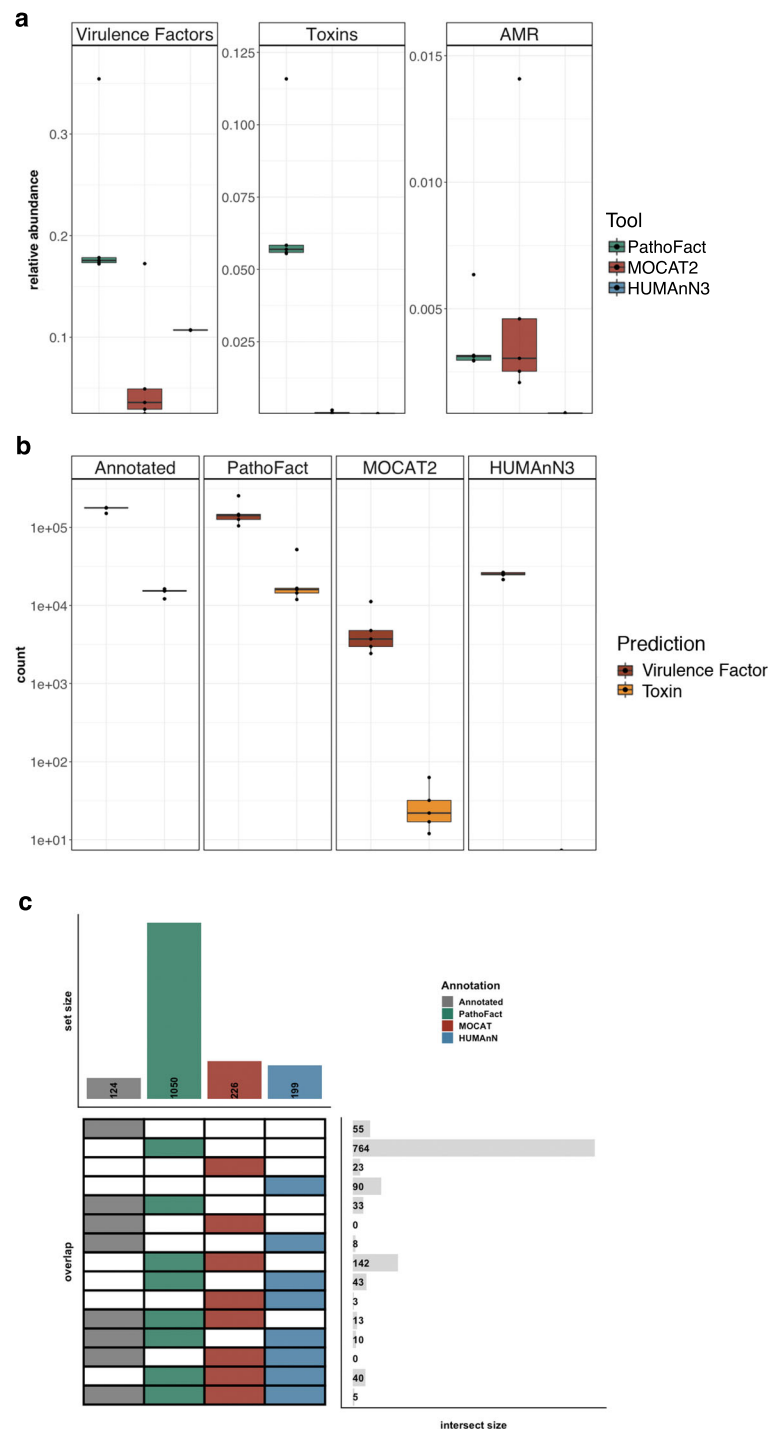


Fig. 2 Performance evaluation of PathoFact on a high-complexity simulated dataset. **a** The relative abundances (%) of virulence factors, including bacterial toxins, as well as antimicrobial resistance, as predicted by PathoFact, MOCAT2, and HUMAnN3, * two-way ANNOVA, *P* value < 0.05. **b** Total number of virulence factors and toxin genes identified in the annotated genome and as predicted by PathoFact, MOCAT2, and HUMAnN3 **c** Number of unique ARGs as annotated by the NCBI resistance database and as predicted by PathoFact, MOCAT2, and HUMAnN3

sum test) different relative abundance of the different virulence factors was found in any of the three studies when comparing diseased state and control (Fig. 3).

In addition to the general prediction of virulence factors using PathoFact, we identified bacterial toxins, as well as their corresponding HMM domain by which they

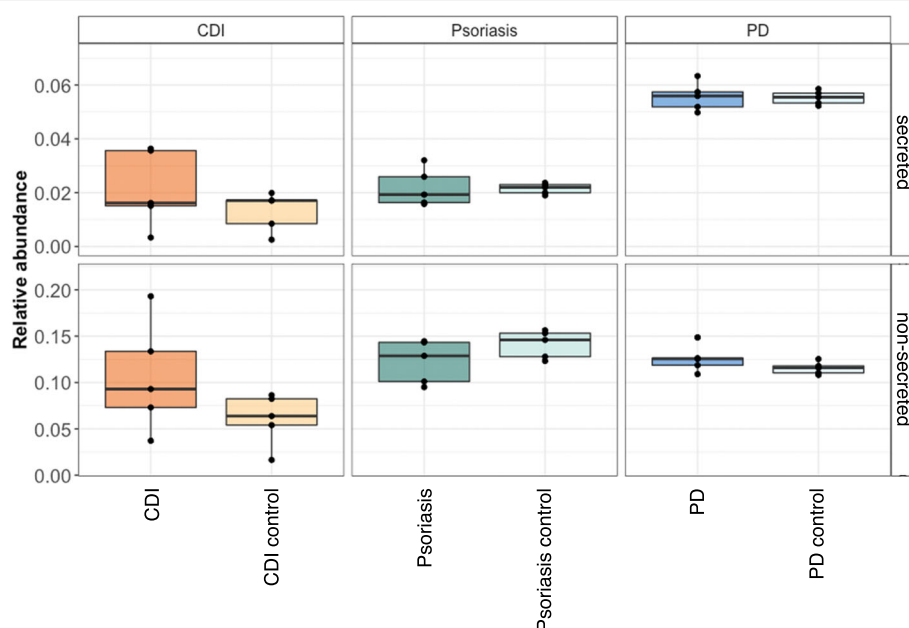


Fig. 3 Virulence factors in three case-control metagenomic datasets. The relative abundances (%) of both secreted and non-secreted virulence factors as well as non-pathogenic sequences in three metagenomic datasets (*Clostridioides difficile* infection, Psoriasis, Parkinson's disease)

were identified. Furthermore, both secreted and non-secreted toxins were identified in both diseased and control groups in all datasets (Fig. 4a), and we identified several differentially abundant bacterial toxins (Additional File 1: Table S3-S5). Within the CDI dataset, three distinct toxin domains, PF13953, PF13954, and PF06609, were identified to be differentially abundant in CDI over control (Fig. 4b). Interestingly, none of these toxin domains have yet been reported to be linked to CDI and therefore are of interest for further research. Four distinct toxin domains (K12340, PF13935, PF14449, and K11052) were found to be significantly abundant in psoriasis over controls (Fig. 4c). Of these toxin domains, only K12340 was previously linked to psoriasis [66]. Finally, regarding the PD study we found several differentially abundant bacterial toxins when comparing PD and control samples (Fig. 4d). Of these bacterial toxins, one containing the PF09599 domains was more abundant in PD and is among others found in invasins proteins in *Salmonella typhimurium* which has been hypothesized to be involved in Parkinson's disease [67]. Interestingly, in all three datasets additional “unknown” toxin domains were identified to be linked to the diseases, therefore representing interesting candidates for further research.

Prediction of antimicrobial resistance

Using the PathoFact pipeline, we predicted the presence of antimicrobial resistance genes in all three case-control metagenomic datasets. Within the CDI datasets, 23 ARG

categories were identified (Additional File 1: Figure S2a) of which six, i.e. diaminopyrimidine, elfamycin, fluoroquinolone, nucleoside, peptide, and multidrug, were significantly higher abundant in individuals with CDI over control (Fig. 5a). Antimicrobial resistance has previously been found to be associated with CDI infections [68]. In the metagenomic data of the skin microbiome, 22 categories of ARGs were identified (Additional File 1: Figure S2b). Interestingly, none of these resistance categories were found to be significantly different, neither with the diseased nor the control group. Within the PD study, 33 ARG categories were identified (Additional File 1: Figure S2c) with glycopeptide resistance significantly abundant in PD over controls, while tetracycline resistance was found to be enriched in the control group (Fig. 5c). The link between antimicrobial resistance and Parkinson's disease has been mostly unexplored thus far. However, a recently published study by Mertsalmi et al. [69] suggests a role for antibiotics in PD through the influence on the gut microbiome.

Although we propose the primary usage of PathoFact for metagenomic analyses, as seen with these three case-control metagenomic datasets, it can also be applied to single genome assemblies. Using the *Klebsiella pneumoniae* subsp. *Pneumoniae* HS11286 reference genome, we identified 86 resistance genes of which 6 contained SNPs contributing to resistance (Additional File 1: Table S6).

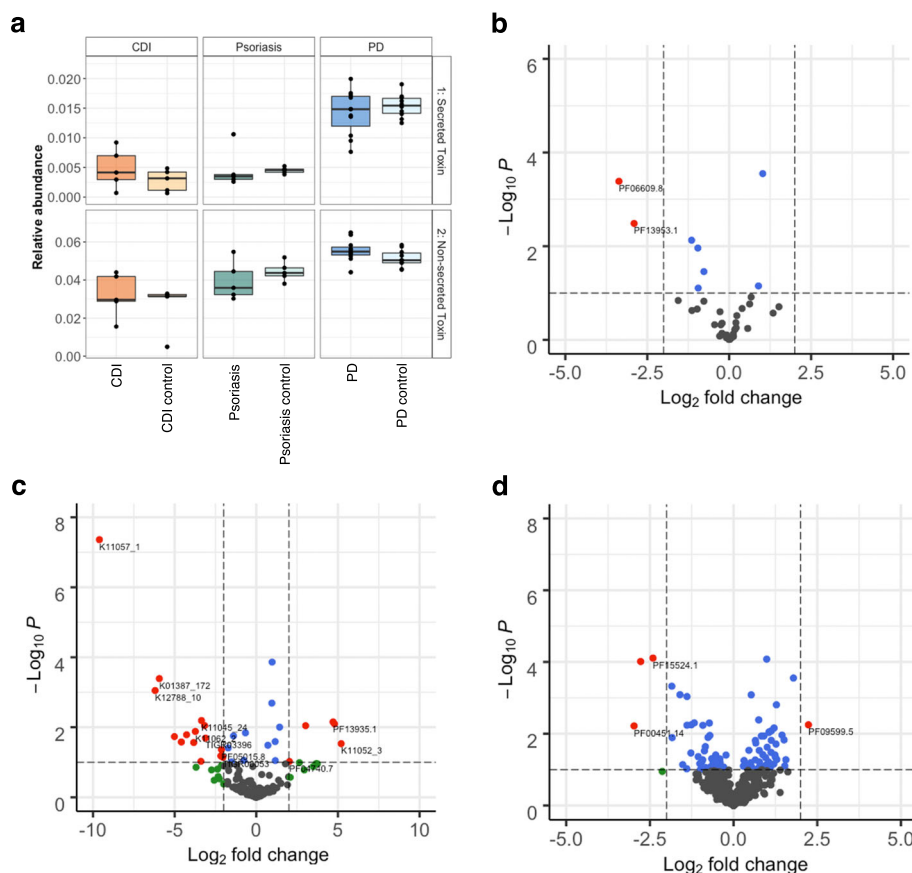


Fig. 4 Bacterial toxins in three case-control metagenomic datasets. Bacterial toxins in disease versus control datasets. **a** The relative abundance (%) of both secreted and non-secreted bacterial toxins in diseased versus control subjects. **b** Volcano plot depicting differentially abundant bacterial toxins in *Clostridioides difficile* infections versus control. **c** Volcano plot depicting differentially abundant bacterial toxins in Psoriasis versus control. **d** Volcano plot depicting differentially abundant bacterial toxins in Parkinson's disease versus control

Prediction of mobile genetic elements linked to virulence factors

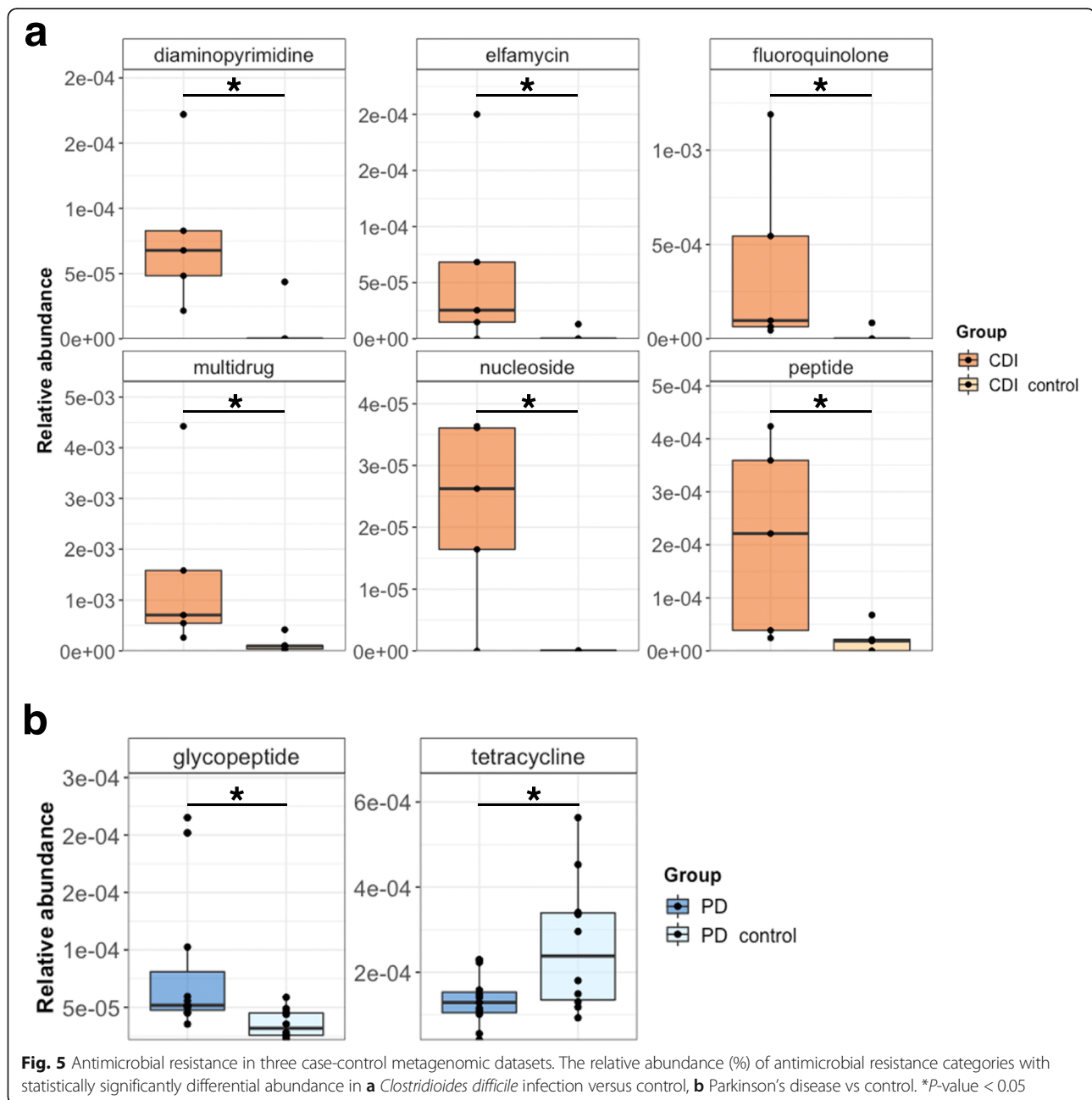
Using the predictions generated by PathoFact, we resolved the genomic contexts and identified MGEs in all three case-control metagenomic datasets (Fig. 6a) (Additional File 1: Figure S3). Within all three datasets, the presence of both phage- and plasmid-derived sequences was detected, although no significant difference was observed between diseased and control. We found that in all datasets the majority of MGEs were found to be both linked to virulence factors as well as AMR (~50%), closely followed by MGEs linked solely to virulence factors, including bacterial toxins, with AMR contributing to the remaining MGEs (Fig. 6b). Furthermore, a number of MGEs were found to be both linked to virulence factors as well as AMR.

Of the ARGs linked to MGEs, the prevalence of the different resistance categories were identified using our tool. Within the CDI dataset, the majority of the MGEs were linked to phenicol and beta-resistance in both

diseased and control groups (Additional File 1: Figure S4a). Additionally, plasmids linked to diaminopyrimidine and sulfonamide resistance were identified within the disease group while found to be absent in the control. Within the skin metagenomes, the majority of the predicted resistance genes linked to MGEs included beta-lactam, tetracycline, and multidrug resistance in both diseased and control groups (Additional File 1: Figure S4b). However, MGEs linked to beta-lactam resistance were found to be enriched in the diseased group. Finally, of the resistance genes within the PD study, both peptide and tetracycline resistances were found to be linked to phage and plasmids. Peptide resistance was abundant in controls whereas tetracycline was identified primarily in diseased (Additional File 1: Figure S4c).

Conclusions

The identification of virulence factors, toxins, and antimicrobial resistance genes are of immediate importance for understanding the pathogenic state of microbiomes.



Using our newly developed tool, PathoFact, we were able to identify virulence factors and bacterial toxins within three publicly available case-control metagenomic datasets. Furthermore, we were able to identify differentially abundant bacterial toxins when comparing diseased and control groups in all datasets. Additionally, antimicrobial resistance genes were identified in two of the datasets with a significant difference of certain resistance categories between diseased and control individuals. The inclusion of MGEs is of particular importance in understanding the

possible transmission of MGE-born virulence factors. With PathoFact, we identified MGEs in all three datasets and were able to link these simultaneously to the corresponding virulence factors, toxins, and antimicrobial resistance genes.

Until now, no single tool has existed which has combined these distinct aspects. Although several prediction tools exist for AMR, DeepARG and RGI have been chosen for their accuracy and ability to identify mutation contribution to resistance, and were included in our pipeline.

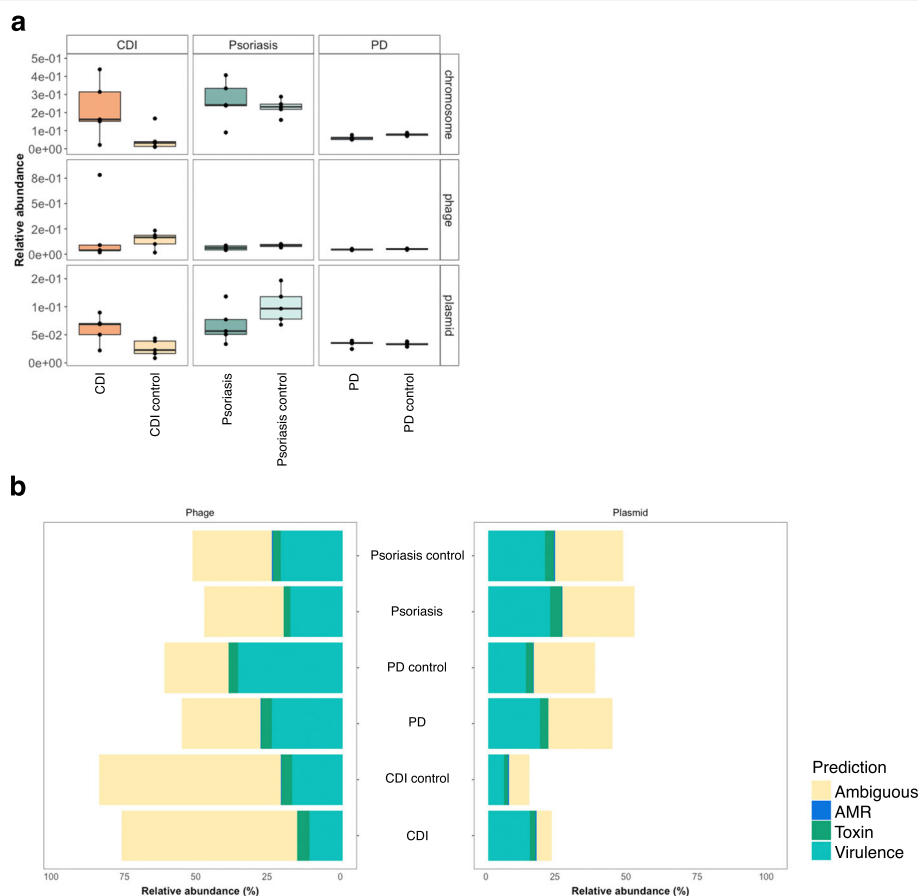


Fig. 6 Identification of MGEs within three case-control metagenomic datasets. Relative abundance of MGEs within three metagenomic datasets (*Clostridioides difficile* infection, psoriasis (skin), and PD). **a** The overall relative abundance of phage and plasmids within the *Clostridioides difficile* infection, psoriasis, and Parkinson's disease datasets. **b** The distribution of virulence factors, incl. toxins, and AMR between phage and plasmid in all datasets

Limited or no tools were available on the other hand for the prediction of toxins and virulence factors. PathoFact utilizes the wealth of currently available software (e.g., AMR and MGE predictions) as well as newly generated tools (e.g., virulence factors and toxins). Furthermore, PathoFact can conveniently integrate updates and newly developed prediction tools. In conclusion, our tool combines the strength of AMR predictions linked to MGE predictions and integrates this with the prediction of toxins and virulence factors. PathoFact is a versatile and reproducible pipeline by its ability to run either the complete workflow or each module on its own, giving the investigator flexibility in their analysis.

Availability and requirements

Project name: PathoFact

Project home page: <https://pathofact.lcsb.uni.lu>

Operating system(s): Platform independent

Programming language: python

Other requirements: snakemake (version ≥ 5.5), conda (version ≥ 4.7)

License: GNU GPLv3.

Restrictions to use by non-academics: see License

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40168-020-00993-9>.

Additional file 1. PathoFact supplementary materials.

Abbreviations

AMR: Antimicrobial resistance; ARG: Antimicrobial resistance gene; CDI: *Clostridioides difficile* infection; PD: Parkinson's disease

Acknowledgements

We are grateful for the feedback and beta-testing by Susana Martinez Arbas. The experiments presented in this paper were carried out using the HPC facilities of the University of Luxembourg.

Authors' contributions

LdN, SL, AHB, and PW designed this study. LdN with support of SL, CCL, PM, and AHB created the application. PathoFact was beta-tested by SB and VG. LdN and PW wrote the manuscript; CCL, PM, and AHB contributed to the review of the manuscript before submission. All authors read and approved the manuscript.

Funding

This work was supported by the Luxembourg National Research Fund (FNR) under grant CORE/BM/11333923, the Michael J. Fox Foundation under grant No. 14701, and the European Research Council (ERC-CoG 863664) to PW, and PRIDE/11823097 to LdN, CCL, PM, and PW.

Availability of data and materials

PathoFact, its models, and databases are available at <https://pathofact.lcsb.uni.lu>.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Systems Ecology Research Group, Luxembourg Centre for Systems Biomedicine, Esch-sur-Alzette, Luxembourg. ²Metagenomics Support Unit, German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany. ³Department of Soil Ecology, Helmholtz Centre for Environmental Research GmbH-UFZ, Halle (Saale), Germany. ⁴Bioinformatics Core, Luxembourg Centre for Systems Biomedicine, Esch-sur-Alzette, Luxembourg.

Received: 21 September 2020 Accepted: 29 December 2020

Published online: 17 February 2021

References

- Beceiro A, Tomás M, Bou G. Antimicrobial resistance and virulence: a successful or deleterious association in the bacterial world? *Clin Microbiol Rev.* 2013;26:185–230.
- Wu H-J, Wang AH-J, Jennings MP. Discovery of virulence factors of pathogenic bacteria. *Curr Opin Chem Biol.* 2008;12:93–101.
- Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, et al. VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.* 2005;33:D325–8.
- Finlay BB, Falkow S. Common themes in microbial pathogenicity revisited. *Microbiol Mol Biol Rev.* 1997;61:136–69.
- Chakraborty A, Ghosh S, Chowdhary G, Maulik U, Chakrabarti S. DBETH. a Database of Bacterial Exotoxins for Human. *Nucleic Acids Res.* 2012;40: D615–20.
- Schiavo G, van der Goot FG. The bacterial toxin toolkit. *Nat Rev Mol Cell Biol.* 2001;2:530–7.
- Martínez JL, Baquero F. Interactions among strategies associated with bacterial infection: pathogenicity, epidemicity, and antibiotic resistance. *Clin Microbiol Rev.* 2002;15:647–79.
- Mediavilla JR, Patrawalla A, Chen L, Chavda KD, Mathema B, Vinnard C, et al. Colistin- and Carbapenem-Resistant *Escherichia coli* Harboring *mcr-1* and *bla*NDM-5, Causing a Complicated Urinary Tract Infection in a Patient from the United States. *MBio.* 2016;7. Available from: <https://doi.org/10.1128/mBio.01191-16>
- O'Neill J. Antimicrobial resistance: tackling a crisis for the health and wealth of nations. Review on antimicrobial resistance. 2014;
- Brogan DM, Mossialos E. A critical analysis of the review on antimicrobial resistance report and the infectious disease financing facility. *Global Health.* 2016;12:8.
- MacLean RC, San Millan A. The evolution of antibiotic resistance. *Science.* 2019;365:1082–3.
- Sommer MOA, Dantas G, Church GM. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science.* 2009;325: 1128–31.
- Burrus V, Waldor MK. Shaping bacterial genomes with integrative and conjugative elements. *Res Microbiol.* 2004;155:376–86.
- Eloe-Fadrosh EA, Ivanova NN, Woyke T, Kyrpides NC. Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nat Microbiol.* 2016;1:15032.
- Alteio LV, Schulz F, Seshadri R, Varghese N, Rodríguez-Reillo W, Ryan E, et al. Complementary Metagenomic approaches Improve Reconstruction of Microbial Diversity in a Forest Soil. *mSystems.* 2020;5. Available from: <https://doi.org/10.1128/mSystems.00768-19>
- D'Costa VM, King CE, Kalan L, Morar M, Sung WWL, Schwarz C, et al. Antibiotic resistance is ancient. *Nature.* 2011;477:457–61.
- Tsai Y-K, Fung C-P, Lin J-C, Chen J-H, Chang F-Y, Chen T-L, et al. Klebsiella pneumoniae outer membrane porins OmpK35 and OmpK36 play roles in both antimicrobial resistance and virulence. *Antimicrob Agents Chemother.* 2011;55:1485–93.
- Barbosa TM, Levy SB. Differential expression of over 60 chromosomal genes in *Escherichia coli* by constitutive expression of MarA. *J Bacteriol.* 2000;182: 3467–74.
- Cabot G, Zamorano L, Moyà B, Juan C, Navas A, Blázquez J, et al. Evolution of *Pseudomonas aeruginosa* Antimicrobial Resistance and Fitness under Low and High Mutation Rates. *Antimicrob Agents Chemother.* 2016;60: 1767–78.
- Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome.* 2018;6:23.
- Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, et al. CARD 2020: antibiotic resistance surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 2020; 48:D517–25.
- Kleinheinz KA, Joensen KG, Larsen MV. Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and *E. coli* virulence genes in bacteriophage and prophage nucleotide sequences. *Bacteriophage.* 2014;4:e27943.
- Yin X, Jiang X-T, Chai B, Li L, Yang Y, Cole JR, et al. ARGs-OAP v2.0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. *Bioinformatics.* 2018;34:2263–70.
- Gupta A, Kapil R, Dhakan DB, Sharma VK. MP3: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data. *PLoS One.* 2014;9:e93907.
- Garg A, Gupta D. VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC Bioinformatics.* 2008;9:62.
- Kultima JR, Coelho LP, Forslund K, Huerta-Cepas J, Li SS, Driessen M, et al. MOCAT2: a metagenomic assembly, annotation and profiling framework. *Bioinformatics.* 2016;32:2520–3.
- Franzosa EA, McIver LJ, Rahnvarad G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods.* 2018;15:962–8.
- Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics.* 2018;34:3600.
- Anaconda INC. Conda. [cited 2018]. Available from: <https://anaconda.com>
- Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* 2013;41:e121.
- Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol.* 2019;37:420–3.
- Zhang R, Ou H-Y, Zhang C-TDEG. a database of essential genes. *Nucleic Acids Res.* 2004;32:D271–2.
- Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics.* 2006;22:1658–9.
- Rowe WPM, Winn MD. Indexed variation graphs for efficient and accurate resistance profiling. *Bioinformatics.* 2018;34:3601–8.
- Li W, Wooley JC, Godzik A. Probing metagenomics by rapid cluster analysis of very large datasets. *PLoS One.* 2008;3:e3375.
- Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat Microbiol.* 2016;2:16180.
- Finn RD, Coghill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 2016;44:D279–85.
- Lee Y, Tsai J, Sunkara S, Karamycheva S, Perteu G, Sultana R, et al. The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res.* 2005;33:D71–4.
- Kanehisa M, Goto S. KEGG. Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000;28:27–30.

40. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* 2004;32:D438–42.
41. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 2000;28:45–8.
42. Breiman L. Random Forests. *Mach Learn.* 2001;45:5–32.
43. Hastie T, Tibshirani R, Friedman J. *Random Forests*. Springer: The Elements of Statistical Learning; 2009. p. 567–603.
44. Chen Z, Zhao P, Li F, Leier A, Marquez-Lago TT, Wang Y, et al. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics.* 2018;34:2499–502.
45. McKinney W. Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*. From: <https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf>. Accessed 30 Sept 2019.
46. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods.* 2020;17:261–72.
47. Pedregosa F. Scikit-learn: Machine Learning in Python. *J Mach Learn Res.* 2011;12:2825–30.
48. Wishart D, Arndt D, Pon A, Sajed T, Guo AC, Djoumbou Y, et al. T3DB: the toxic exposome database. *Nucleic Acids Res.* 2015;43:D928–34.
49. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, et al. The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother.* 2013;57:3348–57.
50. Liu B, Pop M. ARDB—Antibiotic Resistance Genes Database. *Nucleic Acids Res.* 2009;37:D443–7.
51. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 2019;47:D506–15.
52. Krawczyk PS, Lipinski L, Dziembowski A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.* 2018;46:e35.
53. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. *PeerJ.* 2015;3:e985.
54. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, et al. Identifying viruses from metagenomic data by deep learning. *arXiv [q-bio.GN]*. 2018. from: <http://arxiv.org/abs/1806.07810>. Accessed 30 Sept 2019.
55. Pierce NT, Irber L, Reiter T, Brooks P, Brown CT. Large-scale sequence comparisons with sourmash. *F1000Res.* 2019;8:1006.
56. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, et al. Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrob Agents Chemother.* 2019;63. from: <https://doi.org/10.1128/AAC.00483-19>. Accessed 25 Oct 2020.
57. Milani C, Ticinesi A, Gerritsen J, Nouvenne A, Lugli GA, Mancabelli L, et al. Gut microbiota composition and *Clostridium difficile* infection in hospitalized elderly individuals: a metagenomic study. *Sci Rep.* 2016;6:25945.
58. Tett A, Pasolli E, Farina S, Truong DT, Asnicar F, Zolfo M, et al. Unexplored diversity and strain-level structure of the skin microbiome associated with psoriasis. *NPJ Biofilms Microbiomes.* 2017;3:14.
59. Bedarf JR, Hildebrand F, Coelho LP, Sunagawa S, Bahram M, Goeser F, et al. Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naïve Parkinson's disease patients. *Genome Med.* 2017;9: 39.
60. Narayanasamy S, Jarosz Y, Muller EEL, Heintz-Buschart A, Herold M, Kaysen A, et al. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.* 2016;17:260.
61. Liao Y, Smyth GK, Shi W. featureCounts: An efficient general-purpose program for assigning sequence reads to genomic features. *arXiv [q-bio.GN]*. 2013. from: <http://arxiv.org/abs/1305.3347>. Accessed 20 Oct 2019.
62. Hu Y, Yang X, Qin J, Lu N, Cheng G, Wu N, et al. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nat Commun.* 2013;4:2151.
63. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15:550.
64. Szczyba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, et al. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. *Nat Methods.* 2017;14:1063–71.
65. Forsyth CB, Shannon KM, Kordower JH, Voigt RM, Shaikh M, Jaglin JA, et al. Increased intestinal permeability correlates with sigmoid mucosa alpha-synuclein staining and endotoxin exposure markers in early Parkinson's disease. *PLoS One.* 2011;e28032.
66. Trepod CM, Mott JE. Identification of the *Haemophilus influenzae* *tolC* gene by susceptibility profiles of insertionally inactivated efflux pump mutants. *Antimicrob Agents Chemother.* 2004;48:1416–8.
67. Chaudhuri D, Roy Chowdhury A, Biswas B, Chakravorty D. *Salmonella Typhimurium* Infection Leads to Colonization of the Mouse Brain and Is Not Completely Cured With Antibiotics. *Front Microbiol.* 2018;9:1632.
68. Shah D, Dang M-D, Hasbun R, Koo HL, Jiang Z-D, DuPont HL, et al. *Clostridium difficile* infection: update on emerging antibiotic treatment options and antibiotic resistance. *Expert Rev Anti Infect Ther.* 2010;8:555–64.
69. Mertsalmi TH, Pekkonen E, Scheperjans F. Antibiotic exposure and risk of Parkinson's disease in Finland: A nationwide case-control study. *Mov Disord.* 2020;35:431–42.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



Appendix A.3

Persistence of birth mode-dependent effects on gut microbiome composition, immune system stimulation and antimicrobial resistance during the first year of life

ARTICLE OPEN



Persistence of birth mode-dependent effects on gut microbiome composition, immune system stimulation and antimicrobial resistance during the first year of life

Susheel Bhanu Busi^{1,10}, Laura de Nies^{1,10}, Janine Habier¹, Linda Wampach¹, Joëlle V. Fritz^{1,2,9}, Anna Heintz-Buschart^{1,3,4}, Patrick May⁵, Rashi Halder¹, Carine de Beaufort^{1,6,7} and Paul Wilmes^{1,8}✉

© The Author(s) 2021

Caesarean section delivery (CSD) disrupts mother-to-neonate transmission of specific microbial strains and functional repertoires as well as linked immune system priming. Here we investigate whether differences in microbiome composition and impacts on host physiology persist at 1 year of age. We perform high-resolution, quantitative metagenomic analyses of the gut microbiomes of infants born by vaginal delivery (VD) or by CSD, from immediately after birth through to 1 year of life. Several microbial populations show distinct enrichments in CSD-born infants at 1 year of age including strains of *Bacteroides caccae*, *Bifidobacterium bifidum* and *Ruminococcus gnavus*, whereas others are present at higher levels in the VD group including *Faecalibacterium prausnitzii*, *Bifidobacterium breve* and *Bifidobacterium kashiwanohense*. The stimulation of healthy donor-derived primary human immune cells with LPS isolated from neonatal stool samples results in higher levels of tumour necrosis factor alpha (TNF- α) in the case of CSD extracts over time, compared to extracts from VD infants for which no such changes were observed during the first year of life. Functional analyses of the VD metagenomes at 1 year of age demonstrate a significant increase in the biosynthesis of the natural antibiotics, carbapenem and phenazine. Concurrently, we find antimicrobial resistance (AMR) genes against several classes of antibiotics in both VD and CSD. The abundance of AMR genes against synthetic (including semi-synthetic) agents such as phenicol, pleuromutilin and diaminopyrimidine are increased in CSD children at day 5 after birth. In addition, we find that mobile genetic elements, including phages, encode AMR genes such as glycopeptide, diaminopyrimidine and multidrug resistance genes. Our results demonstrate persistent effects at 1 year of life resulting from birth mode-dependent differences in earliest gut microbiome colonisation.

ISME Communications (2021)1:8; <https://doi.org/10.1038/s43705-021-00003-5>

INTRODUCTION

The rate of caesarean section delivery is constantly increasing worldwide, which is partly driven by increases in overall income and access to health facilities.¹ According to a 2015 report, 29.7 million births occurred via CSD in that year accounting for ~18% of the births in 169 countries.¹ At 25% in Europe, this number is higher than the global average.² The short-term risks of CSD include delayed or altered development of the immune system,³ reduced gut microbiome diversity,⁴ limited transmission of bacterial strains from mother to neonate^{5,6} and microbiome-borne functional deficiencies.^{7–10} Although few studies associate CSD with metabolic disorders^{11,12} and allergies,^{13,14} the long-term effects of birth mode are not well understood. Shao et al. reported that CSD may predispose individuals to colonisation by opportunistic pathogens including those carrying antimicrobial resistance

(AMR) genes.¹⁵ On the one hand, several reports including our previously published study⁸ addressed questions concerning the very early development of the neonate's gut microbiomes^{14,16} and immune system priming³ in relation to disease development.^{17,18} On the other hand, only few reports^{19–22} follow the effects of birth mode during the first year of life especially in relation to immune system priming, development and evolution of AMR, and the contribution of mobile genetic elements to the persistence of AMR genes.

Factors including environmental exposure,¹⁴ breast feeding and diet^{19,23,24} and genetics²⁵ play crucial roles in the development of an infant. Aside from this, it is now generally accepted that birth mode, i.e. vaginal delivery (VD) or CSD, has a pronounced impact on early microbiome structure^{3,8,11,26,27} While the majority of these studies focus on overall microbiome structure, analyses of

¹Systems Ecology Group, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg. ²Translational Neuroscience group, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg. ³Metagenomics Support Unit, German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Halle, Germany. ⁴Department of Soil Ecology, Helmholtz-Centre for Environmental Research GmbH - UFZ, Halle, Germany. ⁵Bioinformatics Core, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg. ⁶Centre Hospitalier de Luxembourg, Department of Pediatric Endocrinology and Diabetes, Luxembourg, Luxembourg. ⁷Department of Pediatric Endocrinology, UZ Brussel, Vrije Universiteit Brussel, Brussels, Belgium. ⁸Department of Life Sciences and Medicine, Faculty of Science, Technology and Medicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg. ⁹Present address: Transversal Translational Medicine, Luxembourg Institute of Health (LIH), 1445 Strassen, Luxembourg. ¹⁰These authors contributed equally: Susheel Bhanu Busi, Laura de Nies. ✉email: paul.wilmes@uni.lu

Received: 27 October 2020 Revised: 9 December 2020 Accepted: 2 January 2021

Published online: 26 March 2021

the functional contribution of the microbiome has attracted attention due to its sensitivity to perturbation.²⁸ For example, we previously reported that the microbiome in VD-born babies was enriched in bacterial genes encoding for lipopolysaccharide (LPS) biosynthesis, cationic antimicrobial peptide resistance as well as two-component systems.⁸ Interestingly, higher levels of LPS biosynthesis genes were associated with increased immune responses in VD neonates, whereas CSD neonates had reduced levels of TNF- α and IL-18 immediately after birth. Noteworthy in this context is previous work by Vatanen et al. which showed that differing LPS immunogenicity contributes to autoimmunity thereby affecting the long-term health outcomes of infants exposed to different antigens.²⁹ Furthermore, others have hypothesized and reported^{3,30} a similar phenomenon, whereby the gut microbiome contributes to the development of the immune system during a "critical window" of development.^{3,30–36} In a neonatal cohort at risk for the development of asthma, bacterial metabolites were shown to specifically impede immune tolerance.³⁷ However, some of the reports described above do not elaborate on the continuous effect of early immune system priming in the context of the birth mode and especially over the course of the first year of life, including whether these effects normalize over time.

Aside from the well-studied factors and consequences of development described above, the role of commensal microbiota in the emergence and spread of AMR is not well understood. Recent studies have reported that antibiotic exposure in infancy affects microbial diversity, and enriches AMR genes. Interestingly, Ravi et al. have suggested that the infant gut microbiome acts as a reservoir for multidrug resistance that persists throughout infancy up to 2 years of age.³⁸ They reported that integrons (*int1* gene) in the gut could potentially be responsible for this phenomenon. Nevertheless, the effect of birth mode, CSD or VD, on the transmission and occurrence of AMR remains unresolved.

Here, we address the aforementioned gaps in knowledge concerning the effect of birth mode on the persistence of the gut microbiota over the first year of life including their inherent functions, immunogenic properties and their role in conferring AMR. Our results highlight birth mode-dependent differences in gut microbiome structure and their association with immune function. We found that the gut microbiota becomes similar between CSD and VD babies at 1 year of age, with the exception of an immunostimulatory commensal, *Faecalibacterium prausnitzii*, which was enriched in the VD group. In addition, we identified an increased abundance in AMR genes directed against synthetic and semi-synthetic antibiotics in CSD as early as 5 days *postpartum*. Strikingly, we found that mobile genetic elements (MGEs) including plasmids and bacteriophages are key contributors to the establishment and persistence of AMR, irrespective of birth mode. Collectively, our findings suggest that birth mode-dependent effects persist through the first year of life including the delayed immunostimulation of CSD infants likely affecting tolerance mechanisms as well as the apparent role of bacteriophages in conferring AMR.

RESULTS

Birth mode-dependent gut microbiota differences during the first year

We previously described the initial seeding and colonisation processes within the human gut microbiome and identified differences in microbiome structure and function as well as linked immunogenicity and immune system priming, which stratified according to birth mode.^{8,39} Building on this work, we aimed to understand the long-term effects in relation to the observed differences, especially through the first year of life which represents a "critical window" of development including physiological growth and immune system maturation. To achieve this,

we followed VD and CSD neonates in our cohort and collected faecal samples at crucial intervals after birth, including 5 days, 1 month, 6 months and at 1 year of age (Fig. 1a). In one of our previous studies,⁸ a multivariate analysis was performed to compare the profiles of CSD (\pm SGA) to VD neonates. The results of these analyses demonstrated that delivery mode was the strongest determining factor in the microbial profile and predicted functions, irrespective of the infants were born SGA or not.⁸ In light of these analyses, we included the SGA samples within the CSD group. We reconstructed microbial genomes and identified differentially abundant taxa and functions between the groups using metagenomic sequencing data. Based on metagenomic operational taxonomic units (mOTUs), we calculated the Jensen–Shannon divergence index and found that the intra-group variability within CSD or VD was minimal while the inter-group variability between CSD and VD groups was significantly different (Supplementary Fig. 1). At the genus level, our data also recapitulated previously described⁸ significantly increased levels of *Bacteroides* (FDR-adjusted $p < 0.05$, Wilcoxon rank-sum test) in the VD neonates compared to CSD at the early timepoints (day 5 after birth and at 1 month). *B. caccae* also showed an increasing trend in the CSD group at 6 months and after 1 year of age, while *B. caecimuris* was significantly increased in CSD at 5 days after birth (Fig. 1b; FDR-adjusted $p < 0.05$, Wilcoxon rank-sum test) and showed an increasing trend at 1 year of age. However, at 1 year of age, the abundance of this genus in samples from CSD neonates was comparable to the levels in the VD group. In contrast, the levels of *Bifidobacterium* were increased in VD after 6 months, while *Faecalibacterium prausnitzii*, a commensal associated with healthy human microbiomes,⁴⁰ was found to be significantly increased in the VD group at 1 year of age (Fig. 1b, FDR-adjusted $p < 0.05$, Wilcoxon rank-sum test). We further found that both birth mode and the neonatal age affect the gut microbiome community structure, whereby the latter contributes highly to variation within and between the groups (Fig. 1c). The taxonomic profiles at 1 year of age were distinct when compared to day 5, 1 month and 6 months from both groups.

Assessment of differences in metagenomic functional potential at 1 year of age

Taxonomic differences within the gut microbiome populations may not always manifest as differences in functional diversity due to the redundancy in the latter. To address this, we assigned KEGG⁴¹ orthology identifiers (KOs) to each gene identified from both groups. We found 84 differentially abundant KOs between VD and CSD samples at 1 year of age (Fig. 2a). In addition, we linked all identified KOs ($n = 7103$) to their corresponding KEGG orthology pathways (Fig. 2b) and performed differential pathway analyses. We found that the VD group showed an increase in the gene copy numbers of pathways involved in carbapenem and phenazine biosyntheses (Fig. 2c). We found that 21 unique genera were associated with carbapenem biosynthesis across both groups (Supplementary Data 1) spanning all major phyla found within the gut.

Pro-inflammatory immune responses elevated in CSD after 1 year of life

In the early stages of neonatal development, we found that the immune activation potential of LPS was significantly increased in samples from VD neonates,⁸ whereby the isolated LPS triggered the secretion of TNF- α and IL-18 by monocyte-derived dendritic cells (MoDCs) from four healthy adult donors. To determine if the immunostimulatory potential persisted at 1 year of age, we stimulated the MoDCs (obtained from four healthy adult donors) with LPS isolated from the faecal samples of the CSD and VD groups. In addition to TNF- α and IL-18, we also tested the potential of LPS to stimulate secretion of pro- and anti-inflammatory cytokines such as IL-1 β , IL-12, IL-8 and IL-10 (Fig.

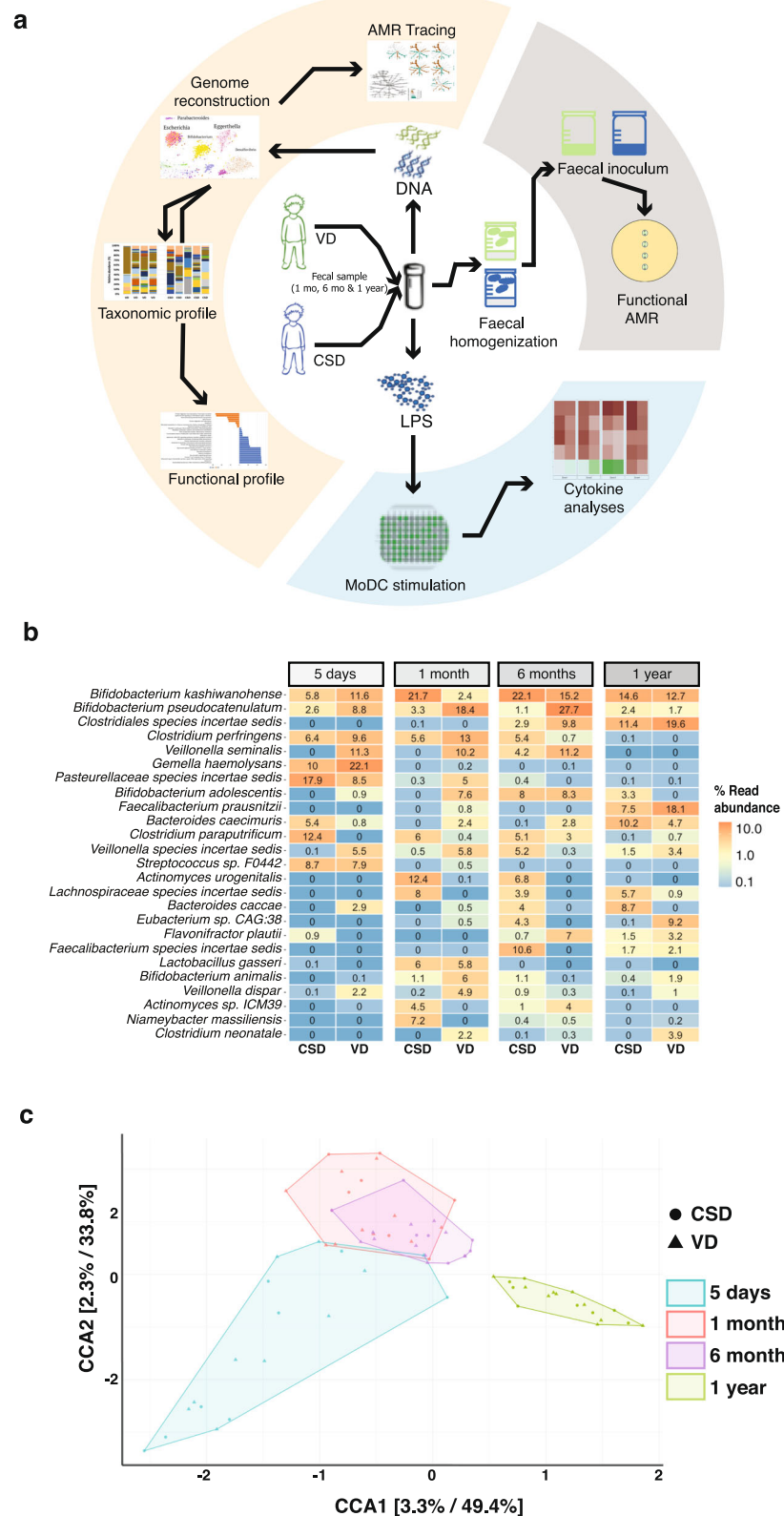


Fig. 1 Gut microbiome profiles throughout the first year of life. **a** Workflow representation of DNA and LPS isolation from faecal samples for metagenomic, immune and functional AMR analyses. **b** Relative abundances of metagenomic operational taxonomic units (mOTUs) >1% abundance at day 5 after birth, 1 month, 6 months, and at 1 year of age. **c** Canonical correlation analyses (CCA) resolving the stratification of taxonomic profiles based on two covariates, i.e. birth mode and time when samples were sequenced.

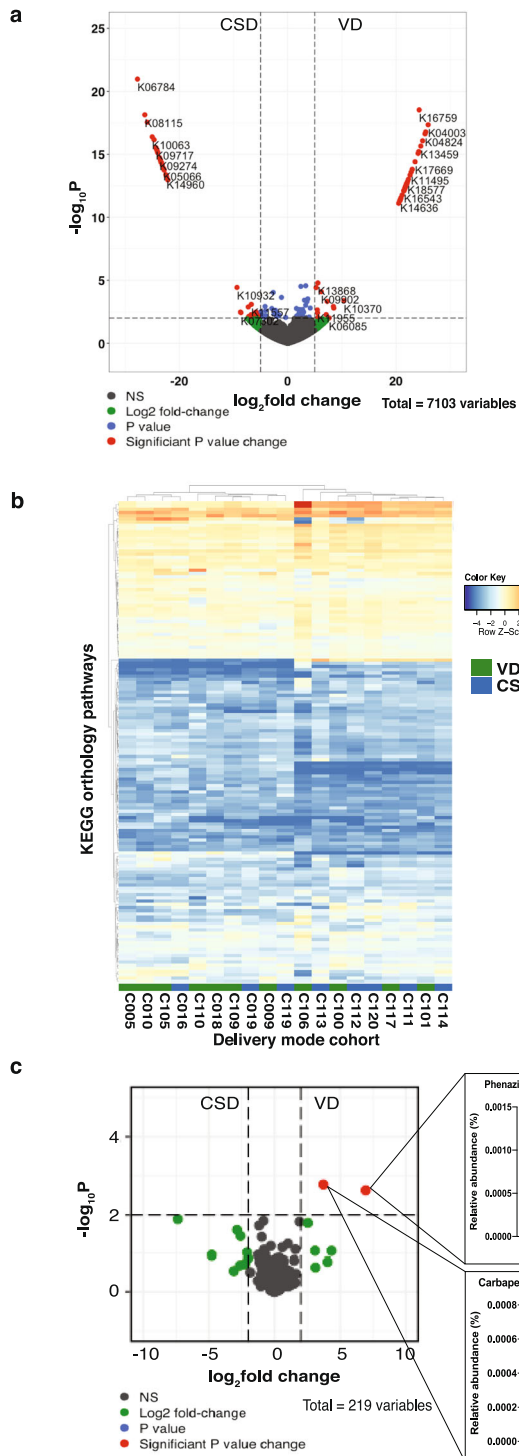


Fig. 2 Functional differences at 1 year of age. **a** Volcano plot depicting the statistically significantly different KEGG orthologs found in both CSD and VD groups at 1 year of age. A total of 6413 variables were tested, with $-\log_{10}(p\text{-value})$ shown on the y-axis. Green dots indicate KOs with a fold change >2 . Red dots indicate KOs with a significant fold-change cut-off of 2, and with a false-discovery rate-adjusted p -value cut-off of 0.01. **b** Heatmap based on the KEGG pathways found in both CSD and VD samples at 1 year of age. Each row denotes a pathway represented by the KO genes, with the hierarchical clustering being based on Euclidean distances using Ward's clustering algorithm. **c** Volcano plot of the 219 KEGG pathways to which the KOs were mapped, tested for significance with a fold-change cut-off of 2, and with a false-discovery rate-adjusted p -value cut-off of 0.01. The insets show carbapenem and phenazine biosynthesis pathways that were statically significantly different. *** p -value < 0.001 .

TNF- α levels at 1 year of age were positively correlated with the abundance of several mOTUs, including *Bacteroides caecimuris* and *Haemophilus influenzae* (Fig. 3c). Previous reports⁴² suggest that Enterobacteriaceae levels correlate with inflammatory levels. However, we did not find a correlation of these taxa with LPS levels in our study (Supplementary Fig. 2). Our data also indicate an increase in the number of Gram-negative (G-ve) bacteria at 1 year of age compared to day 5 after birth in the CSD group (Supplementary Fig. 3). The increase in Shannon diversity at 1 year of age compared to day 5 coupled with the increase in G-ve bacteria provides a mechanistic explanation why the LPS stimulation of donor cells from faecal samples of CSD resulted in similar levels of TNF- α (Fig. 3b), as observed with faecal samples from the VD group at 1 year of age.

Antimicrobial resistance modulated by birth mode

The analyses of the functional potential based on KEGG orthology revealed a stratification of antibiotic biosynthesis pathways based on whether an infant was born by CSD or VD (Fig. 2c). To assess and validate the impact of birth mode on the presence and persistence of AMR, we used a deep-learning approach⁴³ to annotate antibiotic resistance genes in our metagenomic data.⁴⁴ We determined the presence and relative abundance of AMR genes in samples collected from both CSD and VD at day 5 after birth, 1 month, 6 months and at 1 year of age. The samples collected from CSD neonates exhibited an increased abundance in AMR genes at the earliest time point (day 5) compared to the VD group. In addition, we found that the number of AMR genes detected in CSD infants at 1 year of age was significantly reduced in comparison to the CSD samples at day 5 (FDR-adjusted $p < 0.0021$, Wilcoxon rank-sum test; Fig. 4a, Supplementary Fig. 4a). To corroborate our observations on the levels of AMR genes, we assessed the abundance of AMR genes using a random subset of samples from the resistome study by Gasparri et al.⁴⁵ We found that the overall levels of AMR genes starting at 1 month through to 1 year of age were similar in their study to those observed in our own cohort (Supplementary Fig. 4b). Meanwhile in our study, at 1 year of age, we found several genes that were differentially abundant between the CSD and VD groups (FDR-adjusted $p < 0.05$, Wilcoxon rank-sum test; Fig. 4b). Since various genes can confer resistance to the same antibiotic, we regrouped the genes into their respective categories such as multidrug, tetracycline resistance etc. We found that genes conferring glycopeptide, phenicol, pleuromutilin, bacitracin, sulfonamide and diaminopyrimidine resistance were significantly increased in CSD compared to VD at day 5 after birth (Fig. 4c, Supplementary Fig. 5; FDR-adjusted $p < 0.05$, Wilcoxon rank-sum test). Interestingly, diaminopyrimidine, phenicol, pleuromutilin, and sulfonamide are synthetic or semi-synthetic antibiotics and, likely prevalent in the hospital environment.^{46–49} However, these differences did not persist

3a). Interestingly, at 1 year of age, IL-18 was below the detection limits. We did not find any significant differences between the CSD and VD groups with respect to the levels of secreted TNF- α at 1 year of age. However, contrary to the patterns observed at 5 days after birth, we found that the levels of TNF- α stimulated by LPS were significantly increased at 1 year of age within the CSD group ($p < 3.5 \times 10^{-5}$, paired two-way ANOVA; Fig. 3b). Interestingly, the increase in stimulated TNF- α levels in CSD at 1 year of age was similar to the level of the cytokine stimulated by LPS from the day 5 VD samples (Fig. 3b). In addition, we found that the stimulated

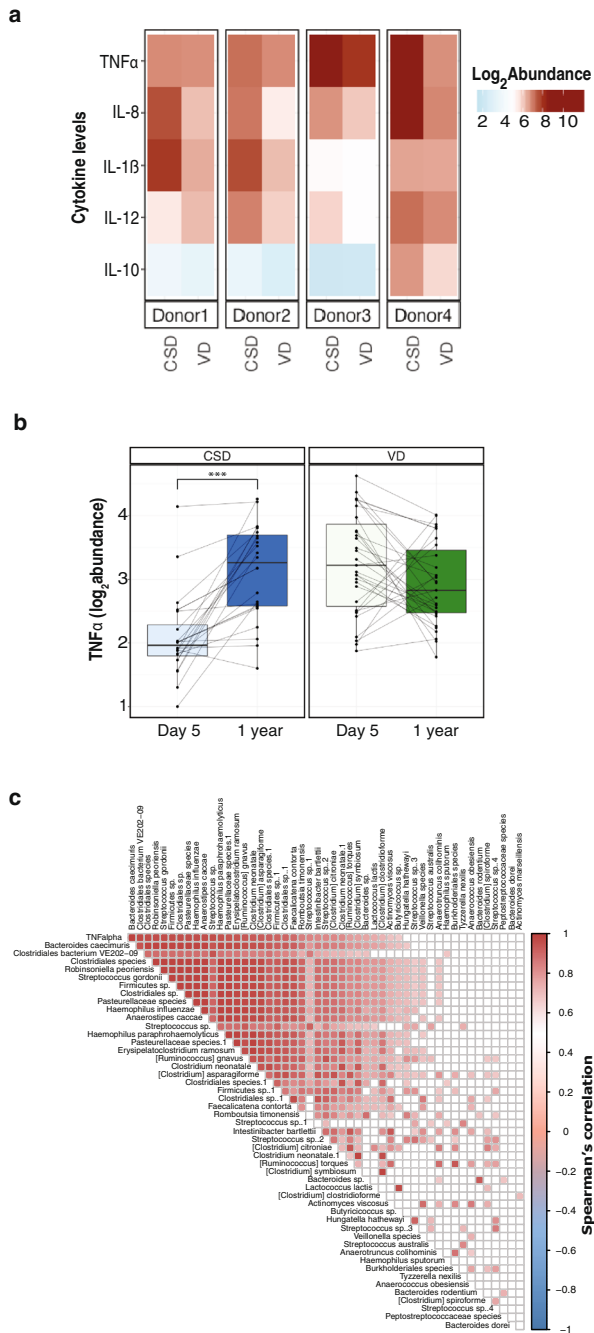


Fig. 3 Immunostimulatory potential at 1 year of age. **a** Heatmap depicting the abundance (log₂) of pro- and anti-inflammatory cytokines levels were measured by stimulating MoDCs from healthy donors (Donor 1–4) with LPS isolated from faecal samples of CSD and VD neonates. **b** Boxplots depicting the TNF- α levels in both groups (CSD and VD) at day 5 after birth and 1 year of age. Paired two-way ANOVA (analysis of variance) *p*-values are listed in the plot to depict significant differences. ****p*-value < 0.001. **c** Correlation of TNF- α levels (row 1) with the relative abundance of metagenomic OTUs based on canonical correlation analysis. Filled squares indicate significantly correlated taxa, whereas colour indicates positive (red) or negative (blue) correlation.

over time. In addition, the mothers in our cohort across both groups (CSD:6 and VD:1) received prophylactic treatment against group B *Streptococcus* (Supplementary Data 2) in the form of cephalosporin. However, we did not find any distinguishing

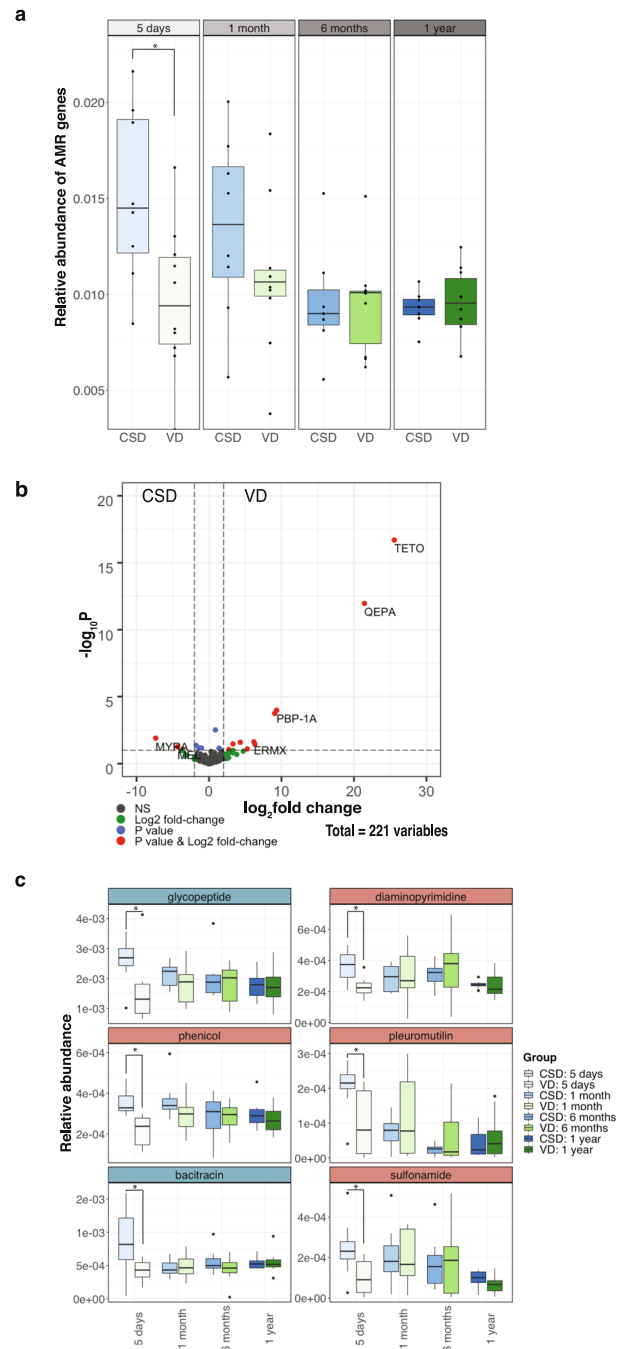


Fig. 4 Antimicrobial resistance gene abundances over time. **a** Boxplots of the overall AMR gene abundance in CSD and VD samples at different timepoints including 5 days after birth and at 1 year of age. Wilcoxon rank-sum tests were used to test for significance. **p* < 0.05. **b** Volcano plot depicting the significantly enriched genes in either CSD or VD samples at 1 year of age. **c** AMR categories which are significantly different between the groups at any of the timepoints are shown. Wilcoxon rank-sum tests were used to test for significance. **p* < 0.05.

patterns within the resistance categories corresponding to this treatment regimen. Albeit a limited sample size, we also tracked the diet including feeding method (bottle- or breast-fed), antibiotic regimen and physical characteristics through the first year and did not find any significant correlations with functional pathways including AMR (Supplementary Fig. 6 and Supplementary Data 3).



Fig. 5 Taxa and mobile genetic elements associated with antimicrobial resistance. **a** Tree plots depicting the median proportion of AMR-associated taxa at day 5 after birth and at 1 year of age comparing the CSD and VD groups. Values used for plotting the trees are an average of all the samples within the group. **b** The relative abundances of AMR genes found on the bacterial chromosome, plasmids or phages at the different timepoints, ranging from day 5 after birth through to 1 year of age. CSD = blue (left panel), VD = green (right panel). Paired two-way ANOVA was used to assess significant differences; adjusted *p*-values are shown in the plot to depict significant differences. **p*-value < 0.05. **c** Stacked bar plot depicting the AMR categories transmitted via phages and plasmids at all timepoints. Each colour in the plot is associated with a category listed in the legend on the right. The plot represents mean values for all samples in each group.

We compared samples from day 5 after birth versus 1 year of age within each birth mode group independently to differentiate between taxonomic groups contributing to AMR. Within CSD samples, we found that *Enterobacteriales* and *Staphylococcaceae* were enriched at day 5 after birth while major AMR contributors at 1 year of age were *Lachnospiraceae*, *Bacteroidaceae*, *Actinobacteria* and *Oscillospirales*. Conversely, within VD samples early AMR resistance was mainly attributed to the abundance of *Bacteroidales*, *Lactobacillales*, *Propionibacteraceae* and *Enterobacteriaceae* at day 5 after birth. Meanwhile, at 1 year of age, VD samples were enriched in taxa including *Lachnospiraceae*, *Ruminococcaceae*, *Veillonellales* and *Eggerthellaceae* with respect to contribution of AMR genes to the resistome (Fig. 5a). These data suggest that AMR genes are also encoded by commensals apart from pathogens which, in the context of the present study, sustain their presence throughout the first year of life.

Role of mobile genetic elements in antimicrobial resistance

Bacterial genomes have been fine-tuned over evolutionary timescales,⁵¹ potentially refining their defence mechanisms against various biocidal agents including chemicals. Aside from these, bacterial components such as MGEs are known to be potent factors in the spread of AMR⁵² and can transfer genes across distinct taxonomic clades. An example of such MGEs are plasmids as well as viruses including bacteriophages, which actively drive the transfer of genetic material.⁵³ To determine the role of MGEs in conferring AMR in our neonate cohort, we analysed the genomic context of the AMR genes. The contigs were classified as chromosomal, plasmid, phage, ambiguous (those that could not be resolved) and unclassified. In this study, chromosomal sequences refer to the bacterial genome excluding plasmids, in accordance with the PlasFlow⁵⁴ methodology. These criteria were used to assess the role of MGEs at all timepoints. The majority (average of ~75%) of the AMR genes were encoded on the bacterial chromosome (Fig. 5b). This phenomenon was prominent in the VD samples irrespective of sampling time point. On the other hand, the mean relative abundance of AMR genes encoded on plasmids (~5%) was marginally increased in the CSD group at both 5 days after birth and at 1 year of age. Overall, we found that phages encoded lower levels (1–3%) of AMR genes compared to the plasmids. However, we found that the relative abundances of phages encoding AMR were significantly increased after 1 year of age (Fig. 5b). Interestingly, we did not find any significant

differences between the birth modes in relation to the virome profiles at any of the timepoints (FDR-adjusted $p > 0.05$, two-way ANOVA, Supplementary Fig. 7). However, a large proportion of the contigs were either ambiguous or unclassified but demonstrated an even distribution across all timepoints (Supplementary Fig. 8a). When ambiguous sequences mapping to both the bacterial chromosome and phages are included in the phage abundance metrics, it results in a higher abundance of AMR genes conferred by phage compared to plasmids (Supplementary Fig. 8b).

Distribution of AMR categories encoded by mobile genetic elements

Assessing AMR conferred by MGEs, we found that both plasmids and phages encoded genes conferring resistance to several classes of antibiotics (Fig. 5c). Though significant differences were not apparent, we found that phage-encoded AMR genes against vancomycin (glycopeptide) and numerous other antimicrobials were dominant in both birth mode groups. In addition, plasmids conferred resistance to diaminopyrimidine and bacitracin, as well as β -lactams, phenicol, MLS and tetracyclines (Fig. 5c). Strikingly, these data suggest that MGE-mediated AMR, encoded by phages, is a potential factor in conferring AMR or serve as a reservoir for antimicrobial resistance throughout the first years of life.

Phage-mediated horizontal gene transfer (HGT)

To understand phage-mediated horizontal gene transfer of AMR we analysed, in detail, phage contigs encoding AMR genes. We identified several genes that were horizontally transferred within the CSD and VD groups (Supplementary Data 4 and Supplementary Fig. 9). CSD samples ($n = 3$; C112, C113 and C119), exhibited HGT involving AMR genes including resistance to glycopeptide and multidrug (Fig. 6a–d). The majority (~88%) mapped to the bacterial chromosome. However, two genes encoding multidrug resistance were encoded by both chromosome and phage (C119: contig 2568, Supplementary Data 4). The contig was found to be a candidate prophage based on detailed inspection and was found to encode several genomic regions with prophage signatures flanking the multidrug resistance genes (Fig. 6f). In addition, the coverage of the contig across its entire length was more variable in the genomic regions where the prophage and AMR gene sequences were identified. Resolution of the taxa involved indicated HGT between the *Intestinimonas butyriciproducens* (GCA 003096335) and *Clostridium bolteae* (ATCC BAA 613 GCA 000154365) strains belonging to the *Oscillospirallales* and *Lachnospirallales* orders, respectively (Supplementary Fig. 10).

DISCUSSION

Birth mode is postulated to represent a major factor in shaping earliest gut microbiome colonisation and the linked development of neonates especially in relation to the priming of the neonates' immune system.³ Apart from birth mode, additional aspects such as diet and medical factors have been described to have a significant effect on neonate colonisation and succession.¹⁹ Whether or not such effects persist during the first year of life remains an essential question. Here, we performed an in-depth longitudinal analysis of the gut microbiome using high-resolution metagenomics on samples collected during the first few days *postpartum* through to the first year of age. We specifically assessed the pervasive effect of birth mode-dependent microbiome differences in relation to immune system priming and AMR. In addition, we analysed the contribution of mobile genetic elements and the role of horizontal gene transfer in conferring AMR.

Our previous findings⁸ along with the longer-term trends of the present study underpin the notion that persistent structural and functional differences exist in the gut microbiomes of neonates born by CSD. More specifically, our results agree with other studies

which have highlighted a reduced abundance and colonisation by taxa such as *Bifidobacterium* and *Bacteroides* in CSD neonates.^{6,8,24,55,56} In addition, we found that the levels of *Faecalibacterium prausnitzii* were significantly elevated in VD infants after 1 year of age. *F. prausnitzii* is a highly abundant commensal in the human gut including those with higher levels of diversity and richness when compared to individuals following a Western lifestyle.⁵⁷ Concurrently, this taxon has been reported to be reduced in the gut of patients with ulcerative colitis and Crohn's disease,⁵⁸ which in turn may be linked to it being a keystone taxon conferring anti-inflammatory properties in humans.^{59,60} Further studies are necessary to effectively understand the longer-term consequences of the differential abundance of *F. prausnitzii* in humans beyond the first year of age.

Our previously published study⁸ highlighted a higher potential for LPS-mediated immune priming in VD compared to CSD at day 5 after birth. Conversely, we found that LPS extracts from the CSD samples taken at 1 year of age resulted in significantly higher TNF- α levels compared to 5 days after birth. Our results indicate that a reduction in the earliest immune system priming through key immunogenic molecules occurs in CSD neonates. This might lead to persistent effects throughout the first year of life which, in turn, may explain the higher rates of immune system-linked diseases observed in CSD infants in later life including metabolic disorders^{11,55} and allergies.^{12,13} Along these lines, Jakobsson et al. previously showed that children born via CSD have reduced Th1 responses.⁴ Furthermore, other groups have reported that early-life immune system stimulation impacts immune disorders including asthma,⁶¹ allergies,⁶² diabetes and IBD.⁶³ In this context, our findings indicate that birth mode-dependent gut microbiota alterations affect the status of the immune system throughout the first year of life, and likely beyond. This in turn may explain immunological deficits linked to numerous chronic diseases for which a higher propensity is observed in individuals born by CSD.⁶⁴

Birth mode-associated alterations of the gut microbiota may facilitate colonisation by opportunistic pathogens, including those encoding antimicrobial resistance.¹⁵ Functional analyses of our metagenomic data highlighted enrichments in carbapenem and phenazine biosynthesis genes in the VD group after the first year of life, potentially a consequence of endogenous gut bacteria-mediated resistance mechanisms against opportunistic pathogens in the gut. Both carbapenem and phenazine are known to be bacterial compounds that are used clinically in fighting Gram-positive and Gram-negative pathogens.^{65,66} This data suggests that the indigenous gut microbiota plays a crucial, early role in conferring colonisation resistance against pathogens. In addition, we found that CSD is associated with resistance against semi- and synthetic antibiotics as early as 5 days after birth. It is well established that mothers undergoing CSD are administered antibiotics to prevent nosocomial infections, as a prophylactic policy.^{42,67,68} Interestingly, we did not find resistance towards the antibiotic treatment administered to mothers in our cohort. It, however, remains plausible that the enrichment in AMR genes especially against phenicol, pleuromutilin and diaminopyrimidine classes at day 5 after birth in CSD neonatal samples is linked to the hospital environment including the actual caesarean section.

In conjunction with the observed differences in AMR between CSD and VD our study also highlights the potential mode of AMR transmission via mobile genetic elements including via plasmids and/or bacteriophages.^{69–71} Parnanen et al. reported the presence of AMR genes and MGEs in infant faecal samples at 1 and 6 months of age.⁷² Our findings agree with their results and expand on these by additionally providing data on the abundance of AMR genes and MGEs at 5 days after birth. Furthermore, we identified 27 categories of AMR genes and linked these to both bacterial taxonomy and MGEs. We found that both plasmids and phages encoded genes which confer resistance to several classes

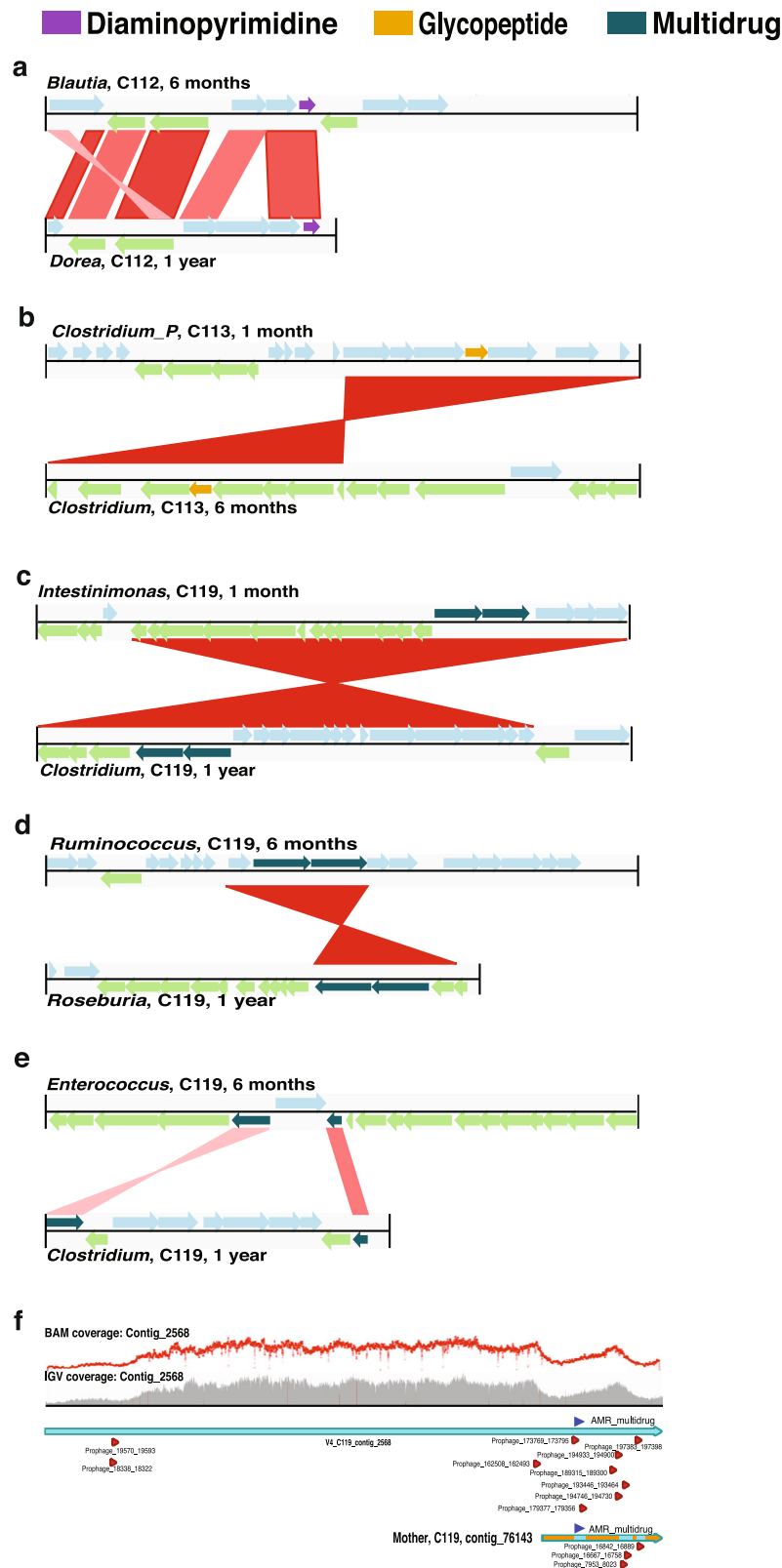


Fig. 6 Horizontal gene transfer (HGT) events. Schematics depicting the **a** diaminopyrimidine resistance gene transfer between *Blautia* and *Dorea* in sample C112 (CSD). The lighter blue and green arrows represent genes localized on the forward and reverse strands, respectively. Red bars indicate matching regions between contigs with darker shades representing higher similarity. **b** HGT event transferring glycopeptide resistance between *Clostridium_P* and *Clostridium* in C113 (CSD). **c–e** Multidrug resistance genes were transferred horizontally between the following genera: *Intestinimonas* and *Clostridium*, *Ruminococcus* and *Roseburia*, *Enterococcus* and *Clostridium*, in sample C119 (CSD). The timepoints at which the genomes were recovered and to which the HGT event corresponds are indicated next to the genera nomenclature. **f** Linear representation of 'contig_2568', which mapped to both bacterial chromosome and phage. Area plot (grey) and dot plot (red) indicating the coverage of the genomic regions. Triangles (red) indicate the prophage sequences (Supplementary Data 4) and their localisation coordinates along the contig, while the dark-blue triangle indicates the location of the multidrug AMR genes ($n = 2$) on the contig.

of antibiotics. Of all MGEs, plasmids conferred resistance to a variety of antimicrobial compounds. Furthermore, we found that glycopeptide and multidrug resistance were transferred via phages, in accordance with previous reports.^{73–75} We also found that horizontal gene transfer plays a critical role in the continued transmission of AMR during the first year of life. While ~88% of HGT occurred via canonical methods involving the bacterial chromosome and plasmids, we found that prophages contributed to multidrug resistance in one CSD sample (C119). We detected prophage signature sequences flanking two multidrug resistance genes, horizontally transferred between bacteria from two distinct orders. Our findings thereby highlight the role of prophages, typically thought to mediate AMR in human pathogens including *Staphylococcus aureus*,⁷⁶ *Salmonella* and shiga toxin-producing *Escherichia coli*,⁷⁷ as mediators of HGT even among commensals. Intriguingly, HGT events in the VD samples did not indicate any AMR gene transmission. On the other hand, considering the smaller sample size, further studies with an increased power are needed to clarify the role of phage-mediated AMR resistance especially during the first few days of life.

The persistence of differences in early-life exposure is an important but challenging research question, not least because of the paucity of long-term, longitudinal studies ranging from immediately after birth until early childhood. Our findings imply that birth mode leads to persistent gut microbiota structural and functional differences. We acknowledge that the limited sample size and the lack of detailed dietary information cannot rule out other confounding factors such as the in utero and postpartum environments of the infants. However, and importantly, our data suggest that gut microbiota structural and functional effects may predispose infants delivered by CSD to delayed immune priming resulting in a deficiency in tolerance. Our results pave the way for future, rational interventions aimed at restoring key functional features of the microbiota. In this context, further studies including following the children over extended periods of time are needed to understand birth mode-mediated manifestations of disease. Concurrently, an important research direction which arises from our study centres on the role of the gut mobilome in conferring AMR and how this affects microbiome trajectories and linked phenotypic outcomes in humans. Considering current global efforts directed at limiting the emergence of antibiotic resistance,⁷⁸ appreciation of the role of phages as an additional source of resistance may be necessary for success in reducing the overall burden of AMR in the future.

METHODS

Ethics statement

All aspects concerning the recruitment and collection of mother–neonate pairs including handling, processing and storing of samples as well as data were approved by the Luxembourg Comité national d'éthique de recherche, under reference number 201110/06 and by the Luxembourg National Commission for Data Protection under reference number A005335/R000058. Prior to specimen collection, following a detailed consultation; written and informed consent was obtained from all mothers enrolled in the study.

Sample collection

Based on our previous study,⁸ the present study design aimed at testing the hypothesis that birth mode elicits longer-term functional microbiome changes which may impact neonatal health and development (with particular foci on antimicrobial resistance and lipopolysaccharide biosynthesis) and we performed the corresponding power analyses using data from our previous study. Founded on the increase in fold-change [caesarean section delivery (CSD) versus vaginal delivery (VD)] in antimicrobial resistance genes, a sample size calculation revealed a minimum number of four individual mother–infant pairs per group to achieve a power of 80% with a significance threshold of 5%. For the LPS-mediated functional cytokine measurements, we estimated a minimum sample size per group

of six pairs based on a fold-change of 1.40x in TNF- α , i.e., a 40% difference of means between the samples (Supplementary Fig. 11). As previously published^{8,39} we found that the functional microbiome differences provide clearer delineations when comparing groups than the typically reported taxonomic profiles. Based on our hypothesis, we focused on functional endpoints, in particular on the emergence and acquisition of AMR genes and the LPS-mediated immune stimulation. As per the results of the power analyses highlighting a minimum requirement of 6 mother–infant pairs per group, we further inflated the per-group sample size by 50% leading to a minimum of 9 mother–infant pairs per group. In the present study, babies delivered via caesarean (CSD, $n = 11$) and vaginal (VD, $n = 9$) deliveries were sampled during the first days of life and were followed-up at 1 month, 6 months and at 1 year of age (Supplementary Data 2). Samples were collected during follow-up visits into sterile plastic vials and immediately flash-frozen in liquid nitrogen. Faecal samples were stored until further processing at -80°C .

Faecal processing and nucleic acid extraction

Genomic DNA was isolated from 50 mg of frozen stool samples aseptically weighed into sterile vials, prior to processing with the DNeasy PowerSoil Kit (Qiagen, Luxembourg) including an additional incubation step at 65°C and milling, as described previously.⁸ All the study samples yielded sufficient DNA for metagenomic sequencing including artefact-curated metagenomic data as described previously⁸ for subsequent analyses. DNA extracted from all timepoints was thereafter stored at -80°C until further use.

DNA sequencing

All DNA samples were subjected to random shotgun sequencing. Briefly, 250 ng of DNA was sheared using Bioruptor NGS (Diagenode, UCD300) with 30-s ON and 30-s OFF for 15 cycles. The sequencing libraries were prepared using TruSeq Nano DNA library preparation kit (Illumina, FC-121-4002) using the protocol provided with the kit. The libraries were prepared considering 350 bp average insert size. Prepared libraries were quantified using Qubit (Invitrogen) and the quality was checked on a Bioanalyzer (Agilent). Sequencing was performed on the NextSeq500 (Illumina) instrument using 2×150 bp read length at the LCSB Sequencing Platform.

Data processing for metagenomics, including genome reconstruction

Paired forward and reverse sequences were processed using the metagenomic workflow of the Integrated Meta-omic Pipeline⁷⁹ (IMP). The metagenomic processing workflow includes pre-processing, assembly, genome reconstruction and functional annotation of genes based on custom databases in reproducible manner. Briefly, the adapter sequences were trimmed in the pre-processing step including the removal of human reads. Thereafter the de novo assembly was performed using the MEGAHIT (version 2.0) assembler.⁸⁰ Default IMP parameters were retained for all samples. Subsequently, we used MetaBAT2⁸¹ and MaxBin2⁸² for binning in addition to an in-house binning methodology previously described.⁸³ This involved ignoring ribosomal RNA sequences in kmer profiles, clustering from VizBin embeddings,⁸⁴ using density-based non-hierarchical clustering algorithms and depth of coverage for genome reconstructions. The reconstructed genomes are hereafter referred to as bins or metagenome-assembled genomes (MAGs). We obtained a non-redundant set of MAGs using DASTool⁸⁵ with a score threshold of 0.7 for downstream analyses.

Metagenomic taxonomic classification, virome and functional analyses

Trimmed and pre-processed read pairs were used as input to determine the microbial abundance and population genomic profiles based on the mOTUs⁸⁶ (version 2) tool. Based on the marker genes in the mOTU2 database taxonomic profiling was performed. The relative abundances of the mOTUs were estimated using a minimum alignment length of 125 basepairs (bp), where the read counts were normalized to the gene length while also accounting for base coverage of the genes. This was done using the *motus profile* option with the built-in option (-c) for relative abundance values per samples. Simultaneously, to improve specificity and minimise false positives, a cut-off of seven genes that deviated from the median was used as an additional parameter to improve both sensitivity and precision. For the reconstructed MAGs, completeness and contamination were determined using CheckM,⁸⁷ while the taxonomy for each MAG was assigned using the GTDB (Genome Taxonomy Database) toolkit (gtdb-tk)⁵⁰

using the *lineage_wf* option and by using the fasta files as inputs for the MAGs.

For the analyses of functional potential from the assembled contigs, open-reading frames were predicted from the assembled contigs using a modified version of Prokka⁸⁸ that includes Prodigal⁸⁹ gene predictions for complete and incomplete open-reading frames. The identified genes were annotated with a hidden Markov models⁹⁰ (HMM) approach, trained using an in-house database⁸⁴ including all KO,⁴¹ TIGRFAM and SWISS-PROT⁹¹ groups and using *hmmsearch* from HMMER 3.1.⁹² Where multiple functional groups were assigned to genes, the best hits based on bit scores were selected. FeatureCounts⁹³ was used to extract the number of reads per functional category, using the arguments -p and -O, thus yielding counts for each functional category. After the LPS-cytokine analysis, insufficient faecal sample for one of the CSD samples (C118) remained for metagenomic sequencing. Therefore, the sample was removed from subsequent metagenomic analyses. For the virome analyses, we used an iterative annotation method to recover microbial (bacterial and archaeal) viruses,⁹⁴ and subsequently taxonomically annotated using a network-based classification protocol defined by Bolduc et al.⁹⁵ Samples C109 was not included in the virome analyses due to viral contigs being below detection confidence thresholds.

Identification of antimicrobial resistance genes and association with mobile genetic elements

We used a deep-learning approach, DeepARG,⁴³ to predict and identify AMR genes within our metagenomic data. The output from Prokka, i.e. the translated fasta sequence files for all open-reading frames, was used as input for the AMR analyses. AMR genes were collapsed into categories based on the Comprehensive Antibiotic Resistance Database (CARD)¹⁴ and identified using DeepARG. Thereafter, the relative abundance of the AMR genes was calculated using the Rnum_Gi method described by Hu et al.⁹⁶

Identified AMR genes and their categories were consecutively linked to associated bacterial taxonomy using the metagenomic bin classification. Furthermore, AMR genes were linked to predicted mobile genetic elements (MGEs; phages and plasmids) to identify probable transmission of AMR between taxa. For the identification of plasmids in the metagenomic data, PlasFlow⁵⁴ was used with a threshold for filtering set to 0.7. Simultaneously, DeepVirFinder⁹⁷ and VirSorter⁹⁸ were used to identify phage sequences within the VD and CSD groups. Predictions from both these tools were subsequently merged to obtain a comprehensive catalogue of phage sequences. For the prediction of phage sequences, the DeepVirFinder thresholds for filtering were set at a *p*-value of < 0.05 and a score of 0.7, while for VirSorter the category 1 and 2 predictions were used for downstream analyses. To link both the MGEs and the taxonomy to the AMR genes, we mapped the genes to assembled contigs, followed by identifying the corresponding bins (MAGs) to which the contigs belonged. By considering all different predictions of MGEs, a final classification was made based on the genomic contexts of the AMR genes encoded on plasmids, phages or chromosomes, including classification of those that could not be resolved (ambiguous). Those AMR genes that could not be assigned to either the MGEs or bacterial chromosomes were further referred to as unclassified genomic signatures. Certain AMR genes were encoded on both the bacterial chromosome and phage genomes (Supplementary Data 4). In such cases, we recorded the encoded AMR gene as being ambiguous. The confirmation of AMR genes and their associated mode of transfer was performed manually alongside the mapping of identical 1 Kbp flanking regions, via the MetaCHIP analyses pipeline.⁹⁹ Briefly, groups of genes among all input MAGs with maximum average identity were considered putative HGT genes. To validate the predicted candidates, a pairwise BLASTN was used to assess each pair of flanking regions of 10 Kbp. Visual representations of the genomic regions were extracted alongside the results for visual interpretation and inspection. Coverage of the genomic regions was additionally assessed through the IGV viewer using the bam file, and manually plotted based on per base coverage statistics for the latter.

LPS isolation and in vitro immunostimulation for cytokine profiling

From the 1 year of age time point, 150 mg faecal samples were weighed aseptically and lipopolysaccharide (LPS) was extracted alongside an extraction blank to serve as a negative control. We also used an in-house pure culture of *E. coli*, from which extracted LPS was used as a positive control. To maximise yields, the samples were divided into triplicates, i.e. 50 mg per vial, prior to LPS extraction using the hot phenol-

water protocol as previously described.⁸ After extractions the triplicates from each sample were pooled and quantified using an endotoxin-detection assay (Endolisa, #609033, Hyglos GmbH, Germany). All samples produced sufficient quantities of LPS. The purified LPS was used to stimulate monocyte-derived dendritic cells (MoDCs). Briefly, primary human monocytes were derived from blood samples from four healthy donors obtained through the Luxembourg Red Cross. The monocytes were further differentiated into MoDCs, in RPMI 1640 medium (ThermoFisher Scientific) supplemented with 10% foetal bovine serum (ThermoFisher Scientific), 20 ng ml⁻¹ of granulocyte-macrophage colony-stimulating factor (PeproTech, London, UK), 20 ng ml⁻¹ IL-4 (PeproTech) and 1% penicillin-streptomycin (Invitrogen). Subsequently, the immunostimulatory potential of the LPS fractions isolated from the 1 year of age faecal samples was determined. For this, MoDCs were treated with LPS extracts from VD and CSD samples. The amount of LPS from each sample that was used to stimulate the MoDCs was adjusted as described by Wampach et al.⁸ Briefly, the MoDCs were stimulated with 7.5 µl/well of LPS while a positive control was established using 15 EU/well LPS isolated from *E. coli*, and a negative control was set up by incubating MoDCs with 7.5 µl/well of the LPS extraction blank. For the in vitro stimulation, the amount of MoDCs was 1 × 10⁵ cells/well. Treatments were performed on cells from all the healthy donor-derived samples, and analysed for the presence of pro- and anti-inflammatory cytokines (TNF-α, IL-8, IL-18, IL-1β, IL-12 and IL-10) using both Human Instant and uncoated ELISA kits (ThermoFisher Scientific).

Data analysis

All figures for the study including visualizations derived from the taxonomic, functional and cytokine profiling were created using version 3.6 of the R statistical software package.¹⁰⁰ DESeq2¹⁰¹ and Wilcoxon rank-sum tests with FDR-adjustments for multiple testing were used to assess significant differences for the AMR and taxonomic analyses whereas a paired two-way ANOVA (analysis of variance) within the *nlme* package was used for identifying statistically significant differences in the cytokine profiles. Volcano plots were generated using the *EnhancedVolcano* package.¹⁰² Corplots were generated using the *corrgram* package developed for R.¹⁰³ The *metacoder*¹⁰⁴ package was used to visualize the AMR-linked taxonomy in R.

DATA AVAILABILITY

The sequencing data and the MAGs generated during the current study are available from NCBI under bioproject accession number PRJNA595749. Supplementary Data 1 lists the taxonomic classifications of the carbapenem biosynthesis KEGG pathway identified within the MAGs. A reporting summary for this article is available as a Supplementary Information file (Supplementary Data 2). Supplementary Data 3 provides the clinical characteristics of all babies including details on diet, antibiotic and growth tracking. A description of the HGT events between reconstructed genomes and the manual validation of phage-conferred AMR is available in Supplementary Data 4, while the accession numbers for the Gasparrini et al.⁴⁵ sequence data are listed in Supplementary Data 5. Legends for the supplementary figures are provided in the Supplementary material file. A description of the AMR and HGT analyses including pre-processing steps along with the scripts and config files can be found at GitLab: <https://git-r3lab.uni.lu/susheel.busi/cosmic2>.

REFERENCES

- Boerma, T. et al. Global epidemiology of use of and disparities in caesarean sections. *The Lancet* **392**, 1341–1348 (2018).
- Betrán, A. P. et al. The increasing trend in Caesarean section rates: global, regional and national estimates: 1990–2014. *PLoS ONE* **11**, e0148343 (2016).
- Gensollen, T. et al. How colonization by microbiota in early life shapes the immune system. *Science* **352**, 539–533 (2016).
- Jakobsson, H. E. et al. Decreased gut microbiota diversity, delayed Bacteroidetes colonisation and reduced Th1 responses in infants delivered by Caesarean section. *Gut* **63**, 559–566 (2014).
- Wang, S. et al. Maternal vertical transmission affecting early-life microbiota development. *Trends Microbiol.* **28**, 28–45 (2019).
- Guittar, J. et al. Trait-based community assembly and succession of the infant gut microbiome. *Nat. Commun.* **10**, 1–11 (2019).
- Sandall, J. et al. Short-term and long-term effects of caesarean section on the health of women and children. *The Lancet* **392**, 1349–1357 (2018).
- Wampach, L. et al. Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. *Nat. Commun.* **9**, 5091 (2018).

9. Korpela, K. et al. Fucosylated oligosaccharides in mother's milk alleviate the effects of caesarean birth on infant gut microbiota. *Sci. Rep.* **8**, 1–7 (2018).
10. Asnicar, F. et al. Studying vertical microbiome transmission from mothers to infants by strain-level metagenomic profiling. *mSystems* **2**, e00164-16 (2017).
11. Bouhanick, B. et al. Mode of delivery at birth and the metabolic syndrome in midlife: the role of the birth environment in a prospective birth cohort study. *BMJ Open* **4**, e005031 (2014).
12. Magne, F. et al. The Elevated Rate of Cesarean Section and Its Contribution to Non-Communicable Chronic Diseases in Latin America: The Growing Involvement of the Microbiota. *Front. Pediatr.* **5**, e192 (2017).
13. Loo, E. X. L. et al. Associations between caesarean delivery and allergic outcomes: Results from the GUSTO study. *Ann. Allergy Asthma Immunol.* **118**, 636–638 (2017).
14. Tamburini, S. et al. The microbiome in early life: implications for health outcomes. *Nat. Med.* **22**, 713–722 (2016).
15. Shao, Y. et al. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**, 117–121 (2019).
16. Stearns, J. C. et al. Intrapartum antibiotics for GBS prophylaxis alter colonization patterns in the early infant gut microbiome of low risk infants. *Sci. Rep.* **7**, 1–9 (2017).
17. Rivas, M. N. et al. The microbiome in asthma. *Curr. Opin. Pediatr.* **28**, 764–771 (2016).
18. Martínez, K. A. et al. Increased weight gain by C-section: functional significance of the primordial microbiome. *Sci. Adv.* **3**, eaao1874 (2017).
19. Bäckhed, F. et al. Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* **17**, 690–703 (2015).
20. Muinck, E. J. de & Trosvik, P. Individuality and convergence of the infant gut microbiota during the first year of life. *Nat. Commun.* **9**, 1–8 (2018).
21. Ferretti, P. et al. Mother-to-infant microbial transmission from different body sites shapes the developing infant gut microbiome. *Cell Host Microbe* **24**, 133–145.e5 (2018).
22. Wopereis, H. et al. The first thousand days—intestinal microbiology of early life: establishing a symbiosis. *Pediatr. Allergy Immunol.* **25**, 428–438 (2014).
23. Baumann-Dudenhoeffer, A. M. et al. Infant diet and maternal gestational weight gain predict early metabolic maturation of gut microbiomes. *Nat. Med.* **24**, 1822–1829 (2018).
24. Bokulich, N. A. et al. Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci. Transl. Med.* **8**, 343ra82 (2016).
25. Mueller, N. T. et al. The infant microbiome development: mom matters. *Trends Mol. Med.* **21**, 109–117 (2015).
26. Stokholm, J. et al. Cesarean section changes neonatal gut colonization. *J. Allergy Clin. Immunol.* **138**, 881–889.e2 (2016).
27. Keoning, J. E. et al. Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl Acad. Sci. USA* **15**, 4578–4585 (2011).
28. Heintz-Buschart, A. & Wilmes, P. Human gut microbiome: function matters. *Trends Microbiol.* **26**, 563–574 (2018).
29. Vatanen, T. et al. Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell* **165**, 842–853 (2016).
30. Jennewein, M. F. et al. Neonate-omics: charting the unknown immune response in early life. *Cell* **174**, 1051–1053 (2018).
31. Romano-Keeler, J. & Weitkamp, J.-H. Maternal influences on fetal microbial colonization and immune development. *Pediatr. Res.* **77**, 189–195 (2015).
32. Spencer, S. J. et al. Early-life immune challenge: defining a critical window for effects on adult responses to immune challenge. *Neuropsychopharmacology* **31**, 1910–1918 (2006).
33. Torow, N. & Hornef, M. W. The neonatal window of opportunity: setting the stage for life-long host-microbial interaction and immune homeostasis. *J. Immunol.* **198**, 557–563 (2017).
34. Gopalakrishna, K. P. et al. Maternal IgA protects against the development of necrotizing enterocolitis in preterm infants. *Nat. Med.* **25**, 1110–1115 (2019).
35. Perez-Muñoz, M. E. et al. A critical assessment of the “sterile womb” and “in utero colonization” hypotheses: implications for research on the pioneer infant microbiome. *Microbiome* **5**, 48 (2017).
36. Olin, A. et al. Stereotypic immune system development in newborn children. *Cell* **174**, 1277–1292.e14 (2018).
37. Levan, S. R. et al. Elevated faecal 12,13-diHOME concentration in neonates at high risk for asthma is produced by gut bacteria and impedes immune tolerance. *Nat. Microbiol.* **4**, 1851–1861 (2019).
38. Ravi, A. et al. The commensal infant gut meta-mobilome as a potential reservoir for persistent multidrug resistance integrons. *Sci. Rep.* **5**, 1–11 (2015).
39. Wampach, L. et al. Colonization and succession within the human gut microbiome by archaea, bacteria, and microeukaryotes during the first year of life. *Front. Microbiol.* <https://doi.org/10.3389/fmicb.2017.00738> (2017).
40. Miquel, S. et al. Faecalibacterium prausnitzii and human intestinal health. *Curr. Opin. Microbiol.* **16**, 255–261 (2013).
41. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
42. Salguero, M. V., Al-Obaide, M. A. I., Singh, R., Siepmann, T. & Vasylyeva, T. L. Dysbiosis of Gram-negative gut microbiota and the associated serum lipopolysaccharide exacerbates inflammation in type 2 diabetic patients with chronic kidney disease. *Exp. Ther. Med.* **18**, 3461–3469 (2019).
43. Arango-Argoty, G. et al. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* **6**, 23 (2018).
44. Alcock, B. P. et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkz935> (2020).
45. Gasparini, A. et al. Persistent metagenomic signatures of early-life hospitalization and antibiotic treatment in the infant gut microbiota and resistome. *Nat. Microbiol.* **4**, 2285–2297 (2019).
46. Fong, I. W. et al. *Antimicrobial Resistance in the 21st Century* (Springer, 2018).
47. Greenwood, D. *Antibiotic and Chemotherapy* 9th edn (Elsevier, 2010).
48. Paukner, S. & Riedl, R. Pleuromutilins: potent drugs for resistant bugs—mode of action and resistance. *Cold Spring Harb. Perspect. Med.* **7**, a027110 (2017).
49. Wright, P. M. et al. The evolving role of chemical synthesis in antibacterial drug discovery. *Angew. Chem. Int. Ed. Engl.* **53**, 8840–8869 (2014).
50. Chaumeil, P.-A. et al. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
51. Martínez-Cano, D. J. et al. Evolution of small prokaryotic genomes. *Front. Microbiol.* **5**, e742 (2015).
52. Woodford, N. et al. Multiresistant Gram-negative bacteria: the role of high-risk clones in the dissemination of antibiotic resistance. *FEMS Microbiol. Rev.* **35**, 736–755 (2011).
53. Gao, N. L. et al. Prokaryotic genome expansion is facilitated by phages and plasmids but impaired by CRISPR. *Front. Microbiol.* **10**, e2254 (2019).
54. Krawczyk, P. S. et al. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.* **46**, e35 (2018).
55. Reyman, M. et al. Impact of delivery mode-associated gut microbiota dynamics on health in the first year of life. *Nat. Commun.* **10**, 1–12 (2019).
56. Dominguez-Bello, M. G. et al. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl Acad. Sci. USA* **107**, 11971–11975 (2010).
57. Rinninella, E. et al. What is the healthy gut microbiota composition? A changing ecosystem across age, environment, diet, and diseases. *Microorganisms* **7**, e14 (2019).
58. Machiels, K. et al. A decrease of the butyrate-producing species *Roseburia hominis* and *Faecalibacterium prausnitzii* defines dysbiosis in patients with ulcerative colitis. *Gut* **63**, 1275–1283 (2014).
59. Sokol, H. et al. *Faecalibacterium prausnitzii* is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. *Proc. Natl Acad. Sci. USA* **105**, 16731–16736 (2008).
60. Rossi, O. et al. *Faecalibacterium prausnitzii* A2-165 has a high capacity to induce IL-10 in human and murine dendritic cells and modulates T cell responses. *Sci. Rep.* **6**, 18507 (2016).
61. Lloyd, C. M. & Hawrylowicz, C. M. Regulatory T cells in asthma. *Immunity* **31**, 438–449 (2009).
62. Vuillermin, P. J. et al. Microbial exposure, interferon gamma gene demethylation in naïve T-cells, and the risk of allergic disease. *Allergy* **64**, 348–353 (2009).
63. Zhuang, L. et al. Intestinal microbiota in early life and its implications on childhood health. *Genomics Proteomics Bioinformatics* **17**, 13–25 (2019).
64. Keag, O. E. et al. Long-term risks and benefits associated with cesarean delivery for mother, baby, and subsequent pregnancies: Systematic review and meta-analysis. *PLoS Med.* **15**, e1002494 (2018).
65. Coulthurst, S. J. et al. Regulation and biosynthesis of carbapenem antibiotics in bacteria. *Nat. Rev. Microbiol.* **3**, 295–306 (2005).
66. Pierson, L. S. & Pierson, E. A. Metabolism and function of phenazines in bacteria: impacts on the behavior of bacteria in the environment and biotechnological processes. *Appl. Microbiol. Biotechnol.* **86**, 1659–1670 (2010).
67. Vangay, P. et al. Antibiotics, pediatric dysbiosis, and disease. *Cell Host Microbe* **17**, 553–564 (2015).
68. Stokholm, J. et al. Prevalence and predictors of antibiotic administration during pregnancy and birth. *PLoS ONE* **8**, e82932 (2013).
69. Gómez-Gómez, C. et al. Infectious phage particles packaging antibiotic resistance genes found in meat products and chicken feces. *Sci. Rep.* **9**, 1–11 (2019).
70. Brown-Jaque, M. et al. Transfer of antibiotic-resistance genes via phage-related mobile elements. *Plasmid* **79**, 1–7 (2015).
71. Cruz, F. de la & Davies, J. Horizontal gene transfer and the origin of species: lessons from bacteria. *Trends Microbiol.* **8**, 128–133 (2000).
72. Pärnänen, K. et al. Maternal gut and breast milk microbiota affect infant gut antibiotic resistome and mobile genetic elements. *Nat. Commun.* **9**, 1–11 (2018).

73. Wang, M. et al. Metagenomic insights into the contribution of phages to antibiotic resistance in water samples related to swine feedlot wastewater treatment. *Front. Microbiol.* **9**, e2474 (2018).
74. Bearson, B. L. et al. The agricultural antibiotic carbadox induces phage-mediated gene transfer in *Salmonella*. *Front. Microbiol.* **5**, e52 (2014).
75. Torres-Barceló, C. The disparate effects of bacteriophages on antibiotic-resistant bacteria. *Emerg. Microbes Infect.* **7**, e168 (2018).
76. Haaber, J. et al. Bacterial viruses enable their host to acquire antibiotic resistance genes from neighbouring cells. *Nat. Commun.* **7**, 1–8 (2016).
77. Colavecchio, A. et al. Bacteriophages contribute to the spread of antibiotic resistance genes among foodborne pathogens of the *Enterobacteriaceae* family—a review. *Front. Microbiol.* **8**, e1108 (2017).
78. Hoffman, S. J. et al. Strategies for achieving global collective action on antimicrobial resistance. *Bull. World Health Organ.* **93**, 867–876 (2015).
79. Narayanasamy, S. et al. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.* **17**, 260 (2016).
80. Li, D. et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinforma. Oxf. Engl.* **31**, 1674–1676 (2015).
81. Kang, D. D. et al. MetaBAT2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
82. Wu, Y. W. et al. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).
83. Heintz-Buschart, A. et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat. Microbiol.* **2**, 16180 (2017).
84. Laczný, C. C. et al. VizBin—an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome* **3**, 1 (2015).
85. Sieber, M. K. C. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
86. Milanese, A. et al. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**, 1–11 (2019).
87. Parks, D. H. et al. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
88. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinforma. Oxf. Engl.* **30**, 2068–2069 (2014).
89. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
90. Yoon, B.-J. Hidden Markov models and their applications in biological sequence analysis. *Curr. Genomics* **10**, 402–415 (2009).
91. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
92. Eddy, S. R. Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
93. Liao, Y. et al. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
94. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).
95. Bolduc, B. et al. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ* **5**, e3243 (2017).
96. Hu, Y. et al. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nat. Commun.* **4**, 1–7 (2013).
97. Ren, J. et al. Identifying viruses from metagenomic data using deep learning. *Quant Biol.* **8**, 64–77 (2020).
98. Roux, S. et al. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
99. Song, W. et al. MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and phylogenetic approaches. *Microbiome* **7**, 36 (2019).
100. Team, R. C. R.: *A Language and Environment for Statistical Computing* (2013).
101. Love, M. I. et al. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, e550 (2014).
102. Blighe, K. et al. EnhancedVolcano: publication-ready volcano plots with enhanced colouring and labeling. *GitHub Repository*, <https://github.com/kevinblighe/EnhancedVolcano> (2019).
103. Chambers, J. M. *Graphical Methods for Data Analysis* (CRC Press, 2018).
104. Foster, Z. S. L. et al. Metacoder: an R package for visualization and manipulation of community taxonomic diversity data. *PLOS Comput. Biol.* **13**, e1005404 (2017).

ACKNOWLEDGEMENTS

We are grateful to all the parents and neonates who participated in the study. We thank the dedicated clinical staff and neonatologists of the paediatric clinic and gynaecologists at the CHL for participant recruitment and sample collection, especially Alain Noirhomme and all involved study nurses of the Clinical and Epidemiological Investigation Centre (CIEC) who performed sample and data collection at the CHL and at home as well as the scientific staff of the IBBL for sample storage. The experiments presented in this paper were carried out using the HPC facilities of the University of Luxembourg. We are thankful for the insightful conversations about AMR assays and immune profiling with Dr. Camille-Martin Gallausiaux, and for the assistance of Audrey Frachet, Lea Grandmougin, Annegret Daujemont and Laura Lebrun (LCSB) for laboratory support. We are also grateful for the feedback on the manuscript by Dr. Cedric Christian Laczný. The present work was partially financed by the Fondation André et Henriette Losch. It was further supported by an ATTRACT programme grant (ATTRACT/A09/03) and CORE programme grants (CORE/15/BM/104040 and CORE/C15/SR/10404839) to P.W. and (CORE Junior/14/BM/8066232) to J.V.F., Aide à la Formation Recherche grants to L.W. (AFR PHD-2013-5824125) and S.N. (AFR PHD-2014-1/7934898), all funded by the Luxembourg National Research Fund (FNR). P.W. acknowledges the European Research Council (ERC-CoG 863664). L.d.N., P.W. and P.M. were supported by the Luxembourg National Research Fund PRIDE17/11823097. S.B.B. was supported by the Synergia grant (CRSII5_180241) through the Swiss National Science Foundation (in collaboration with Dr. Tom Battin at EPFL, Switzerland). A.H.-B. was funded by the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig of the German Research Foundation (FZT 118 - 202548816). Sample collection, processing and storage were co-funded by the Integrated BioBank of Luxembourg under the Personalised Medicine Consortium Diabetes programme.

AUTHOR CONTRIBUTIONS

S.B.B. and L.d.N. participated in the study design and performed the biomolecular extractions, and analyses including cytokine, metagenomic, including taxonomy, functional, AMR and horizontal gene transfer. J.H. along with L.d.N. prepared the monocyte-derived dendritic cells, while J.H. and S.B.B. performed the cytokine ELISAs and subsequent data analyses. L.W. was instrumental in identifying all samples and previously sequenced metagenomic data. J.F. set up immune profiling protocols. P.M. and A.H.-B. were involved in the curation, trimming, assembly and annotation of the metagenomic data. R.H. was instrumental in sequencing the samples for metagenomics. C.d.B. and P.W. conceived the study, participated in its design, were involved in data interpretation and coordinated the study. S.B.B., L.d.N. and P.W. wrote the manuscript. All authors read and approved the final manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s43705-021-00003-5>.

Correspondence and requests for materials should be addressed to P.W.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons

Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix A.4

Mobilome-driven segregation of the resistome in biological wastewater treatment

Mobilome-driven segregation of the resistome in biological wastewater treatment

Laura de Nies¹, Susheel Bhanu Busi¹, Benoit Josef Kunath¹, Patrick May¹ and Paul Wilmes^{1,2,#}

¹Luxembourg Centre for Systems Biomedicine, 7, avenue des Hauts-Fourneaux, Esch-sur-Alzette, L-4362, Luxembourg

²Department of Life Sciences and Medicine, Faculty of Science, Technology and Medicine, University of Luxembourg, 6, avenue du Swing, Belvaux, L-4367, Luxembourg.

[#]Corresponding author: Paul Wilmes; paul.wilmes@uni.lu

Abstract

Biological wastewater treatment plants (BWWTP) are considered to be hotspots of evolution and subsequent spread of antimicrobial resistance (AMR). Mobile genetic elements (MGEs) promote the mobilization and dissemination of antimicrobial resistance genes (ARGs) and are thereby critical mediators of AMR within the BWWTP microbial community. At present, it is unclear whether specific AMR categories are differentially disseminated via bacteriophages (phages) or plasmids. To understand the segregation of AMR in relation to MGEs, we analyzed meta-omic (metagenomic, metatranscriptomic and metaproteomic) data systematically collected over 1.5 years from a BWWTP. Our results showed a core group of fifteen AMR categories which were found across all timepoints. Some of these AMR categories were disseminated exclusively (bacitracin) or primarily (aminoglycoside, MLS, sulfonamide) via plasmids or phages (fosfomycin and peptide), whereas others were disseminated equally by both MGEs. Subsequent expression- and protein-level analyses further demonstrated that aminoglycoside, bacitracin and sulfonamide resistance genes were expressed more by plasmids, in contrast to fosfomycin and peptide AMR expression by phages, thereby validating our genomic findings. Longitudinal assessment further underlined these findings whereby the log2-fold changes of aminoglycoside, bacitracin and sulfonamide resistance genes were increased in plasmids, while fosfomycin and peptide resistance showed similar trends in phages. In the analyzed communities, the dominant taxon *Candidatus* Microthrix parvicella was a major contributor to several AMR categories whereby its plasmids primarily mediated aminoglycoside resistance. Importantly, we also found AMR associated with ESKAPEE pathogens within the BWWTP, for which MGEs also contributed differentially to the dissemination of ARGs. Collectively our findings pave the way towards understanding the segmentation of AMR within MGEs, thereby shedding new light on resistome populations and their mediators, essential elements that are of immediate relevance to human health.

Introduction

Throughout human history, bacterial infections have been a major cause of both disease and mortality¹. The discovery as well as the subsequent development and medical use of antibiotics have provided effective treatment options which limited the development and spread of bacterial pathogens. However, the use of antibiotics has exacerbated the emergence of antimicrobial resistance (AMR) in both commensal and pathogenic bacteria². As a result, AMR, as the "silent pandemic", has become a prevalent threat to human health^{3–5}.

From a public health perspective, biological wastewater treatment plants (BWWTPs) are considered hotspots of AMR due to the convergence of antibiotics with resistant, potentially pathogenic microorganisms originating from both the general population as well as agriculture and healthcare services^{6,7}. Additionally, the mobilization of antimicrobial resistance genes (ARGs) through rampant horizontal gene transfer (HGT) promotes the dissemination of AMR within the BWWTP microbial community⁸. Therefore, BWWTPs represent an environment exceptionally suited for the evolution and subsequent spread of AMR^{9,10}. To date, more than 32 studies have documented the role of BWWTPs as key reservoirs of AMR¹¹. Furthermore, BWWTPs generally do not contain the necessary infrastructure to remove either ARGs or resistant bacteria, which are released into the receiving water via the effluent, promoting its spread in the environment at large¹². Most often these are surface water bodies such as rivers, which contribute to the further dissemination of AMR and resistant bacteria among environmental microorganisms¹³. Acquired resistance may in turn be carried over to humans and animals using these water resources. In fact, there is strong evidence suggesting that ARGs from environmental bacteria can be taken up by human-associated and pathogenic bacteria^{14,15}. From an epidemiological and surveillance perspective, BWWTPs also provide samples representative of entire populations¹⁶. As such, BWWTPs have recently been crucial for the monitoring of SARS-CoV-2 within the human population¹⁷. Overall, to increase our understanding of the dissemination of AMR and the underlying mechanisms as well as its general prevalence, it is necessary to map the resistome of various environments starting with biological BWWTPs because it is critical to unravel the extent to which they act as reservoirs for the dissemination of antimicrobial resistance genes

(ARGs) to bacterial pathogens. Moreover, understanding the community-level overviews of the ARG potential and its expression, coupled with population-level linking, including to pathogens, may allow for efficient monitoring of pathogenic and AMR potential with broad impacts on human health.

The conditions such as the presence of resistance genes¹⁸, and sub-inhibitory antibiotic selection pressure from various sources⁸ facilitate HGT of ARGs into new hosts through the mobilome, i.e. mobile genetic elements (MGEs). Acquisition of ARGs via MGEs primarily occurs through two mechanisms: conjugation or transduction¹⁹. In conjugation, plasmids carrying one or more resistance genes are transferred between microorganisms²⁰, while in transduction bacteriophages carrying ARGs infect bacteria and integrate their genome into those of the host thereby conferring resistance²¹. Of these mechanisms, conjugation is often thought to have the greatest influence on the dissemination of ARGs, while transduction is deemed less important⁸. In general terms, studies concerning AMR and its dissemination focus either on phage^{22,23} or plasmids solely²⁴. Alternatively, the two are treated collectively^{12,25} without a comprehensive comparative analysis. This circumstance has created a knowledge gap whereby the contributions of plasmids and phages as independent entities to AMR transmission within complex communities, such as those found in biological BWWTPs, is largely unknown.

To shed light on the evolution, dissemination and potential segregation of AMR within MGEs in a WTP microbial community, we leveraged longitudinal meta-omics data (metagenomics, metatranscriptomics and metaproteomics). Samples collected for 51 consecutive weeks over a period of 1.5 years, were used to characterize the resistome. We found that several bacterial orders such as Acidimicrobiales, Burkholderiales and Pseudomonadales were associated with 29 AMR categories across all timepoints. Our longitudinal analysis demonstrated that MGEs are important drivers of AMR dissemination within BWWTPs and that assessing the activity of the ARGs is critical for understanding the underlying mechanisms. More importantly, we reveal that MGEs, i.e. plasmidomes and phageomes, contribute differentially to AMR dissemination. Furthermore, we observed this phenomenon in clinically-relevant taxa such as the ESKAPEE pathogens²⁶, for which plasmids and phages were exclusively associated with specific ARGs. Collectively, our data suggest that BWWTPs are critical reservoirs of AMR

which show clear evidence for the segregation of distinct AMR genes within MGEs especially in complex microbial communities. In general, we believe that these findings may provide crucial insights into the segregation of the resistome via the mobilome in any and all reservoirs of AMR, including but not limited to animals, humans, and other environmental systems.

Results

Longitudinal assessment of the resistome within a BWWTP

To characterize the BWWTP resistome, we sampled a municipal BWWTP on a weekly basis over a 1.5 year period (ranging from 21-03-2011 to 03-05-2012)^{27,28}. Utilizing the PathoFact pipeline we resolved the BWWTP resistome. This analysis revealed the presence of 29 different categories of AMR within the BWWTP. Subsequent longitudinal analyses highlighted enrichments in aminoglycoside, beta-lactam and multidrug resistance genes (Fig. 1a). Concomitantly, we observed specific shifts in the AMR profiles over time. For example, a shift at two timepoints (13-05-2011, 08-02-2012) highlighted a steep increase in resistance genes corresponding to glycopeptide resistance. Other AMR categories, such as diaminopyrimidine resistance, exhibited a less drastic but more fluid change in longitudinal abundance observable over multiple timepoints.

Additionally, AMR categories were found to persist variable over time (Fig. 1b). A core group of 15 AMR categories in total were identified and found to be present across the 1.5 year sampling period. These included aminoglycoside, beta-lactam and multidrug resistance genes, which contributed the most to the pool of ARGs. A further six (aminocoumarin, aminoglycoside:aminocoumarin, elfamycin, nucleoside, triclosan and unclassified) AMR categories were found to be prevalent (>75% of all timepoints), while another three AMR categories were moderately (50 - 75% of all timepoints) present over time (Fig. 1b). Five other categories were rarely present within the BWWTP, with resistance corresponding to acridine dye only present at six of the timepoints. Altogether, this emphasized that the BWWTP resistome varies over time, substantiating the

requirement for a longitudinal analysis to obtain an accurate overview of the community's overall resistome.

Although the data thus far provided a clear overview of the BWWTP from a metagenomic perspective, it did not provide any information regarding AMR expression. We therefore utilized the corresponding metatranscriptomic dataset to investigate the expression of identified ARGs and monitor their changes, within the BWWTP, over time. In contrast to the metagenomic data, we observed a difference in AMR expression levels for several categories. Aminoglycoside, beta-lactam, and multidrug resistance identified at high levels in metagenomic information were also highly expressed within the BWWTP (Fig.1c). However, peptide resistance demonstrated the highest expression levels of all the AMR categories. We further investigated which ARG subtypes contributed to the identified peptide resistance category and found that ~90% of the expressed peptide resistance was directly contributed by a single resistance gene, *YojI* (Supp. Fig. 1), typically associated with resistance to microcins²⁹ a potential adaptive strategy amongst the microbial populations in the BWWTP against these specific stressors.

Microbial community and co-occurrence patterns of AMR

Based on the previously identified microbial community²⁸, we hypothesized that the abundant and prevalent bacterial orders such as Acidimicrobiales were major contributors to the abundance in ARGs observable via metagenomics. To further investigate the contribution to AMR by the distinct microbial populations, we linked AMR genes to the contig-based taxonomic annotations of the assemblies. Herein, we identified a wide variety of taxonomic orders contributing to AMR, with multiple orders often contributing to the same resistance categories (Supp. Fig. 2). Overall, taxa belonging to Acidimicrobiales, followed by Burkholderiales, were found to encode most of the ARGs (Fig. 2a). Additionally, the abundance of ARGs linked to taxonomy varied over time. This was most noticeable during a five-week period (autumn: 02-11-2011 to 29-11-2011), where a decrease in abundance in ARGs linked to Acidimicrobiales and Bacteroidales was observed coinciding with an increase in ARG abundance in Pseudomonadales and Lactobacillales.

Since the family Acidimicrobiales was found to be linked to the highest abundance in ARGs, we further resolved the taxonomic affiliation and identified the species *Candidatus* Microthrix parvicella (hereafter known as *M. parvicella*) to be the main contributor to AMR. *M. parvicella* was previously found to dominate this microbial community²⁷ and is a well-characterized bacterium commonly occurring in the BWWTP³⁰. Overall, aminoglycoside, beta-lactam, multidrug and peptide resistance were found to be abundant in this species (Fig. 2b), with aminoglycoside resistance demonstrating the highest expression levels as confirmed through metatranscriptomic analysis (Fig. 2c). Although it was not surprising to find a high abundance of ARGs linked to this species, the longitudinal variation in the abundances of these ARGs was nevertheless surprising (Fig. 2b). Furthermore, coupled to a decrease in the abundance of *M. parvicella* itself²⁷, we observed an almost complete decrease in ARGs at two timepoints (23-11-2011 and 29-22-2011). However, the *M. parvicella* population recovered to levels resembling the earlier timepoints in conjunction with the abundances in ARGs towards the end of the sampling period (Fig. 2a, Fig. 2b), underlining their overall contribution to AMR within this BWWTP. Alternatively, it is plausible that the dominance of *M. parvicella* is attributable to the encoded ARGs, which in turn, may confer a fitness advantage.

In order to determine whether the abundances in ARGs may be directly associated with the community composition and population sizes over time, co-occurrence patterns between ARG subtypes and taxa (genus level) were explored using the metagenomic data. Bipartite network analyses (Fig. 2c) demonstrated that ARGs, within or across ARG types and microbial taxa, showed clear and distinct co-occurrence patterns within the BWWTP. These patterns indicated a strong segregation of distinct, taxa-specific ARG subtypes within the BWWTP community over time. One clear example was that of *M. parvicella* which encoded different aminoglycoside resistance genes (Fig. 3a). Thus, the abundance of this bacterium along with the aminoglycoside ARGs were highly correlated.

Monitoring pathogenic microorganisms within BWWTPs

In conjunction with the families observed within BWWTPs, we also found that certain ESKAPEE pathogens²⁶, such as *Klebsiella* spp. and *Pseudomonas* spp., demonstrated co-occurring patterns with ARGs (Fig. 2c).

As previously mentioned, BWWTPs represent a collection of potentially pathogenic microorganisms originating from, among others, the human population. Moreover, evidence suggests that ARGs from environmental and commensal bacteria can spread to pathogenic bacteria through HGT¹⁹. Therefore, we assessed the acquisition and dissemination of AMR in the extended priority list of pathogens (Table 1), characterized as such by the WHO³¹, using both metagenomics and metatranscriptomics.

Table 1: WHO priority list for research and development of new antibiotics for antibiotics-resistant bacteria³¹.

Bacteria	Priority	Organism detected	Resistance detected
<i>Acinetobacter baumannii</i>	Critical	+	+
<i>Pseudomonas aeruginosa</i>	Critical	+	+
<i>Enterobacteriaceae</i>	Critical	+	+
<i>Enterococcus faecium</i>	high	+	+
<i>Staphylococcus aureus</i>	high	+	+
<i>Helicobacter pylori</i>	high	+	+
<i>Campylobacter</i> spp	high	+	-
<i>Salmonella</i> spp	high	+	+
<i>Neisseria gonorrhoeae</i>	high	+	-
<i>Streptococcus pneumoniae</i>	medium	+	+
<i>Haemophilus influenzae</i>	medium	+	-
<i>Shigella</i> spp	medium	+	+

Of the identified pathogens (Table 1), we found that *Pseudomonas aeruginosa*, both encoded and expressed the highest abundance of ARGs, followed by *Acinetobacter baumannii*, over time within the BWWTP (Fig. 3). Moreover, an increase in ARG abundance and expression was observed in *Pseudomonas aeruginosa* during the time period, during which the otherwise dominant *M. parvicella* demonstrated reduced abundance (Fig. 2b & Fig. 3).

Differential transmission of antimicrobial resistance via mobile genetic elements

As previously described^{32,33}, the mobilome is a major contributor to the dissemination of AMR within a microbial community. Consequently, to understand (i) the role of MGE-mediated AMR transfer within the BWWTP, and (ii) to identify differential contribution of the mobilome to the dissemination of AMR, we identified both plasmids and phages within the metagenome and linked these to the respective ARGs. Overall, we found that plasmids contributed to an average of 10.8% of all ARGs, while phage contributed to an average of 6.8% of all resistance genes, confirming the general hypothesis that conjugation has the greatest influence on the dissemination of ARGs¹⁹. This phenomenon, however, varied across time within the BWWTP (Fig. 4a).

When investigating the dissemination of AMR via MGEs, most reports typically focus on either phages or plasmids individually, or both as collective contributors to transmission³⁴. To date and to our knowledge, the respective contributions of phage and plasmid to AMR transmission have not been subjected to a comprehensive comparative analysis. To facilitate a systematic, comparative view of MGE-mediated AMR, we assessed the segregation of MGEs with respect to AMR and found that phages and plasmids contributed differentially to AMR (Supp. Fig. 3). Specifically, we found a significant difference in six AMR categories when comparing ARGs encoded by phages and plasmids (Fig. 4b). Aminoglycoside, bacitracin, MLS (i.e. macrolide, lincosamide and streptogramin) and sulfonamide resistance were found to be primarily encoded by plasmids, whereas fosfomycin and peptide resistance were found to be associated with phages.

To further understand AMR in relation to the community dynamics, we investigated the abundance and segregation of the above-mentioned significant resistance categories at different timepoints within the BWWTP. We observed ARG abundances varied over time both in phages (Fig. 4c) as well as plasmids (Fig. 4d). For instance, the abundance in aminoglycoside and sulfonamide resistance, which was encoded primarily by plasmids (Supp. Fig. 4a), fluctuated widely over time in both phages and plasmids (Fig. 4c). Additionally, plasmid-mediated sulfonamide resistance was reduced at 23-11-2011, followed by its highest abundance a week later (20-11-2011), while subsequently again

decreasing. Similarly, in line with the above observations, fosfomycin and peptide resistance genes, while segregating within phages, demonstrated significant fluctuations over time (Fig. 4d). In addition to the metagenome, we also contextualized the localization of the expressed ARGs within MGEs based on the metatranscriptomic information. Specifically, we found that plasmids demonstrated a significantly increased expression of aminoglycoside along with bacitracin and sulfonamide resistance genes, while the expression of glycopeptide, mupirocin and peptide resistance genes were primarily enriched in phages (Fig. 5a). These observations pertaining to plasmid-mediated AMR were in line with the metagenomic findings (Fig. 4b). Only peptide resistance was observed to be expressed via phages in contrast to the differential enrichment of fosfomycin resistance observable in the metagenomic data.

Taxonomic affiliations of MGE-derived resistance genes

When assessing the differential contributions of MGEs to AMR, we found congruency between plasmids and phages to the AMR categories and taxonomic affiliations (Fig. 5b). For example, in the metagenomic data MGEs (phage and plasmid) were predominantly associated with the same AMR category and subsequently the same taxa. However, some exceptions were observed with specific taxa associated with AMR either through plasmids or phages. For instance, MLS resistance in Bacteroidales and Nostocales was mediated solely through plasmids, whereas the same resistance category was mediated by phage in Bifidobacteriales, indicating a mechanistic basis for the segregation of AMR between taxa and MGEs.

As most bacteria harbor MGEs, we queried whether the MGE-mediated AMR categories were linked to the abundance of some of the earlier reported taxa. Interestingly, we found that peptide resistance encoded by *M. parvicella* was solely associated with phages, while aminoglycoside resistance was primarily correlated with plasmids (Supp. Fig. 4b). Other highly abundant taxa such as *Pseudomonas* and *Comamonas* (Supp. Fig. 4c-d), on the other hand, were correlated with sulfonamide resistance in addition to aminoglycoside resistance encoded on plasmids (Fig. 5b). This was further reflected within the metatranscriptome data where in taxa such as Acidimicrobiales the expression levels of aminoglycoside resistance were solely

associated with plasmids (Supp. Fig. 5a). Additionally, in the Burkholderiales family, peptide and glycopeptide resistance were found to be expressed through phages (Supp. Fig. 5b).

We also found a clear segregation of the mobilome with respect to individual pathogens in the metagenome. Interestingly, plasmids were exclusively associated with AMR in six out of the fourteen relevant taxa (Fig. 5c). These included *Streptococcus pneumoniae*, *Staphylococcus aureus*, *Shigella flexneri*, *Klebsiella pneumoniae*, *Enterobacter kobei* and *Enterobacter hormaechei*. Furthermore, the plasmids were also associated with conferring peptide, multidrug, MLS, beta-lactam, fluoroquinolone, bacitracin, aminoglycoside, aminoglycoside:aminocumarin and sulfonamide resistance. Phages were exclusively associated with glycopeptide and aminoglycoside resistance in *Salmonella enterica*. Overall, our results revealed for the first time the key segregation patterns of AMR via the mobilome in taxa that are of relevance to human health and disease. Moreover, substantiating the metagenomic data, the pathogenic bacteria *S. pneumoniae*, *S. aureus*, *K. pneumoniae*, *E. kobei* and *E. hormaeche* were found to express ARGs solely associated with plasmids (Fig. 5c). Collectively, these findings represent an imminent threat to global health due to their potential for dissemination across reservoirs.

Metaproteomic validation of AMR abundance and expression

In order to validate our findings with the expression (metatranscriptomic) analyses on the BWWTP, we further used the corresponding metaproteomic data to offer complementary information at the protein level. Similar to the metagenome data we found protein expression linked to aminoglycoside, beta-lactam and multidrug resistance, over time within the BWWTP (Supp. Fig. 6). Proteins linked to multidrug resistance especially were found to increase over time.

To further improve upon the understanding of the AMR expression and assess its stability across the time, we estimated the normalized protein index (NPI) per gene, as discussed in the Methods, by integrating all of the multi-omic data. The estimated NPI demonstrated stable levels of aminoglycoside and multidrug resistance within the BWWTP (Fig. 6a). Specifically, proteins conferring multidrug resistance were found to

increase over time, which is in line with the gene- and expression-level observations. Furthermore, we contextualized the normalized proteins conferring AMR to their localization on MGEs. We identified five resistance categories, i.e. aminoglycoside, beta-lactam, sulfonamide, multidrug and tetracycline resistance, to be expressed through MGEs (Fig 6b). Of these categories we found that aminoglycoside resistance, in concordance with the gene and expression levels, was significantly higher mediated through plasmids compared to phages. We further found that the MGE-mediated AMR categories were associated with specific microbial taxa. with plasmid-mediated aminoglycoside resistance found to be strongly associated with the previously mentioned *M. parvicella* (Fig. 6b). On the other hand, we did not identify any peptides associated with the ESKAPEE pathogens via metaproteomics.

Discussion

The surveillance of wastewaters for the identification of microbial molecular factors is a critical tool for identifying potential pathogens. This has been highlighted recently with the tracking of SARS-CoV-2 within wastewater treatment plants to assess viral prevalence and load within a given community³⁵. Such approaches have also been employed for screening for antimicrobial resistance at a population level^{36,37}. So far, several studies^{10,16,38,39} have characterized the proliferation of ARGs and antibiotic resistant bacteria in BWWTP. Szczepanowski *et al.*³⁸ identified 140 clinically relevant plasmid-derived ARGs in a BWWTP metagenome³⁸ while Parsley *et al.*³⁹ characterized ARGs from bacterial chromosomes, plasmids and in viral metagenomes found in a BWWTP³⁹. Further studies have shown that conventional BWWTP processes at best only partially remove ARGs from the effluent and may find their way into the urban water cycle^{40–42}. Wastewater treatment plants, therefore, are crucial reservoirs of AMR, whose monitoring may allow for early-detection of AMR within the human population feeding into the system. Here, we leveraged a systematic and longitudinal sampling scheme from a BWWTP to identify diverse AMR categories prevalent within the BWWTP microbial community. In line with the studies by Szczepanowski *et al.*³⁸ and Parsley *et al.*³⁹, we found up to 29 AMR categories with several ARGs within the BWWTP. More importantly,

and unlike the previous studies, we linked the identified ARGs to clinically-relevant ESKAPEE pathogens, which represent a growing global threat to human health.

In our BWWTP samples, we identified a core group of 15 AMR categories that were ubiquitous at all timepoints. In line with the above-mentioned reports, the observed core resistance categories may reflect their abundance in the surrounding human population⁴³. This has previously been reported by Pärnänen *et al.*⁴⁴, Su *et al.*⁴⁵ and Hendriksen *et al.*¹⁶ where they showed that BWWTP AMR profiles correlate with clinical antibiotic usage as well as other socio-economic and environmental factors. On the other hand, bacteria are known to have innate defense mechanisms against inhibitory bacteriocins from other taxa⁴⁶. Therefore, one must be cognizant of the phenomenon that the observed core group of AMR categories may also be a proxy for the abundance of specific resistant bacteria. Despite this observation, it is plausible that both anthropogenic and microbial sources for AMR play a role in the observed resistance categories within the BWWTP. Interestingly, we found that several AMR categories, including ancillary (prevalent, moderate, and rare) groups, were associated with *M. parvicella* within the BWWTP. Similar to the findings by Munck *et al.*⁴⁷, we found a wide range of bacteria associated with AMR categories including Acidimicrobiales, Burkholderiales and Rhodocyclales. On the other hand, we report that taxa, including ESKAPEE pathogens, belonging to 25 bacterial orders were associated with 29 AMR categories, compared to the eight bacterial orders reported previously.

It is important to note that the mobilome plays a critical role in the dissemination of AMR within microbial communities. AMR from resistant bacteria within the BWWTP can quickly disseminate within the BWWTP^{11,25}, including transmission from pathogenic to commensal species^{48,49}. As a result, mediated through HGT, the BWWTP becomes a hotspot for resistant bacteria, which are then released back into the receiving environment⁵⁰, and eventually the human population^{11,51}. Therefore, to limit the dissemination of AMR, it is important to understand the role of MGEs within the BWWTP. Our comprehensive analyses identified the differential contributions of AMR transmission mediated via phage and plasmid (Fig. 7). Specifically, we identified clear segregation of aminoglycoside, bacitracin, MLS and sulfonamide resistance categories with plasmids, while fosfomycin and peptide resistance were increasingly encoded and conferred via

phages. While the association between these AMR categories and plasmids^{52–55} or phages⁵⁶ are in line with previously reported results, differential analysis between MGEs has not been previously reported and has not been performed on multi-omic levels. As such, in this study we report for the first time the systematic and extensive comparison of AMR encoded and expressed by phages versus plasmids. Our results indicating the segregation of ARGs within the ESKAPEE taxa via the MGEs further provide insights into potential modes of AMR transmission among pathogens. Though one cannot exclude the possibility of transmission of the above-mentioned ARGs via other MGEs, identifying potential segregation of MGEs in the transmission of ARGs brings us one step closer to identifying specific transmission paths and limiting the spread of AMR. For example, some studies have reported plasmid “curing”, the process by which plasmids are removed from bacterial populations, as a strategy against dissemination of AMR^{57,58}. As described by Buckner *et al.*⁵⁹ plasmid curing, as well as other anti-plasmid strategies, could both reduce AMR prevalence, and (re-)sensitize bacteria to antibiotics⁵⁹. Combining these strategies with AMR categorization according to preference for specific MGEs will give us novel strategies for removing MGE-mediated resistance in the fight against AMR.

By complementing the metagenomic analyses, metatranscriptomics conferred essential information regarding gene expression within the resistome. For instance, when comparing AMR expression levels of aminoglycoside, bacitracin, and sulfonamide mediated via MGEs, it is noticeable that expression levels in plasmids mirror the genomic content, i.e. they exhibited higher levels of expression when compared to phage. On the other hand, glycopeptide and mupirocin resistance genes which were highly expressed in phages were not reflected within the metagenomic data. Additionally, we found the *YojI* resistance gene to be more highly expressed than any other ARGs. To facilitate resistance against the peptide antibiotic microcin J25, the outer membrane protein, TolC, in combination with *YojI* is required to export the antibiotic out of the cell²⁹. Microcin J25 belongs to the group of ribosomally synthesized and post-translationally modified peptides (RiPPs) and has antimicrobial activity against pathogenic genera such as *Salmonella* spp. and *Shigella* spp.⁶⁰. Interestingly, it has only recently been proposed as a treatment option against *Salmonella enterica* and has been discussed in recent years as a potential novel antibiotic⁶¹. Based on these results, by considering that BWWTPs

may reflect both the presence of AMR within the human population as well as be a hotspot of dissemination and generation of new AMR, surveillance of BWWTPs must be emphasized when developing new antibiotics. Our findings collectively suggest that the differential capacity of MGEs to disseminate AMR, coupled with longitudinal and expression-level analyses are crucial for monitoring human health conditions. More importantly, we report for the first time that BWWTP monitoring for AMR may allow for early detection of previously undescribed and previously undescribed resistance mechanisms.

Finally, we applied an integrated multi-omic approach to improve our knowledge on the functional potential of AMR and simultaneously validate the abundance and expression findings of the ARGs. By normalizing the metaproteomic results with the normalized expression of genes we were able to assess the stability of expressed AMR across time. We find that our methodology allows for an unbiased assessment of overall expression accounting for gene copy abundance and expression. These findings support the notion that the AMR genes may serve as sentinels or indicators of the presence of particular antimicrobial agents. However, it is plausible that we are only identifying the most abundant proteins and/or proteins that are more stable over time, and do not capture the entirety of the proteome profiles. Factors such as protein decay rates⁶² among others, may additionally influence this assessment. Irrespective of these observations, we identified segregation of AMR categories with respect to plasmids and phages. Our findings also highlighted the potential for identifying segregation of AMR via specific MGEs with an aim towards possible therapeutic and mitigation strategies via for example plasmid curing. Furthermore, we demonstrate that longitudinal analyses are required to survey AMR within BWWTPs due to the variations in the resistome across time. These shifts may either be representative of a shift within the human population itself, which in turn could be associated with the concurrent use of antibiotics at a given time, or competition within the microbial community. In any case, an independent or static analysis of the various time points may show an incomplete view of the BWWTP resistome, thus underlining the importance of our longitudinal resistome analyses. Overall, our findings suggest that BWWTPs are critical reservoirs of AMR, potentially allowing for early detection and monitoring of pathogens and novel resistance mechanisms linked to the

introduction of new antimicrobials, whilst serving as a model for understanding the separation of MGEs through AMR.

Methods

Sampling and biomolecular extraction

From within the anoxic tank of the Schiffange municipal biological wastewater treatment plant (located in Esch-sur-Alzette, Luxembourg; 49° 30' 48.29" N; 6° 1' 4.53" E) individual floating sludge islets were sampled according to previous described protocols²⁸. Sampling was performed starting on 21-03-21 till 03-05-2012 in approximately one-week intervals resulting in a total of 51 samples. DNA, RNA and proteins were extracted from the samples in a sequential co-isolation procedure as previously described⁶³.

Sequencing and data processing for metagenomics and metatranscriptomics

Paired-end libraries were generated for metagenomics with the AMPure XP/Size Select Buffer Protocol following a size selection step recommended by the standard protocol. Libraries for metatranscriptomics were prepared from RNA after washing stored extractions with ethanol and depletion of rRNAs with the Ribo-Zero Meta-Bacteria rRNA Removal Kit (Epicenter). Subsequently, the ScriptSeq v2 RNA-seq library preparation kit (Epicenter) was used for cDNA library preparation, followed by sequencing on an Illumina Genome Analyses IIx instrument with 100-bps paired-end protocol. Processing and assembly of metagenomic and metatranscriptomic reads was done using the Integrated Meta-omic Pipeline⁶⁴ (IMP v1.3; available at <https://r3lab.uni.lu/web/imp/>). For the IMP processing, Illumina Truseq2 adapters were trimmed, and reads of human origin were filtered out, followed by a de novo assembly with MEGAHIT⁶⁵ v1.0.6. Both metagenomic and metatranscriptomic reads were co-assembled to increase contiguity of the assemblies⁶⁴.

Identification of antimicrobial resistance genes and association with mobile genetic elements

The assembled contigs from IMP were used as input for PathoFact⁶⁶, for the prediction of antimicrobial resistance genes, and to annotate MGEs. ARGs were further collapsed into their respective AMR categories, as identified by PathoFact in accordance with those provided by the Comprehensive Antibiotic Resistance Database (CARD)⁶⁷. Thereafter, the raw read counts per ORF, as given by PathoFact, were determined with featureCounts. The relative abundance of the ARGs was calculated using the RNum_Gi method described by Hu *et al.*⁶⁸ This method was applied to the BAM files generated by mapping using bwa⁶⁹ which were further processed by samtools⁷⁰, both for the metagenomic and metatranscriptomic reads independently, to extract gene copy number and transcriptome expression respectively, per sample.

Identified ARGs and their categories were further linked to associated bacterial taxonomies using the taxonomic classification system Kraken2⁷¹. Kraken2 was run on the contigs using the maxikraken2_1903_140GB (March 2019, 140GB) (https://lomanlab.github.io/mockcommunity/mc_databases.html) database⁷¹. Furthermore, utilizing PathoFact, AMR genes were linked to predicted mobile genetic elements (i.e. plasmids and phages) to track transmission of AMR between taxa. Specifically, to link both the MGEs and the taxonomy to the AMR genes, we mapped the genes to assembled contigs. By considering all different predictions of MGEs, a final classification was made based on the genomic contexts of the AMR genes encoded on plasmids, phages or chromosomes, including classification of those that could not be resolved (ambiguous). The AMR genes that could not be assigned to either the MGEs or bacterial chromosomes were subsequently referred to as *unclassified* genomic elements.

Metaproteomics and data analyses

Raw mass spectrometry files were converted to MGF format using MSconvert⁷² with default parameters. The metaproteomic searches were performed using SearchGUI / PeptideShaker⁷³ for each time point. To generate the databases, each predicted protein sequence file was concatenated with the cRAP database of contaminants (*common*

Repository of Adventitious Proteins, v 2012.01.01; The Global Proteome Machine) and with the human UniProtKB Reference Proteome⁷⁴. In addition, inversed sequences of all protein entries were concatenated to the databases for the estimation of false discovery rates (FDRs). The search was performed using SearchGUI-3.3.20⁷⁵ with the X!Tandem⁷⁶, MS-GF+⁷⁷ and Comet⁷⁸ search engines using the following parameters: Trypsin was used as the digestion enzyme and a maximum of two missed cleavage sites was allowed. The tolerance levels for identification were 10 ppm for MS1 and 15ppm for MS2. Carbamidomethylation of cysteine residues was set as a fixed modification and oxidation of methionines was allowed as variable modification. Peptides with a length between 7 and 60 amino acids and with a charge state composed between +2 and +4 were considered for identification. The results from SearchGUI were merged using PeptideShaker-1.16.45⁷³ and all identifications were filtered in order to achieve a peptide and protein FDR of 1%.

Each predicted protein sequence corresponded to the predicted ORFs generated by the Prodigal (version 2.6.3) predictions included in PathoFact. As such predicted protein sequences matched the ARG annotation of the ORFs as provided by PathoFact.

Multi-omic integration

To further improve upon the understanding of the AMR expression and assess its stability across time, we estimated the normalized protein index (NPI) per gene, by integrating the multi-omic data. To estimate the NPI, we first normalized the metaT abundance based on per gene copy numbers obtained via the metagenomic abundance:

$$NPI = \frac{N_{metaproteome}}{N_{metatranscriptome} / N_{metagenome}}$$

This, the normalized expression of genes, yields the per copy expression of ARGs within each AMR category. Subsequently, the normalized expression was used to standardize the metaP abundances for those genes where the necessary data was available.

MGE partition assessment

To assess the segregation of MGEs through AMR we determined niche regions and overlap using the *nicheROVER* R package⁷⁹. *nicheROVER* uses Bayesian methods to calculate niche regions and pairwise niche overlap using multidimensional niche indicator data (i.e. stable isotopes, environmental variables). As such, using AMR as the indicator data, we extended the application of *nicheRover* to calculate the probability for the size of the niche area of one MGE inside that of the other, and vice versa. We calculated the segregation size estimate for each MGE and additionally generated the posterior distributions of μ (population mean) for each AMR category in all omics. We further computed the niche overlap estimates between MGEs with a 95% confidence interval over 10 000 iterations.

Data analysis

Figures for the study including visualizations derived from the taxonomic and functional analyses were created using version 3.6 of the R statistical software package⁸⁰. A paired two-way ANOVA (Analysis of Variance) within the *nlme* package was used for identifying statistically significant differences for the AMR and taxonomic analyses. Tripartite and Bipartite networks were generated using the *SpiecEasi*⁸¹ R package where a weighted adjacency matrix was generated using the Meinhausen and Buhlmann (*mb*) algorithm, with a $n\lambda$ of 40, and lambda minimum ratio at 0.001. The analyses were bootstrapped with $n=999$ to avoid overfitting, autocorrelations and false network associations. The network was further refined, selecting for positive edges, with a degree greater than the mean-degree of the initial network. The *igraph*⁸² package was used in R to render the graphics for the network. All code for visualization and analysis is available at: https://git-r3lab.uni.lu/laura.denies/lao_scripts.

Data availability

The genomic FASTQ files from this work are publicly available at NCBI BioProject PRJNA230567. Metaproteomic data is publicly available at the PRIDE database under accession number PXD013655.

Code availability

The open-source tools and algorithms used for the data analyses are reported in the Methods section, including relevant flags used for the various tools. Additionally, custom code for further analysis and generation of the figures can be found at:

https://git-r3lab.uni.lu/laura.denies/lao_scripts

Funding

This work was supported by the Luxembourg National Research Fund (FNR) under grant CORE/BM/11333923 and the European Research Council (ERC-CoG 863664) to PW, and PRIDE/11823097 to LdN and PW. SBB was supported by a Synergia grant (CRSII5_180241: Swiss National Science Foundation to Tom Battin at EPFL and PW).

Acknowledgements

We are thankful for the assistance of Audrey Frachet Bour, Lea Grandmougin, Janine Habier, Laura Lebrun (LCSB) for laboratory support. We acknowledge the valuable input from Rashi Halder at the LCSB Sequencing Platform with respect to library preparation. The computational analyses were performed at the HPC facilities at the University of Luxembourg (<https://hpc.uni.lu>)⁸³.

Figures

Figure 1. Longitudinal metagenomic and metatranscriptomic assessment of AMR
a) ARG relative abundances over time within the BWWTP. b) ARG categories at various timepoints categorized in 4 distinct groups based on presence/absence: Core (all timepoints), Prevalent (>75% of timepoints), Moderate (50-75% of timepoints) and Rare (< 50% of all timepoints). c) Relative abundance levels of expressed AMR categories over time within the BWWTP.

Figure 2. Microbial population-linked AMR

a) Longitudinal ARG relative abundance levels linked to their corresponding microbial taxa (order level). b) Relative abundance of ARG categories linked to *Candidatus Microthrix parvicella*. c) Bi-partite network depicting co-occurrence patterns of individual antimicrobial resistance genes (ARGs) and microbial taxa on genus level.

Figure 3. Assessment of AMR associated with clinical pathogens
ARG relative abundance encoded and expressed by clinical pathogens over time within the BWWTP.

Figure 4. MGE-derived AMR within the BWWTP resistome

a) Overall relative abundance of MGEs encoding ARGs. b) Boxplots depicting significant ($adj.p < 0.05$, Two-way ANOVA) differential abundances of ARGs encoded by plasmids vs phages. c) Relative abundance of the 6 significantly different AMR categories encoded on phages over time. d) Relative abundance of the 6 significantly different AMR categories encoded on plasmids over time.

Figure 5 Taxonomic affiliations of MGE-derived resistance genes

a) Boxplot depicting significant differential abundance ($adj.p < 0.05$, Two-way ANOVA) of ARGs expressed in plasmids vs phages. b) Tripartite network assessing the association of MGE-derived ARGs with the microbial taxa. Thickness of the lines representing potential niche-partitioning of the AMR category to one MGE over the other. Color of the line representing which MGE the AMR is linked to: green (phage), blue (plasmid) or black (both phage and plasmid). Asterisk denominates taxonomic orders which include known clinical pathogens. c) Alluvial plot depicting relative abundances of MGE-derived ARGs encoded (metagenome) and/or expressed (metagenome) by clinical pathogens.

Figure 6. Integrative multi-omic assessment of AMR

a) Longitudinal metaproteomic assessment of AMR within the WTP. b) metagenomic and metatranscriptomic normalized protein levels linked to AMR within the WTP over time. c) Tripartite network assessing the normalized protein levels derived from MGEs and associated taxa. Boxplots depicting significant differential ($adj.p < 0.05$, Two-way

ANOVA) abundance of aminoglycoside resistance in plasmid versus phage in *Candidatus* *Microthrix parvicella* as well as overall.

Figure 7. Separation of MGE-derived AMR within the BWWTP.

A graphical summary highlighting AMR categories found significantly increased in phage versus plasmid in all three omes.

Supplementary figure 1. Expression levels of individual ARGs

Expression levels of individual ARGs over time within the BWWTP.

Supplementary figure 2. Taxonomic diversity of AMR

The plot indicates the number of taxa (order level) in which the corresponding AMR categories are identified.

Supplementary figure 3. Partitioning of MGEs through AMR

The boxplots indicate the niche sizes (left) for the MGEs (plasmids and phages) based on metagenomic assessment. Niche plots (right) reveal that plasmids tend to differentiate from phages based on their capacity to encode for aminoglycoside resistance.

Supplementary figure 4. Differential AMR abundance in MGEs

The barplot reports the log2foldchange of AMR categories over time in MGEs (plasmid versus phage) in: a) the general microbial population, b) *M. parvicella*, c) *Pseudomonas* spp. And d) *Comamonas* spp.

Supplementary figure 5. Expression of AMR categories in MGEs

The barplot reports the expression levels of AMR categories over time in MGEs (plasmid versus phage) in: a) Acidimicrobiales, and b) Burkholderiales.

Supplementary figure 6. AMR protein abundances

Barplot depicting protein abundances of various AMR categories over time.

References

1. Bonilla, A. R. & Muniz, K. P. *Antibiotic Resistance: Causes and Risk Factors, Mechanisms and Alternatives*. (Nova Science Publishers, 2009).
2. Wright, G. D. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nat. Rev. Microbiol.* **5**, 175–186 (2007).
3. O'Neill, J. Antimicrobial resistance: tackling a crisis for the health and wealth of nations. *Review on antimicrobial resistance* (2014).
4. Brogan, D. M. & Mossialos, E. A critical analysis of the review on antimicrobial resistance report and the infectious disease financing facility. *Global. Health* **12**, 8 (2016).
5. Mahoney, A. R., Safaee, M. M., Wuest, W. M. & Furst, A. L. The silent pandemic: Emergent antibiotic resistances following the global response to SARS-CoV-2. *iScience* **24**, 102304 (2021).
6. Rodríguez-Molina, D. *et al.* Do wastewater treatment plants increase antibiotic resistant bacteria or genes in the environment? Protocol for a systematic review. *Syst. Rev.* **8**, 304 (2019).
7. Alexander, J., Hembach, N. & Schwartz, T. Evaluation of antibiotic resistance dissemination by wastewater treatment plant effluents with different catchment areas in Germany. *Sci. Rep.* **10**, 8952 (2020).
8. von Wintersdorff, C. J. H. *et al.* Dissemination of Antimicrobial Resistance in Microbial Ecosystems through Horizontal Gene Transfer. *Front. Microbiol.* **7**, 173 (2016).
9. Chen, B. *et al.* Metagenomic profiles of antibiotic resistance genes (ARGs) between human impacted estuary and deep ocean sediments. *Environ. Sci. Technol.* **47**, 12753–12760 (2013).
10. Calero-Cáceres, W. *et al.* Sludge as a potential important source of antibiotic resistance genes in both the bacterial and bacteriophage fractions. *Environ. Sci. Technol.* **48**, 7602–

- 7611 (2014).
11. Fouz, N. *et al.* The Contribution of Wastewater to the Transmission of Antimicrobial Resistance in the Environment: Implications of Mass Gathering Settings. *Trop Med Infect Dis* **5**, (2020).
12. Alexander, J., Hembach, N. & Schwartz, T. Evaluation of antibiotic resistance dissemination by wastewater treatment plant effluents with different catchment areas in Germany. *Sci. Rep.* **10**, 8952 (2020).
13. Singer, A. C., Shaw, H., Rhodes, V. & Hart, A. Review of Antimicrobial Resistance in the Environment and Its Relevance to Environmental Regulators. *Front. Microbiol.* **7**, 1728 (2016).
14. Nadeem, S. F. *et al.* Antimicrobial resistance: more than 70 years of war between humans and bacteria. *Crit. Rev. Microbiol.* **46**, 578–599 (2020).
15. Trinh, P., Zaneveld, J. R., Safraneck, S. & Rabinowitz, P. M. One Health Relationships Between Human, Animal, and Environmental Microbiomes: A Mini-Review. *Front Public Health* **6**, 235 (2018).
16. Hendriksen, R. S. *et al.* Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nat. Commun.* **10**, 1124 (2019).
17. Herold, M. *et al.* Genome Sequencing of SARS-CoV-2 Allows Monitoring of Variants of Concern through Wastewater. *Water* **13**, 3018 (2021).
18. Tennstedt, T., Szczepanowski, R., Braun, S., Pühler, A. & Schlüter, A. Occurrence of integron-associated resistance gene cassettes located on antibiotic resistance plasmids isolated from a wastewater treatment plant. *FEMS Microbiol. Ecol.* **45**, 239–252 (2003).
19. MacLean, R. C. & San Millan, A. The evolution of antibiotic resistance. *Science* **365**, 1082–1083 (2019).
20. Carattoli, A. Plasmids and the spread of resistance. *Int. J. Med. Microbiol.* **303**, 298–304 (2013).

21. Chiang, Y. N., Penadés, J. R. & Chen, J. Genetic transduction by phages and chromosomal islands: The new and noncanonical. *PLoS Pathog.* **15**, e1007878 (2019).
22. Lood, R., Ertürk, G. & Mattiasson, B. Revisiting Antibiotic Resistance Spreading in Wastewater Treatment Plants - Bacteriophages as a Much Neglected Potential Transmission Vehicle. *Front. Microbiol.* **8**, 2298 (2017).
23. Strange, J. E. S., Leekitcharoenphon, P., Møller, F. D. & Aarestrup, F. M. Metagenomics analysis of bacteriophages and antimicrobial resistance from global urban sewage. *Sci. Rep.* **11**, 1600 (2021).
24. Li, Q., Chang, W., Zhang, H., Hu, D. & Wang, X. The Role of Plasmids in the Multiple Antibiotic Resistance Transfer in ESBLs-Producing *Escherichia coli* Isolated From Wastewater Treatment Plants. *Front. Microbiol.* **10**, 633 (2019).
25. Che, Y. *et al.* Mobile antibiotic resistome in wastewater treatment plants revealed by Nanopore metagenomic sequencing. *Microbiome* **7**, 44 (2019).
26. Reza, A., Sutton, J. M. & Rahman, K. M. Effectiveness of Efflux Pump Inhibitors as Biofilm Disruptors and Resistance Breakers in Gram-Negative (ESKAPEE) Bacteria. *Antibiotics (Basel)* **8**, (2019).
27. Martínez Arbas, S. *et al.* Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics. *Nat Microbiol* **6**, 123–135 (2021).
28. Herold, M. *et al.* Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance. *Nat. Commun.* **11**, 5281 (2020).
29. Delgado, M. A., Vincent, P. A., Farías, R. N. & Salomón, R. A. Yojl of *Escherichia coli* functions as a microcin J25 efflux pump. *J. Bacteriol.* **187**, 3465–3470 (2005).
30. Calusinska, M. *et al.* A year of monitoring 20 mesophilic full-scale bioreactors reveals the existence of stable but different core microbiomes in bio-waste and wastewater anaerobic digestion systems. *Biotechnol. Biofuels* **11**, 1–19 (2018).

31. Tacconelli, E. *et al.* Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect. Dis.* **18**, 318–327 (2018).
32. Beceiro, A., Tomás, M. & Bou, G. Antimicrobial resistance and virulence: a successful or deleterious association in the bacterial world? *Clin. Microbiol. Rev.* **26**, 185–230 (2013).
33. Wee, B. A., Muloi, D. M. & van Bunnik, B. A. D. Quantifying the transmission of antimicrobial resistance at the human and livestock interface with genomics. *Clin. Microbiol. Infect.* **26**, 1612–1616 (2020).
34. Slizovskiy, I. B., Mukherjee, K., Dean, C. J., Boucher, C. & Noyes, N. R. Mobilization of Antibiotic Resistance: Are Current Approaches for Colocalizing Resistomes and Mobilomes Useful? *Front. Microbiol.* **11**, 1376 (2020).
35. Westhaus, S. *et al.* Detection of SARS-CoV-2 in raw and treated wastewater in Germany - Suitability for COVID-19 surveillance and potential transmission risks. *Sci. Total Environ.* **751**, 141750 (2021).
36. Kwak, Y.-K. *et al.* Surveillance of antimicrobial resistance among *Escherichia coli* in wastewater in Stockholm during 1 year: does it reflect the resistance trends in the society? *Int. J. Antimicrob. Agents* **45**, 25–32 (2015).
37. Reinthaler, F. F. *et al.* Resistance patterns of *Escherichia coli* isolated from sewage sludge in comparison with those isolated from human patients in 2000 and 2009. *J. Water Health* **11**, 13–20 (2013).
38. Szczepanowski, R. *et al.* Detection of 140 clinically relevant antibiotic-resistance genes in the plasmid metagenome of wastewater treatment plant bacteria showing reduced susceptibility to selected antibiotics. *Microbiology* **155**, 2306–2319 (2009).
39. Parsley, L. C. *et al.* Identification of diverse antimicrobial resistance determinants carried on bacterial, plasmid, or viral metagenomes from an activated sludge microbial assemblage. *Appl. Environ. Microbiol.* **76**, 3753–3757 (2010).

40. Hiller, C. X., Hübner, U., Fajnorova, S., Schwartz, T. & Drewes, J. E. Antibiotic microbial resistance (AMR) removal efficiencies by conventional and advanced wastewater treatment processes: A review. *Sci. Total Environ.* **685**, 596–608 (2019).
41. Proia, L. *et al.* Occurrence and persistence of carbapenemases genes in hospital and wastewater treatment plants and propagation in the receiving river. *J. Hazard. Mater.* **358**, 33–43 (2018).
42. Rodriguez-Mozaz, S. *et al.* Occurrence of antibiotics and antibiotic resistance genes in hospital and urban wastewaters and their impact on the receiving river. *Water Res.* **69**, 234–242 (2015).
43. Aarestrup, F. M. & Woolhouse, M. E. J. Using sewage for surveillance of antimicrobial resistance. *Science* **367**, 630–632 (2020).
44. Pärnänen, K. M. M. *et al.* Antibiotic resistance in European wastewater treatment plants mirrors the pattern of clinical antibiotic resistance prevalence. *Sci Adv* **5**, eaau9124 (2019).
45. Su, J.-Q. *et al.* Metagenomics of urban sewage identifies an extensively shared antibiotic resistome in China. *Microbiome* **5**, 84 (2017).
46. Frost, I. *et al.* Cooperation, competition and antibiotic resistance in bacterial colonies. *ISME J.* **12**, 1582–1593 (2018).
47. Munck, C. *et al.* Limited dissemination of the wastewater treatment plant core resistome. *Nat. Commun.* **6**, 8452 (2015).
48. Blake, D. P., Hillman, K., Fenlon, D. R. & Low, J. C. Transfer of antibiotic resistance between commensal and pathogenic members of the Enterobacteriaceae under ileal conditions. *J. Appl. Microbiol.* **95**, 428–436 (2003).
49. Brinkac, L., Voorhies, A., Gomez, A. & Nelson, K. E. The Threat of Antimicrobial Resistance on the Human Microbiome. *Microb. Ecol.* **74**, 1001–1008 (2017).
50. Turolla, A., Cattaneo, M., Marazzi, F., Mezzanotte, V. & Antonelli, M. Antibiotic resistant bacteria in urban sewage: Role of full-scale wastewater treatment plants on environmental

spreading. *Chemosphere* **191**, 761–769 (2018).

51. Newton, R. J. & McClary, J. S. The flux and impact of wastewater infrastructure microorganisms on human and ecosystem health. *Curr. Opin. Biotechnol.* **57**, 145–150 (2019).
52. Dubnau, D. *et al.* Regulation of Plasmid Specified MLS-Resistance in *Bacillus subtilis* by Conformational Alteration of RNA Structure. in *Molecular Biology, Pathogenicity, and Ecology of Bacterial Plasmids* (eds. Levy, S. B., Clowes, R. C. & Koenig, E. L.) 157–167 (Springer US, 1981).
53. Galimand, M., Courvalin, P. & Lambert, T. Plasmid-mediated high-level resistance to aminoglycosides in Enterobacteriaceae due to 16S rRNA methylation. *Antimicrob. Agents Chemother.* **47**, 2565–2571 (2003).
54. Han, X. *et al.* Functional Analysis of a Bacitracin Resistance Determinant Located on ICECp1, a Novel Tn916-Like Element from a Conjugative Plasmid in *Clostridium perfringens*. *Antimicrob. Agents Chemother.* **59**, 6855–6865 (2015).
55. Razavi, M. *et al.* Discovery of the fourth mobile sulfonamide resistance gene. *Microbiome* **5**, 160 (2017).
56. Torres-Barceló, C. The disparate effects of bacteriophages on antibiotic-resistant bacteria. *Emerg. Microbes Infect.* **7**, 168 (2018).
57. Vrancianu, C. O., Popa, L. I., Bleotu, C. & Chifiriuc, M. C. Targeting Plasmids to Limit Acquisition and Transmission of Antimicrobial Resistance. *Front. Microbiol.* **11**, 761 (2020).
58. Bouanchaud, D. H. & Chabbert, Y. A. The problems of drug-resistant pathogenic bacteria. Practical effectiveness of agents curing R factors and plasmids. *Ann. N. Y. Acad. Sci.* **182**, 305–311 (1971).
59. Buckner, M. M. C., Ciusa, M. L. & Piddock, L. J. V. Strategies to combat antimicrobial resistance: anti-plasmid and plasmid curing. *FEMS Microbiol. Rev.* **42**, 781–804 (2018).
60. Naimi, S. *et al.* Fate and Biological Activity of the Antimicrobial Lasso Peptide Microcin J25

Under Gastrointestinal Tract Conditions. *Front. Microbiol.* **9**, 1764 (2018).

61. Ben Said, L. *et al.* Phenomic and genomic approaches to studying the inhibition of multiresistant *Salmonella enterica* by microcin J25. *Environ. Microbiol.* **22**, 2907–2920 (2020).

62. Cameron, D. E. & Collins, J. J. Tunable protein degradation in bacteria. *Nat. Biotechnol.* **32**, 1276–1281 (2014).

63. Roume, H., Heintz-Buschart, A., Muller, E. E. L. & Wilmes, P. Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample. *Methods Enzymol.* **531**, 219–236 (2013).

64. Narayanasamy, S. *et al.* IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.* **17**, 260 (2016).

65. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

66. de Nies, L. *et al.* PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. *Microbiome* **9**, 49 (2021).

67. Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**, D517–D525 (2020).

68. Hu, Y. *et al.* Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nat. Commun.* **4**, 2151 (2013).

69. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

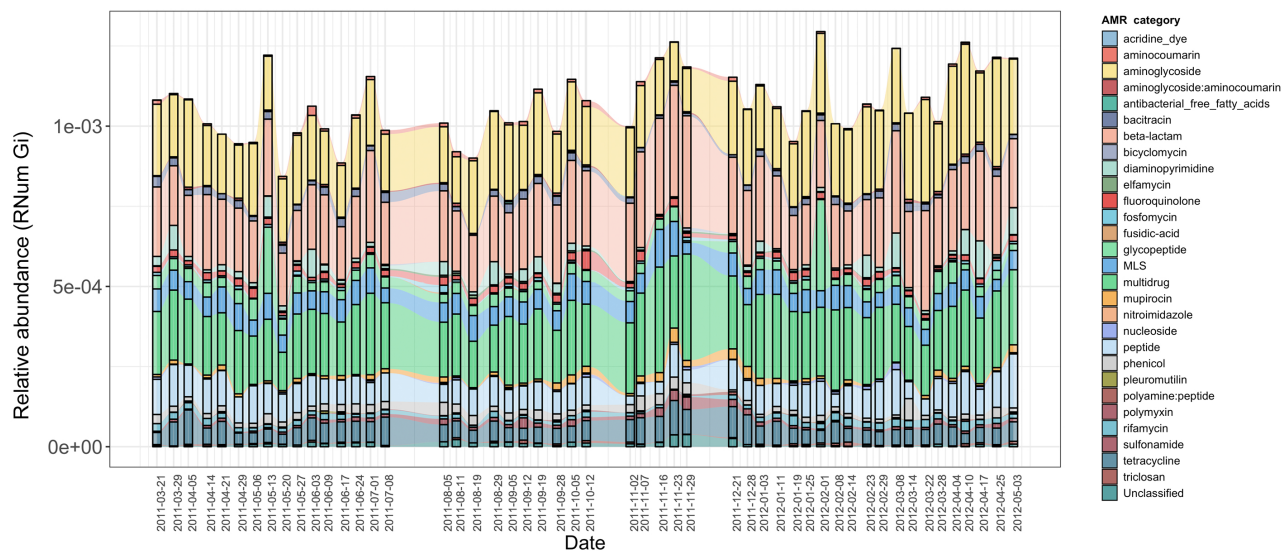
70. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

71. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

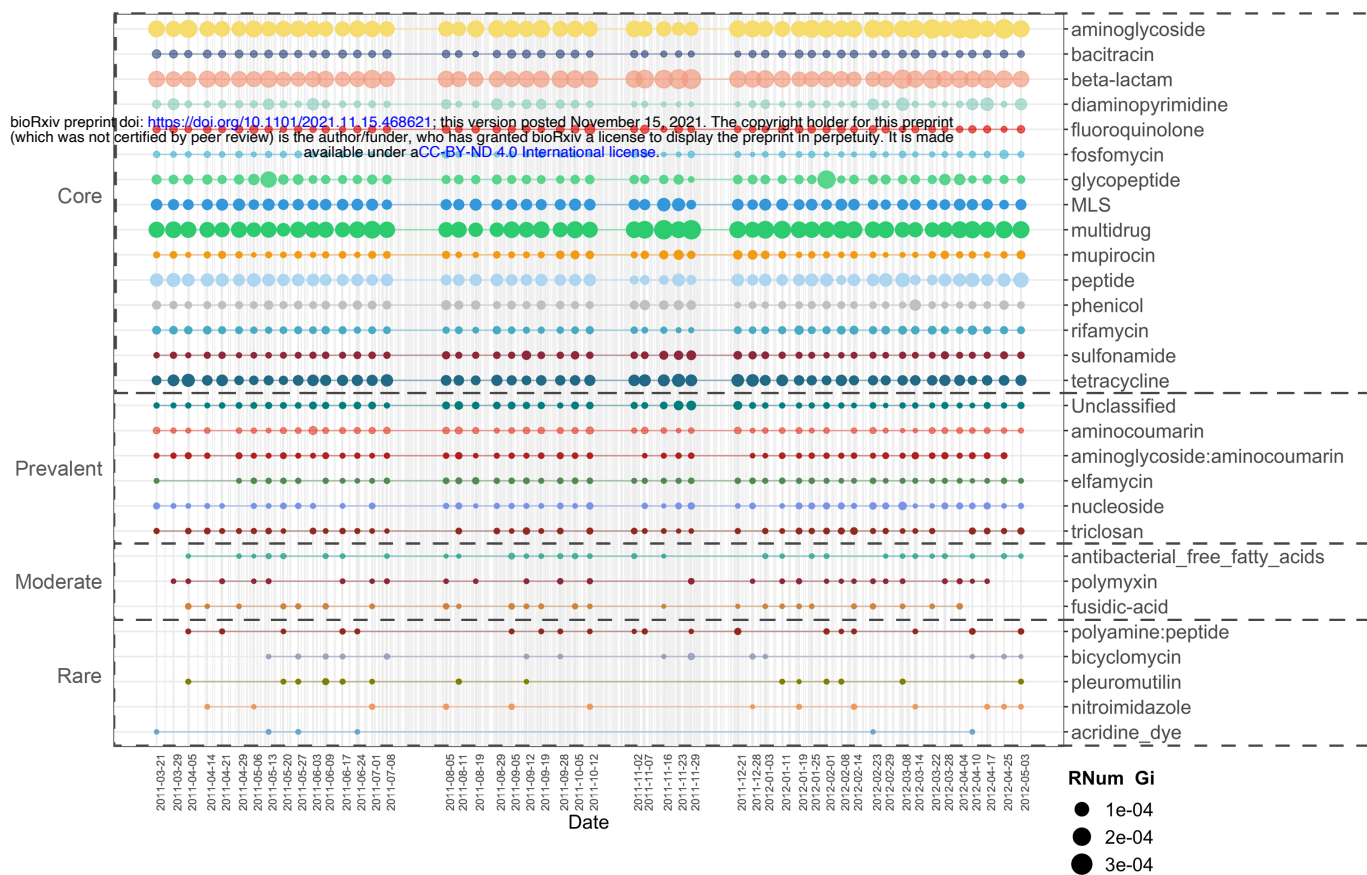
72. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **30**, 918–920 (2012).
73. Vaudel, M. *et al.* PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* **33**, 22–24 (2015).
74. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
75. Barsnes, H. & Vaudel, M. SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. *J. Proteome Res.* **17**, 2552–2555 (2018).
76. Langella, O. *et al.* X!TandemPipeline: A Tool to Manage Sequence Redundancy for Protein Inference and Phosphosite Identification. *J. Proteome Res.* **16**, 494–503 (2017).
77. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277 (2014).
78. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).
79. Swanson, H. K. *et al.* A new probabilistic method for quantifying n-dimensional ecological niches and niche overlap. *Ecology* **96**, 318–324 (2015).
80. Team, R. C. & Others. R: A language and environment for statistical computing. (2013).
81. Kurtz, Z. D. *et al.* Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* **11**, e1004226 (2015).
82. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Systems* **1695**, 1–9 (2006).
83. Varrette, S., Bouvry, P., Cartiaux, H. & Georgatos, F. Management of an academic HPC cluster: The UL experience. in *2014 International Conference on High Performance Computing Simulation (HPCS)* 959–967 (2014).

Figure 1

a



b



c

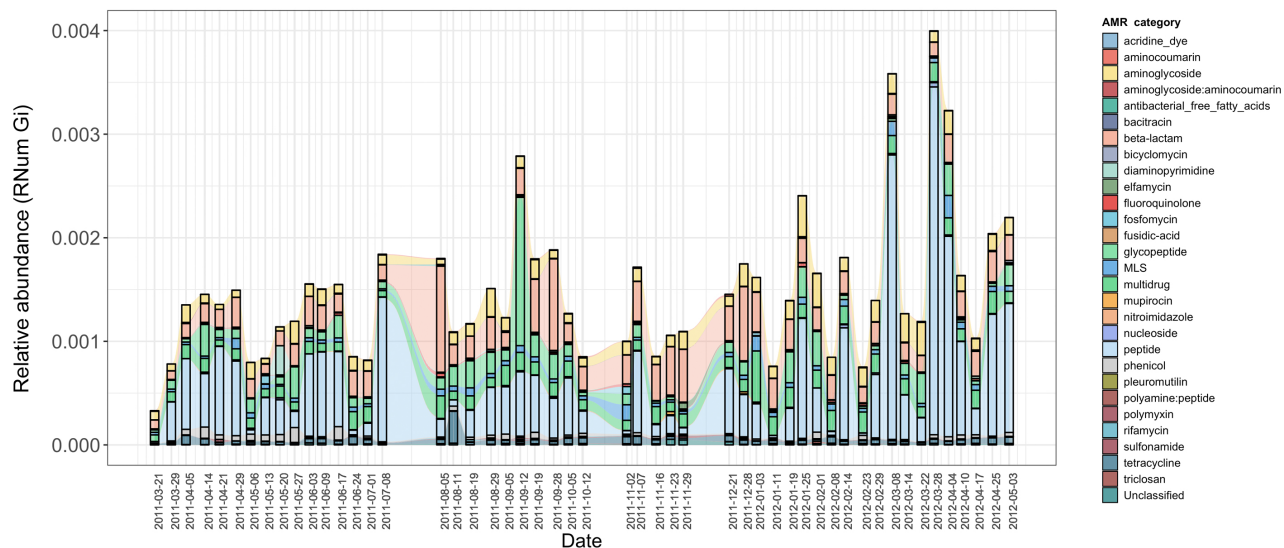


Figure 2

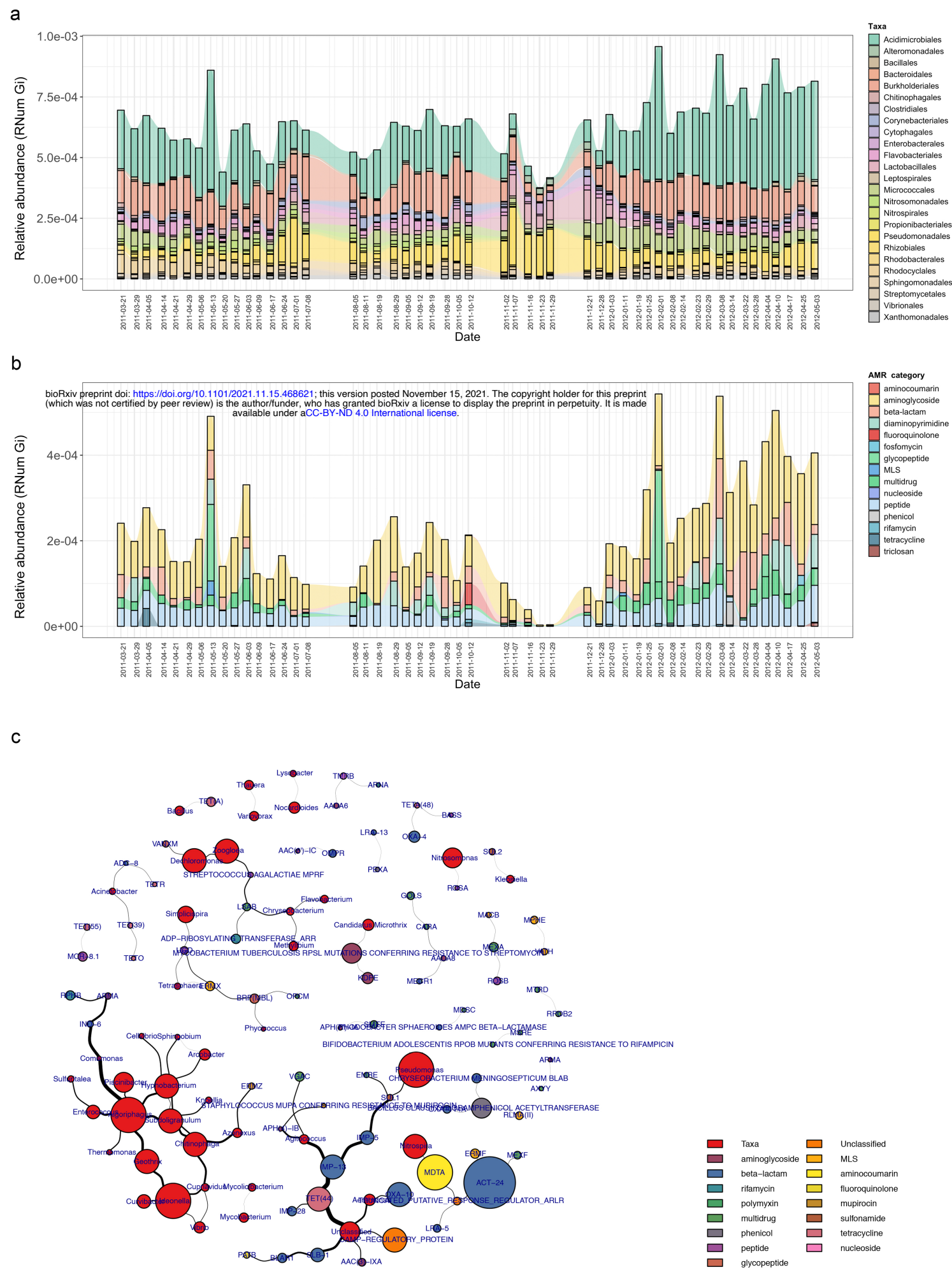


Figure 3

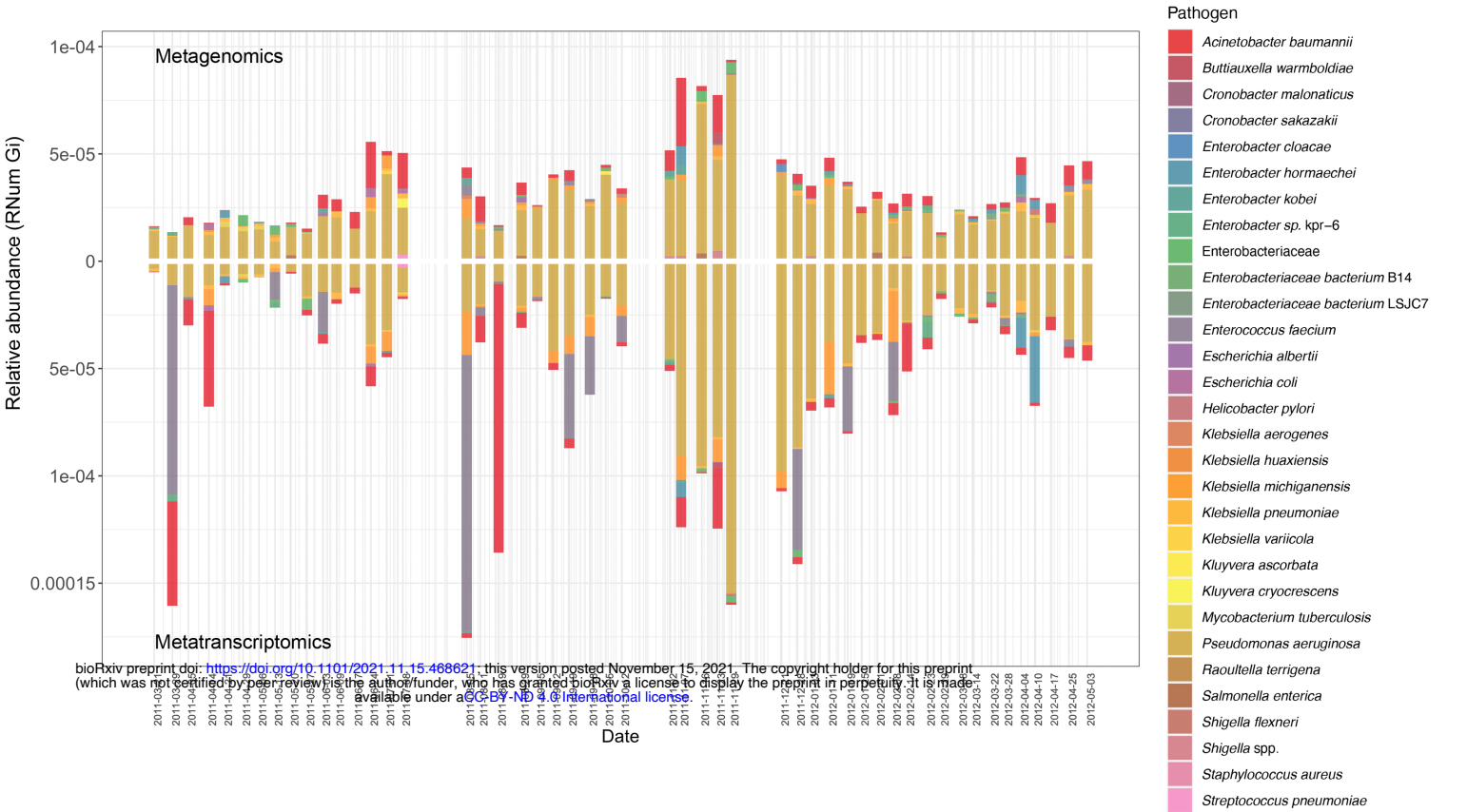
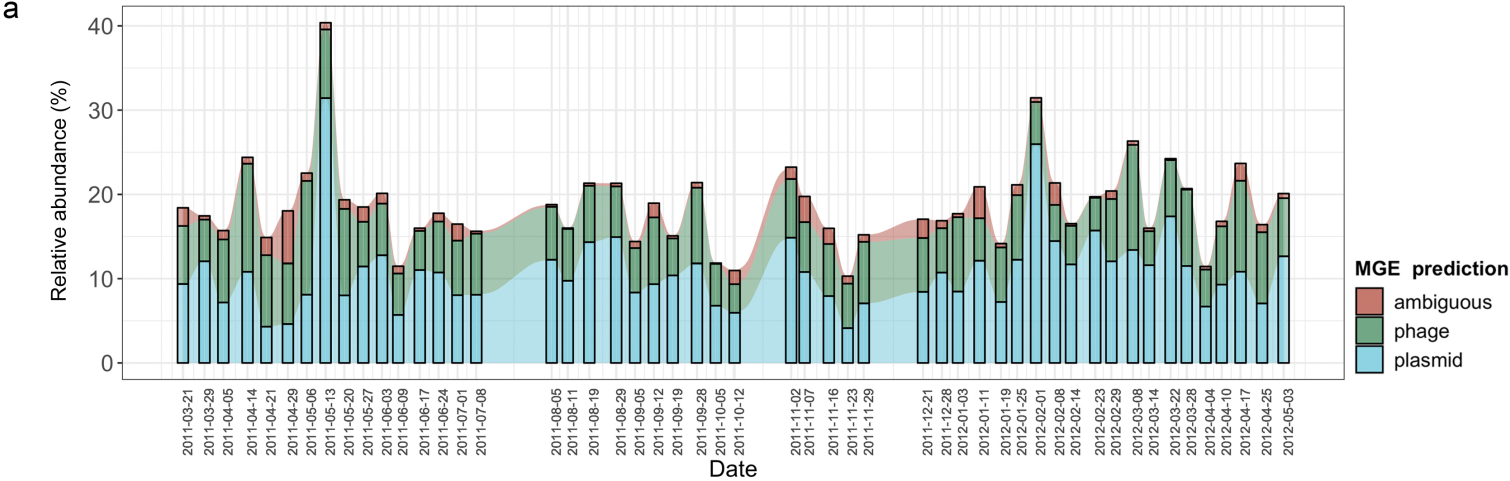
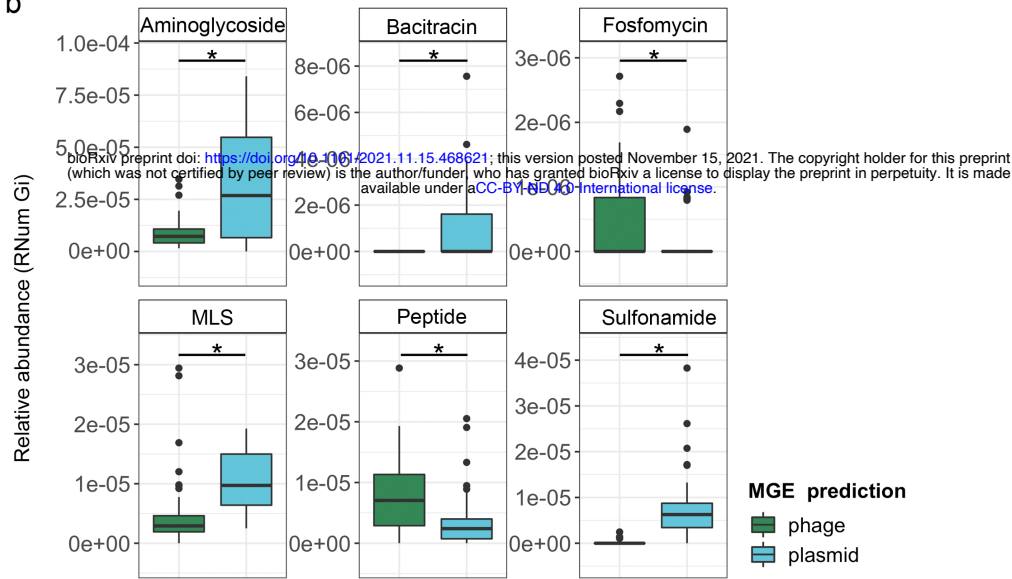


Figure 4

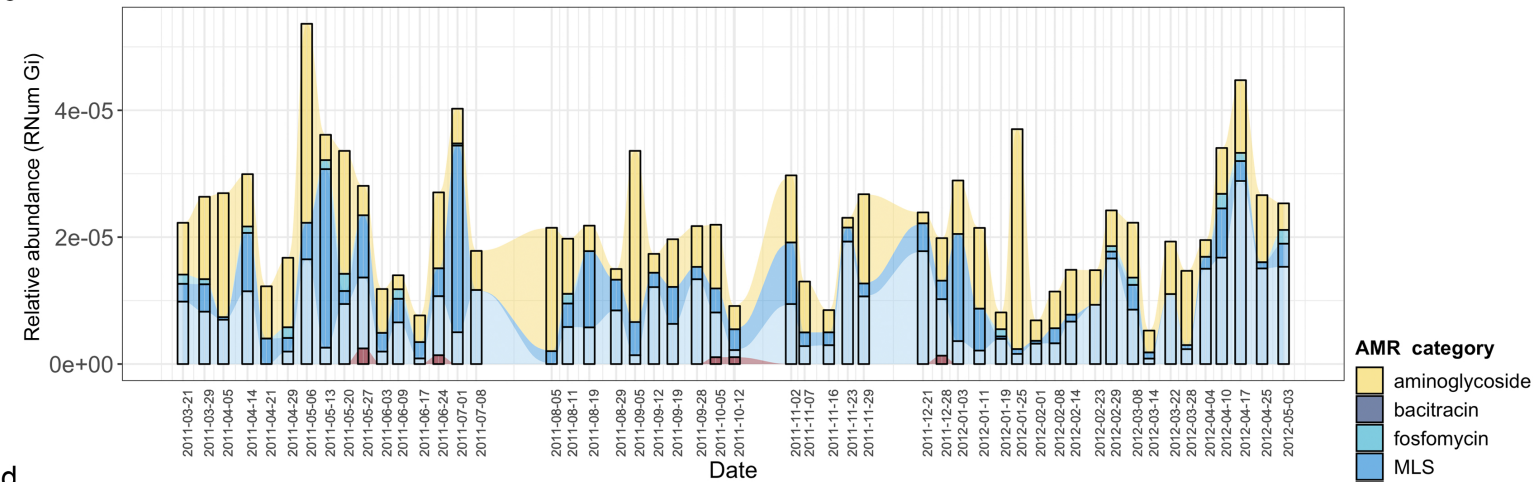
a



b



c



d

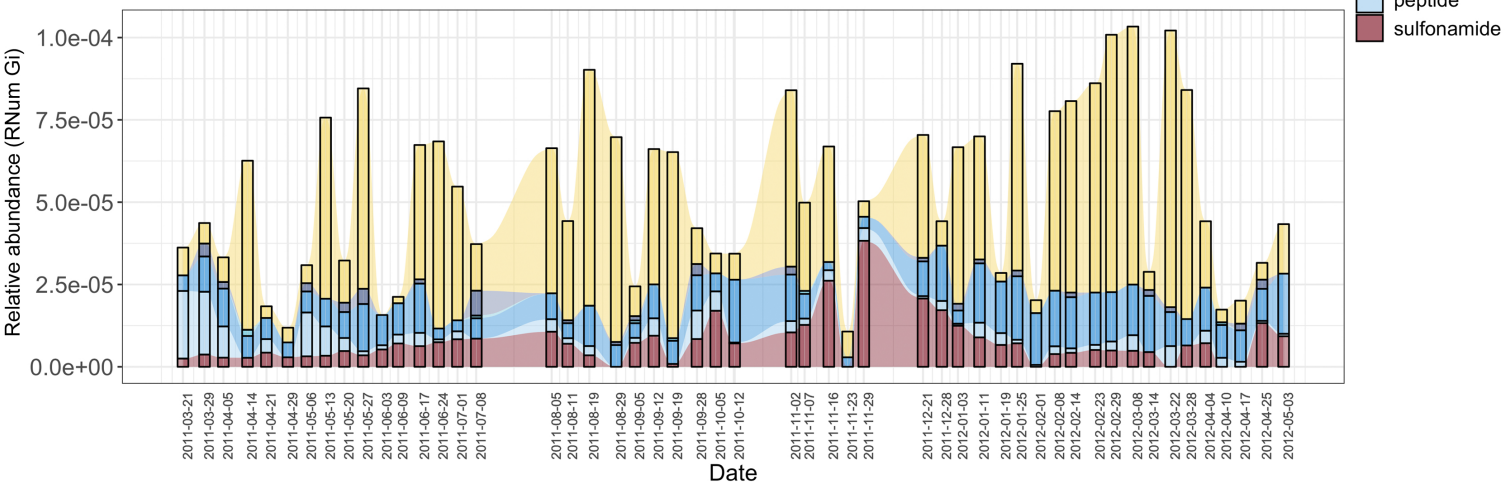
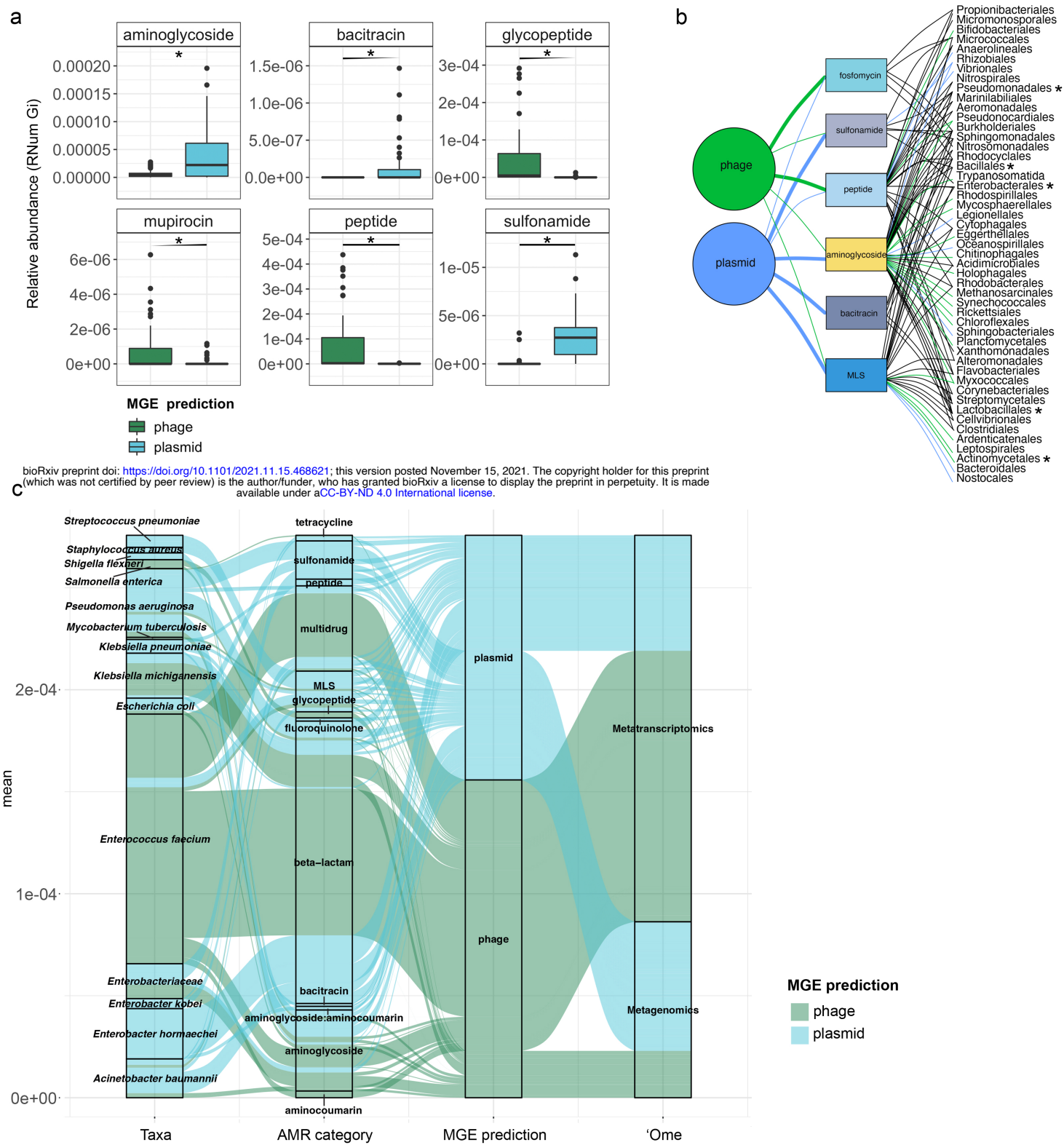
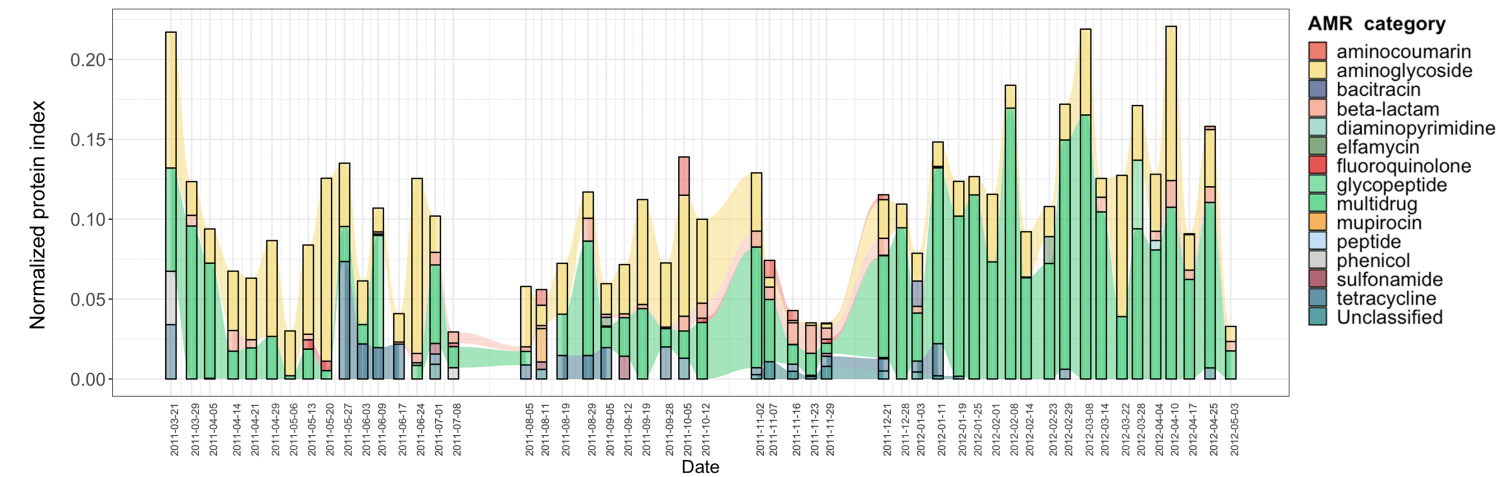


Figure 5



bioRxiv preprint doi: <https://doi.org/10.1101/2021.11.15.468621>; this version posted November 15, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure 6
a



b

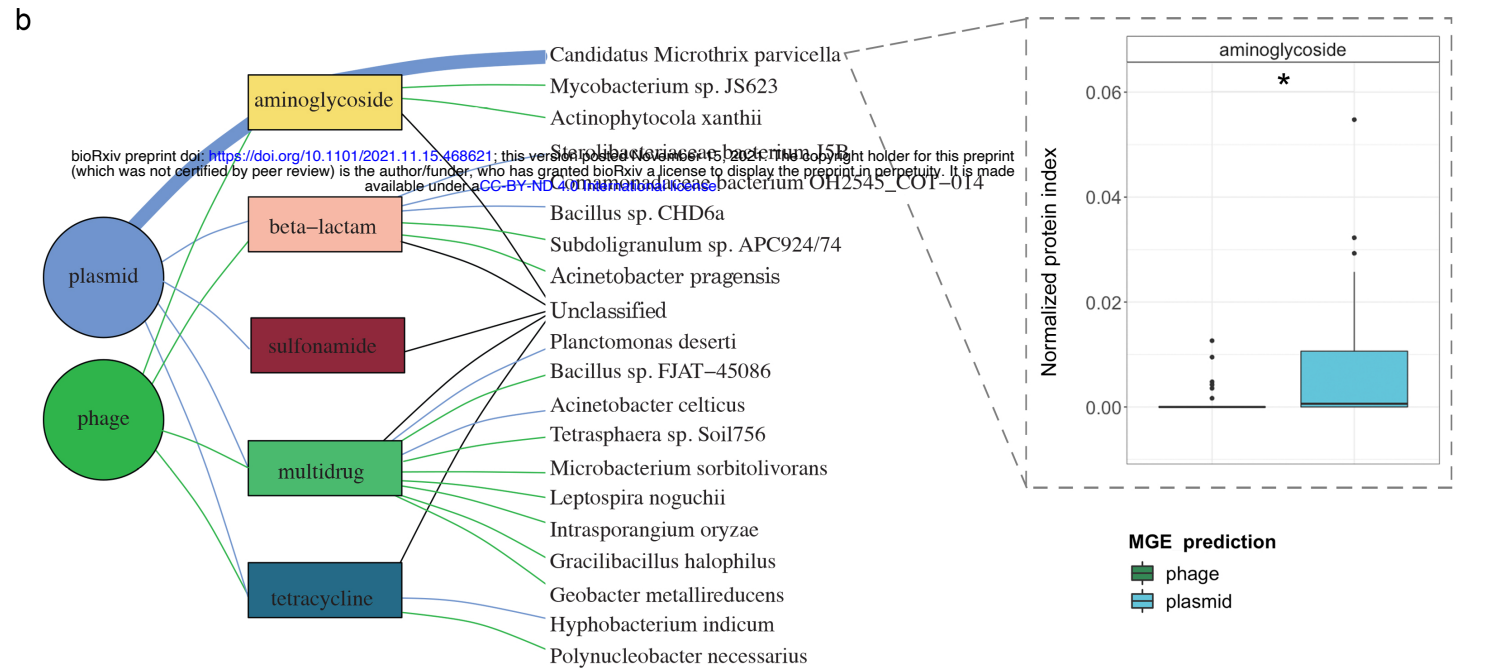
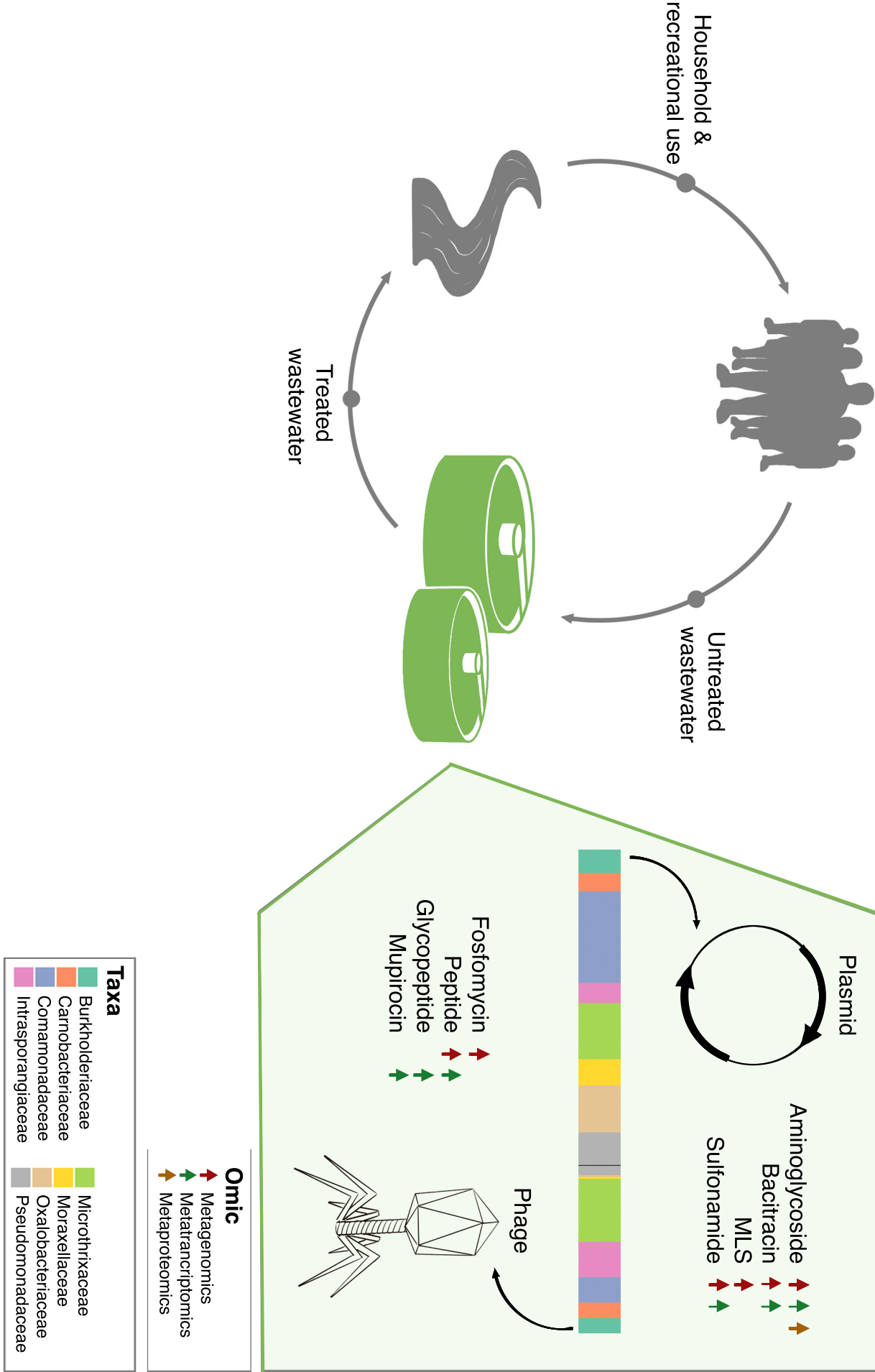


Figure 7



Appendix A.5

Glacier-stream biofilms harbour diverse resistomes and biosynthetic gene clusters

Glacier-stream biofilms harbour diverse resistomes and biosynthetic gene clusters

Susheel Bhanu Busi^{1,#,*}, Laura de Nies^{1,#}, Paraskevi Pramateftaki², Massimo Bourquin², Leïla Ezzat², Tyler J. Kohler², Stilianos Fodelianakis², Grégoire Michoud², Hannes Peter², Michail Styllas², Matteo Tolosano², Vincent De Staercke², Martina Schön², Valentina Galata¹, Paul Wilmes^{1,*} and Tom Battin²

¹Systems Ecology Group, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg

²Stream Biofilm & Ecosystem Research Lab, ENAC, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

#Contributed equally to this work

*Corresponding author(s):

Prof. Paul Wilmes (paul.wilmes@uni.lu)

Susheel Bhanu Busi (susheel.busi@uni.lu)

Running title: Resistome and biosynthetic gene clusters of glacier-fed streams

Keywords: glacier-fed streams, metagenomics, antimicrobial resistance, biosynthetic gene clusters, cross-domain interactions

Abstract

Background

Antimicrobial resistance (AMR) is a universal phenomenon whose origins lay in natural ecological interactions such as competition within niches, within and between micro- to higher-order organisms. However, the ecological and evolutionary processes shaping AMR need to be better understood in view of better antimicrobial stewardship. Resolving antibiotic biosynthetic pathways, including biosynthetic gene clusters (BGCs), and corresponding antimicrobial resistance genes (ARGs) may therefore help in understanding the inherent mechanisms. However, to study these phenomena, it is crucial to examine the origins of AMR in pristine environments with limited anthropogenic influences. In this context, epilithic biofilms residing in glacier-fed streams (GFSs) are an excellent model system to study diverse, intra- and inter-domain, ecological crosstalk.

Results

We assessed the resistomes of epilithic biofilms from GFSs across the Southern Alps (New Zealand) and the Caucasus (Russia) and observed that both bacteria and eukaryotes encoded twenty-nine distinct AMR categories. Of these, beta-lactam, aminoglycoside, and multidrug resistance were both abundant and taxonomically distributed in most of the bacterial and eukaryotic phyla. AMR-encoding phyla included Bacteroidota and Proteobacteria among the bacteria, alongside Ochrophyta (algae) among the eukaryotes. Additionally, BGCs involved in the production of antibacterial compounds were identified across all phyla in the epilithic biofilms. Furthermore, we found that several bacterial genera (*Flavobacterium*, *Polaromonas*, etc.) including representatives of the superphylum Patescibacteria encode both ARGs and BGCs within

close proximity of each other, thereby demonstrating their capacity to simultaneously influence and compete within the microbial community.

Conclusions

Our findings highlight the presence and abundance of AMR in epilithic biofilms within GFSs. Additionally, we identify their role in the complex intra- and inter-domain competition and the underlying mechanisms influencing microbial survival in GFS epilithic biofilms. We demonstrate that eukaryotes may serve as AMR reservoirs owing to their potential for encoding ARGs. We also find that the taxonomic affiliation of the AMR and the BGCs are congruent. Importantly, our findings allow for understanding how naturally occurring BGCs and AMR contribute to the epilithic biofilms mode of life in GFSs. Importantly, these observations may be generalizable and potentially extended to other environments which may be more or less impacted by human activity.

Background

Today, antimicrobial resistance (AMR) has become a well-known threat to human health with an estimated number of 700,000 people per year dying of drug-resistant infections [1]. The dramatic rise of antimicrobial resistance over the past decade has even led to the moniker, “silent pandemic” [2]. Therefore, AMR is often directly associated with human impacted environments with a global increase in resistant bacteria linked to the over- and mis-use of antibiotics [3]. However, contrary to public perception, AMR is a natural phenomenon, which has existed for billions of years [4]. Long before the rather recent use of antibiotics in the clinical setting, microorganisms have used these, along with corresponding protective mechanisms, to establish competitive advantages over other microbes contending for the same environment and/or resources [5].

Microbes, in general, produce a range of secondary metabolites with diverse chemical structures which in turn confer a variety of functions, including antibiotics [6]. Such secondary metabolites including metal transporters and quorum sensing molecules [7,8] are not directly associated with the growth of microorganisms themselves but instead are known to provide benefits by acting as growth inhibitors against competing bacteria. Consequently, many of these natural products have found their uses in industrial settings as well as in human medicine as anti-infective drugs [7,9,10]. The biosynthetic pathways responsible for producing these specialized metabolites are encoded by locally clustered groups of genes known as ‘biosynthetic gene clusters’ (BGCs). Typically, BGCs include genes for expression control, self-resistance, and metabolite export [11]. They can, however, be further divided into various classes including non-ribosomal peptide synthetases (NRPSs), type I and type II polyketide synthases (PKSs), terpenes, and

bacteriocins alongside others [10]. NRPSs and PKSs specifically have been of interest due to their known synthesis of putative antibiotics [12,13]. Furthermore, evidence suggests that within these BGCs at least one resistance gene conferring resistance can be found as a self-defense mechanism against the potentially harmful secondary metabolites encoded by the BGC [14]. For instance, the tylosin-biosynthetic gene cluster of *Streptomyces fradiae* also encodes three resistance genes (*tlrB*, *tlrC* and *tlrD*) [15], while in another example, *Streptomyces toyacaensis*, the *vanHAX* resistance cassette is proximal to the vancomycin biosynthesis gene cluster, thereby encoding inherent resistance [16].

Remote and pristine microbial communities provide a rich genetic resource to explore the historical evolutionary origins of naturally occurring antibiotic resistance from the pre-antibiotic era. Only in few pristine environments with limited anthropogenic influence (e.g., permafrost, glaciers, deep sea, and polar regions) can remnants of the above-described ancient biological warfare mechanisms still be detected. These ARGs and resistant bacteria evolving in pristine environments may therefore be considered the inherent antibiotic resistance present in the environment [5].

We have recently reported the genomic and metabolic adaptations of epilithic biofilms to windows of opportunities in glacier-fed streams (GFSs) [17]. For example, given the short flow season during glacial melt, i.e. summer, the incentive to reproduce quickly while conditions are favourable, is high. During these windows of opportunity, the necessity for taxa to not only acquire physical niches, but also appropriate resources yields a competitive environment. Within these biofilms, we observe complex cross-domain

interactions between microorganisms to potentially mitigate the harsh nutrient and environmental conditions of the GFSs. Additionally, owing to their complex biodiversity [18] and generally oligotrophic conditions [19], epilithic biofilms are ideal model systems for understanding BGCs and AMR. While oligotrophy may provide the basis for competition over resources amongst microorganisms such as prokaryotes and (micro-)eukaryotes. Our previous insights revealed that taxa such as *Polaromonas*, *Acidobacteria*, and *Methylothermobacter* have strong interactions with eukaryotes such as algae and fungi [17]. The inherent diversity allows for understanding the influence of AMR in microbial interactions. For example, the accidental discovery of penicillin by Alexander Fleming in 1928 based on bacterial-fungal interactions, [20], has since been expanded upon by Netzker *et al.* [21]. They reported that microbial interactions lead to the production of bioactive compounds including antibiotics that may shape the microbial consortia within a community.

Here, to shed light on the role of AMR in shaping microbial communities within (relatively) pristine environments, we used high-resolution metagenomics to investigate twenty-one epilithic biofilms from glacier-fed streams. These samples were collected from 8 GFSs spread across the Southern Alps in New Zealand and the Caucasus in Russia (Supplementary Table 1). Herein, we found 29 categories of ARGs within the GFSs across both bacterial and eukaryotic domains. Importantly, most of the AMR was found in bacteria. We also identified antibacterial BGCs that were encoded both in bacterial and eukaryotes suggesting extensive intra- and inter-domain competition. Our findings demonstrate that microorganisms within biofilms from pristine environments not only encode ARGs, but that they may potentially influence several features of epilithic biofilms

such as biofilm formation, community assembly and/or maintenance, including conferring mechanisms for competitive advantages under extreme conditions.

Methods

Sampling and biomolecular extractions

Eight GFSs were sampled in early- to mid-2019 from the New Zealand Southern Alps and the Russian Caucasus, respectively, for a total of 21 epilithic biofilms (Supp. Table 1). The biofilm samples were collected from each stream reach due to biofilms ranging from abundant to absent, depending on stream geomorphology. One to three biofilm samples were collected per reach (Supp. Table 1), taken using sterilized metal spatulas to scrape rocks, followed by their immediate transfer to cryovials. Samples were immediately flash-frozen in liquid nitrogen and stored at -80 °C until DNA was extracted. DNA from the epilithic biofilms was extracted using a previously established protocol [22] adapted to a smaller scale due to relatively high DNA concentrations. DNA quantification was performed for all samples with the Qubit dsDNA HS kit (Invitrogen).

Sequencing and data processing for metagenomics

Random shotgun sequencing was performed on all epilithic biofilm DNA samples after library preparation using the NEBNext Ultra II FS library kit. 50 ng of DNA was enzymatically fragmented for 12.5 min and libraries were prepared with six PCR amplification cycles. An average insert of 450 bp was maintained for all libraries. Qubit was used to quantify the libraries followed by sequencing at the Functional Genomics Centre Zurich on a NovaSeq (Illumina) using a S4 flowcell. The metagenomic data was

processed using the Integrated Meta-omic Pipeline (IMP v3.0; commit# 9672c874 available at <https://git-r3lab.uni.lu/IMP/imp3>) [23]. IMP's workflow includes pre-processing, contig assembly, genome reconstruction (metagenome-assembled genomes, i.e. MAGs) and additional functional analysis of genes based on custom databases in a reproducible manner [23].

Identification of antimicrobial resistance genes, antibiotic biosynthesis pathways and BGCs

For the prediction of ARGs the IMP-generated contigs were used as input for PathoFact [24]. Identified ARGs were further collapsed into their respective AMR categories in accordance with the Comprehensive Antibiotic Resistance Database (CARD) [25]. PathoFact uses an HMM-based search to identify homologous sequences across genomic data, therefore possibly also detecting resistance genes within eukaryotic genomic fragments. Subsequently, the raw read counts per ORF, obtained from PathoFact, were determined using FeatureCounts [26].

To identify pathways for the biosynthesis of antibiotics, we assigned KEGG orthology (KOs) identifiers to the ORFs using a hidden Markov model [27] (HMM) approach using *hmmsearch* from HMMER 3.1 [28] with a minimum bit score of 40. Additionally, we linked the identified KOs to their corresponding KEGG orthology pathways and extracted the pathways annotated as antibiotic biosynthesis pathways by KEGG. Both the identified ARGs and KEGG pathways were then further linked to associated bacterial taxonomies. The bacterial and eukaryotic taxonomies were assigned using the PhyloDB and MMETSP databases associated with EUKulele (commit# fb8726a; available at

<https://github.com/AlexanderLabWHOI/EUKulele>). Consensus taxonomy per contig was then used for downstream analyses including association with ARGs.

We further identified BGCs within the MAGs using antiSMASH (ANTibiotics & Secondary Metabolite Analysis SHell) [29] and annotated these using deepBGC [30]. To link BGCs and ARGs, we linked the resistance genes to their associated assembled contigs, followed by identifying the corresponding bins (MAGs) to which said contigs belonged.

Data analysis

The relative abundance of the ORFs was calculated based on the RNum_Gi method described by Hu *et al.* [31]. Figures for the study, including visualizations derived from the taxonomic and functional analyses, were created using version 3.6 of the R statistical software package [32] and using the *tidyverse* package [33]. Alluvial plots were generated using the *ggalluvial* package [34] while heatmaps were generated using the *ComplexHeatmap* package [35] developed for R. The corresponding visualization and analysis code is available at: https://gitr3lab.uni.lu/laura.denies/Rock_Biofilm_AMR.

Results

Antimicrobial resistance in a pristine environment

We characterised the resistomes of GFS epilithic biofilms and assessed the distribution of AMR in twenty-one epilithic biofilm samples, across 8 individual glaciers originating from the Southern Alps in New-Zealand (SA1, SA2, SA3 and SA4) and the Caucasus in

Russia (CU1, CU2, CU3, CU4). In total, we identified a high number (n=1840) of ARGs within 29 categories of AMR, with similar AMR profiles observed across all GFSs (Fig. 1a, Supp. Fig. 1), except for SA2 and SA3 where the differences were driven by elevated fluoroquinolone, glycopeptide and phenicol resistance, respectively. It is to be noted that while ARGs refer to the genes encoding specific resistance, AMR categories derived from metagenomic data in this context, typically reflect the functional potential associated with respect to the resistance encoded. Of the identified AMR categories, beta-lactam and multidrug resistance (i.e. resistance conferring protection against multiple antibiotic classes), followed by aminoglycoside resistance, were found to be highly abundant in all samples. We subsequently analysed the diversity of ARGs within the various resistance categories and found beta-lactam resistance to represent the largest resistance category, contributing 930 unique ARGs to the resistome. This was followed by multidrug (179 ARGs) and aminoglycoside (176 ARGs) resistance (Supp. Table 2). In contrast, some resistance categories such as polymyxin and pleuromutilin resistance were only detected at very low levels within the epilithic biofilm resistomes.

We further investigated the contribution of microbial populations to the resistome and found contributions from both prokaryotes and eukaryotes (Fig. 1b). Prokaryotes within this study refer to bacteria alone, since archaea encoded for an infinitesimal number of ARGs ($<0.000001\%$ RNum_GI; *Methods*), and therefore were excluded from further analyses. Among the eukaryotes, the phylum Ochrophyta (algae) was the dominant contributor and encoded most of the AMR categories (Fig. 1c, Supp. Fig. 2a). In bacteria, AMR was more evenly distributed with most of the phyla encoding ARGs across all categories (Fig. 1c). However, members of the Alphaproteobacteria, Betaproteobacteria,

and the Bacteroidetes/Chlorobi group encoded the highest overall ARG abundance (Fig. 1c, Supp. Fig. 2b). Additionally, AMR categories such as aminoglycoside, beta-lactam, glycopeptide and rifamycin resistance (among others) were widely distributed in both bacteria as well as among the eukaryotes. On the other hand, categories such as aminocoumarin, bacitracin, and diaminopyrimidine resistance were found to be primarily encoded by bacteria.

Antibiotic biosynthesis pathways and biosynthetic gene clusters

As described above, beta-lactam, multidrug and aminoglycoside resistance were the most abundant resistance categories within GFS epilithic biofilms. This was not surprising as beta-lactams and aminoglycosides are natural and prevalent compounds [36,37]. Furthermore, multidrug resistance is typically conferred via efflux machineries which were also common in the GFS epilithic biofilms. These typically serve dual purposes in particular for protein export within most bacteria [38]. Based on these results, it is therefore highly likely that pristine environments such as GFSs potentially reflect the spectrum of natural antibiotics and their resistance mechanisms, reinforcing their capacity to serve as natural baselines for assessing enrichments and spread of AMR.

To further understand if these encoded resistance genes reflected natural antibiotic pressure, we investigated pathways associated with antibiotic biosynthesis using the KEGG database [39]. In total, we identified seven different pathways corresponding to the biosynthesis of macrolides (MLS), ansamycins, glycopeptides (vancomycin), beta-lactams (monobactam, penicillin and cephalosporin), aminoglycosides (streptomycin),

and tetracyclines, which were present in various abundances in all samples (Supp. Fig. 3a). Importantly, the identified antibiotic synthesis genes thereby corresponded to the resistance categories identified within the epilithic biofilms. Interestingly, in most of the GFSs, antibiotic biosynthesis was primarily encoded by bacteria spanning multiple phyla (Supp. Fig. 3b, Supp. Fig. 3c). Exceptions to these were GL11 and GL15 in which biosynthesis pathways were equally distributed among eukaryotes, specifically Ochrophyta, in addition to bacteria.

To further validate our observations, we assessed the abundance of BGCs, which are known to encode genes for secondary metabolite synthesis, including antibiotics. We found six different structural classes of BGCs by annotating 537 medium-to-high quality (>50% completion and <10% contamination) bacterial and 30 eukaryotic MAGs using antiSmash [29] and DeepBGC [30]. Using this ensemble approach we identified one or more BGCs in most bacterial (n=490, ~91% of all bacterial MAGs) and eukaryotic (n=28) MAGs. Of these BGCs, those annotated with an antibacterial function were dominant across the microbial populations, represented here by the MAGs, and were found across all phyla (Fig. 2a). Overall, a wider variety of BGCs associated with cytotoxic activity, inhibitory, and antifungal mechanisms were also identified in bacteria. Eukaryotes, on the other hand, encoded a high prevalence of antibacterial BGCs (~93% of all eukaryotic MAGs) (Fig. 2a). We further annotated those BGCs identified as antibacterial to determine their subtypes and found that most of them were 'unknown' (Fig. 2b). However, other identified subtypes include ribosomally synthesized and post-translationally modified peptides (RiPPs) such as bacteriocins, along with NRPs, PKs, and terpenes.

According to the resistance hypothesis [14], within or close to, each BGC there is at least one gene conferring resistance to its encoded secondary metabolite. To test this, we assessed whether the MAGs encoding a BGC also encoded corresponding ARGs. In line with this hypothesis, we identified BGCs and their respective resistance genes in close proximity to each other through their localization on the same contig. Consequently, we identified various BGCs encoded together with ARGs in both the bacterial and eukaryotic MAGs. For example, we found that an antibacterial BGC was encoded by *Flavobacterium* spp. on the same contig as both MLS (macrolides, lincosamides and streptogramin) and beta-lactam resistance genes (Fig. 2c). Incidentally, we also found that a candidate phyla radiation (CPR) bacterium (Aalborg-AAW-1; phylum Patescibacteria) also encoded both antibacterial BGC and MLS resistance on the same contig.

Discussion

Microbial reservoirs in pristine environments, with little to no impact from anthropogenic selection pressures, provide the opportunity to investigate the natural propensity and linked evolutionary origins of AMR. Here, by leveraging high-resolution metagenomics on twenty-one epilithic biofilms, we assessed the resistomes of eight individual GFS epilithic biofilms.

To date, while many studies have looked for novel antibiotics and resistance genes in pristine environments such as the deep sea [40] or the polar regions [41], few have explored the full diversity of antibiotic resistance in such environments [42,43]. Van Goethem *et al.* [44] identified 117 naturally occurring ARGs associated with multidrug,

aminoglycoside and beta-lactam resistance in pristine Antarctic soils. Similarly, D'Costa *et al.* [4] identified a collection of ARGs encoding resistance to beta-lactams as well as tetracyclines and glycopeptides in 30,000-year-old Beringian permafrost sediments. In agreement with these previous studies, we identified 29 AMR categories, including the previously mentioned resistance categories, in the studied biofilm communities. Among these, the highest ARG abundance was associated with aminoglycoside and beta-lactam resistance. Our study further suggests that although the overall abundance differs, the epilithic resistome was highly similar in all GFSs, independent of origin (i.e. New Zealand or Russia). Furthermore, our results agree with the results obtained in other resistomes identified in pristine environments such as Antarctic soils and permafrost in terms of the identified ARGs. Unlike previous studies, where ARGs were primarily associated with bacteria, we report for the first time that AMR was associated with both bacteria and eukaryotes in various abundances in environmental samples including GFSs. A previous study by Brown *et al.* [45] reported that the IRS-HR (isoleucyl-tRNA synthetase - high resistance) type gene conferring resistance against mupirocin was identified in *Staphylococcus aureus*. More importantly, they suggested that horizontal gene transfer led to the acquisition of IRS-HR genes by bacteria from eukaryotes [45]. Despite these early reports, the contribution of eukaryotes to most resistomes, including from pristine environments, has largely been unexplored thus far. An exception to this was the report by Fairlamb *et al.* [46] who identified eukaryotic drug resistance, especially encoded by fungi (*Candida* and *Aspergillus*) and parasites (*Plasmodium* and *Trypanosoma*). However, most of these modes of resistance were highly specific towards particular drug treatments [46]. Our results specifically revealed that taxa from the phylum Ochrophyta

encoded resistance to 28 AMR categories and this was also reflected in other (micro-)eukaryotes.

Apart from encoded resistance mechanisms, microalgae such as Ochrophyta have been of interest as a source of (new) antimicrobial compounds [47,48]. In line with this, Martins *et al.* suggested that extracts from different microalgae may potentially serve not only as antimicrobial agents, but also as anti-cancer therapeutics. However, our present results suggest that these taxa may also serve as environmental reservoirs for AMR itself. It is however presently unclear whether this phenomenon confers advantages with respect to niche occupation and protection against bacterial infection as well as whether the eukaryotes are sensitive to the antibiotics produced by them.

Studies delving into the origins of AMR have reported that fecal pollution may explain ARG abundances in anthropogenically impacted environments [49]. This phenomenon was also observed by Antelo *et al.* [50] and others [51] who detected ARGs in soils in Antarctica, especially in proximity to scientific bases. Although it is plausible that some of the GFSs sampled in our study may indeed be under anthropogenic influence, in pristine environments, AMR is most likely derived from natural antibiotics produced by microorganisms as a competitive advantage. Microorganisms acquire resistance either as a protective measure against other microorganisms [52,53] or as a self-defense mechanism to prevent inadvertent suicide by damaging metabolites [14]. Accordingly, we found both antibiotic biosynthesis pathways and BGCs within the epilithic resistomes. We identified pathways for the biosynthesis of glycopeptides, beta-lactams, and aminoglycosides, among others, concurrent with the high abundance of ARGs against

said antibiotics. Additionally, we identified BGCs with a predicted antibacterial function in both eukaryotes and bacteria. While a limited number of studies such as Waschulin *et al.* [54] and Liao *et al.* [55], have shown BGCs in pristine environments, none of these studies have contextualized the co-occurrence of BGCs with AMR. Hence, we not only found that most of our MAGs contain BGCs, of which many have an antibacterial function, but also found all MAGs to encode multiple resistance genes. Additionally, we found several BGCs closely localized to ARGs on the same contig, thereby indicating an immediate self-defense mechanism against the encoded secondary metabolites. This agrees with the resistance hypothesis highlighted by Tran *et al.* stating that a gene conferring resistance to potentially harmful metabolites produced by the organism are to be found within the BGC-encoding operons [14]. We also observed that the recently identified CPR bacteria [56] (in our case, phylum Patescibacteria) not only encoded for AMR but also harboured genes associated with the production of molecules with antibacterial effects. Although Patescibacteria have been identified in oligotrophic environments [57,58] with carbon and/or nutrient limitations similar to those observed for GFSs, it is plausible that their ability to survive with minimal biosynthetic and metabolic pathways may indeed depend on the expression of BGCs and AMR. At the time of writing, a preprint by Maatouk *et al.* [59], described the presence of ARGs across publicly available CPR bacterial genomes. In addition, we report the identification of AMR within GFS-derived CPR genomes, likely as a means of competitive inhibition against other taxa. Alternatively, biofilms may also allow for collective resistance, tolerance, and exposure protection to antibacterial compounds [60]. The AMR and BGCs encoded by most phyla may therefore affect cooperation and/or interactions associated with nutrient exchange, leading to the privatization of public goods [60]. Such a phenomenon may be achieved due to the

competition within taxa, both at the intra- and inter-species levels, via secretion of toxins [53] and occupying spatial niches [61,62] thereafter. Furthermore, Stubbendieck and Straight previously highlighted the multifaceted effects of bacterial competition which include the potential taxation and subsequent increase in bacterial fitness [63]. Thus, the *in-situ* competition within multi-species biofilms may allow for cross-phyla and cross-domain interactions whilst simultaneously increasing the overall fitness of the endogenous epilithic microbial community. Alternatively, these interactions or lack thereof may shape the overall community including spatial organisation [64], especially in energy limited systems such as the GFSs.

Conclusions

Epilithic biofilms are an integral and key mode of survival in extreme environments such as glacier-fed stream ecosystems. Herein, we report that these biofilms provide critical insights into the naturally occurring resistome. Our findings demonstrate that intra- and inter-domain competition and survival mechanisms shed light on the ecological dimension of microbial communities. Furthermore, we reveal the congruence of genes encoding for both BGCs and AMR, in both bacteria and eukaryotes. More importantly, we highlight for the first time the comprehensive AMR profile of CPR bacteria and of (micro-)eukaryotes. Collectively, our results highlight underlying resistance mechanisms, including BGCs, employed in 'biological warfare' in oligotrophic and challenging glacier-fed stream ecosystems.

List of Abbreviations

- AMR: Antimicrobial resistance
- ARGs: Antimicrobial resistance gene(s)
- BGC: Biosynthetic gene clusters
- CA: Caucasus
- CPR: Candidate Phyla radiation
- GFSSs: Glacier-fed stream(s)
- GL: Glacier
- IRS-RS: isoleucyl-tRNA synthetase - high resistance
- IMP: Integrate Meta-Omics Pipeline
- KEGG: Kyoto Encyclopedia of Genes and Genomes
- MAGs: Metagenome-assembled genome(s)
- NRPS: Non-ribosomal peptide synthetases
- PKS: Polyketide synthases (type I and type II)
- RiPPs: Post-translationally modified peptide(s)
- SA: Southern Alps

Declarations

- Ethics approval and consent to participate
- Not applicable
- Consent for publication
- Not applicable

Availability of data and material

The Biosample accession IDs listed under Supp. Table 3 can be found on NCBI under the BioProject accession# **PRJNA733707**. The analyses code for IMP and downstream analyses is detailed at https://git-r3lab.uni.lu/susheel.busi/nomis_pipeline. Binning and manual refinement of eukaryotic MAGs was done as described here: https://git-r3lab.uni.lu/susheel.busi/nomis_pipeline/-/blob/master/workflow/notes/MiscEUKMAGs.md. All visualization and analysis code is available at: https://git-r3lab.uni.lu/laura.denies/Rock_Biofilm_AMR.

Competing interests

The authors declare that they have no competing interests

Funding

This research has been supported by The NOMIS Foundation to TJB and the Swiss National Science Foundation (CRSII5_180241) supporting SBB. LdN and PW are supported by the Luxembourg National Research Fund (FNR; PRIDE17/11823097) awarded to PW.

Authors' contributions

SBB, LdN, PW, and TJB conceived the project. PP extracted DNA, SBB and PP prepared the metagenomic libraries for sequencing. SBB and LdN conceptualized and performed the data analyses. SBB and LdN wrote the manuscript with PW and TJB, with significant input and editing from all coauthors.

Acknowledgements

We gratefully acknowledge the laboratory support from Emmy Marie Oppliger at EPFL and Lea Grandmougin, Janine Habier, Laura Lebrun at the University of Luxembourg. We also acknowledge the key input from Rashi Halder at the LCSB Sequencing Platform regarding library preparation. We thank Patrick May and Cedric Christian Laczny for the crucial insights into metagenomic processing. The computational analyses were performed at the HPC facilities at the University of Luxembourg (<https://hpc.uni.lu>) [65].

Figure legends

Figure 1. Epilithic biofilms in GFSs harbour a diverse resistome

(a) Relative abundance of 29 AMR categories within 21 epilithic biofilms collected from four New Zealand Southern Alps (SA) and four Russian Caucasus (CU) GFSs. (b) Bar plots depicting the relative abundance of bacteria and eukaryotes encoding ARGs. (c) Phylum-level representation of the AMR abundances across bacteria and eukaryotes. Size of the closed circle indicates the normalised relative abundance (Rnum_Gi; see *Methods*), whereby the color represents individual phyla.

Figure 2. Biosynthetic gene clusters indicate the resistome potential

(a) Heatmap depicting the overall abundance of BGCs identified across bacterial and eukaryotic MAGs. The respective phyla are listed on the left while the coloured legend represents the taxonomic order. (b) In-depth characterisation of the ‘antibacterial’ BGCs found within all phyla and orders across medium-to-high quality MAGs. (c) Alluvial plots

depicting the taxa where both BGCs and AMR were found adjacently on the same contig.
Colours indicate the genera associated with the MAGs.

Supplementary figure 1. Ordination analyses reveal the (dis)similarity of the GFS resistomes

(a) Principal component analyses depicting the overall similarity of the individual GFS resistomes. Each dot represents the resistome predicted from a single metagenome. SA: Southern Alps. CU: Caucasus. (b) Biplot demonstrating the underlying factors, i.e. ARG abundances across 29 AMR categories, driving the similarity within the GFS epilithic resistomes.

Supplementary figure 2. Bacterial and eukaryotic phyla encode AMR

(a) Relative abundance of the bacteria associated with AMR. The stacked bar plots are faceted by the individual GFSs where the epilithic biofilms were collected. The colors represent the individual phyla. (b) Stacked bar plots indicating the relative abundance of the AMR encoded by eukaryotes.

Supplementary figure 3. Antibiotic synthesis pathway assessment via KEGG orthology

(a) Relative abundance of KEGG pathways associated with antibiotic synthesis across the 21 epilithic biofilms. (b) Bar plots indicating the relative abundance of the antibiotic associated KEGG pathways mediated by bacteria and eukaryotes. (c) Normalised relative abundance of pathways associated with antibiotic production in the KEGG database, juxtaposed with the various phyla encoding these genes.

478

479 **Supplementary data**

480 **Supplementary table 1. Sample metadata**

481 **Supplementary table 2. List of ARGs identified across 21 GFS epilithic biofilms**

482 **Supplementary table 3. NCBI accession metadata**

483

484 **References**

- 485 1. Stanton IC, Bethel A, Leonard AFC, Gaze WH, Garside R. What is the research evidence
486 for antibiotic resistance exposure and transmission to humans from the environment? A
487 systematic map protocol. *Environ Evid.* 2020;9: 12.
- 488 2. Balasegaram M. Learning from COVID-19 to Tackle Antibiotic Resistance. *ACS Infect Dis.*
489 2021;7: 693–694.
- 490 3. Wright GD. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nat Rev*
491 *Microbiol.* 2007;5: 175–186.
- 492 4. D’Costa VM, King CE, Kalan L, Morar M, Sung WWL, Schwarz C, et al. Antibiotic
493 resistance is ancient. *Nature.* 2011;477: 457–461.
- 494 5. Scott LC, Lee N, Aw TG. Antibiotic Resistance in Minimally Human-Impacted
495 Environments. *Int J Environ Res Public Health.* 2020;17. doi:10.3390/ijerph17113939
- 496 6. Tyc O, Song C, Dickschat JS, Vos M, Garbeva P. The Ecological Role of Volatile and
497 Soluble Secondary Metabolites Produced by Soil Bacteria. *Trends Microbiol.* 2017;25: 280–
498 292.
- 499 7. Chen R, Wong HL, Kindler GS, MacLeod FI, Benaud N, Ferrari BC, et al. Discovery of an

- 500 Abundance of Biosynthetic Gene Clusters in Shark Bay Microbial Mats. *Front Microbiol.*
501 2020;11: 1950.
- 502 8. Demain AL, Fang A. The Natural Functions of Secondary Metabolites. In: Fiechter A, editor.
503 History of Modern Biotechnology I. Berlin, Heidelberg: Springer Berlin Heidelberg; 2000. pp.
504 1–39.
- 505 9. Newman DJ, Cragg GM. Natural Products as Sources of New Drugs from 1981 to 2014. *J*
506 *Nat Prod.* 2016;79: 629–661.
- 507 10. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, et al. Minimum
508 Information about a Biosynthetic Gene cluster. *Nat Chem Biol.* 2015;11: 625–631.
- 509 11. Martinet L, Naômé A, Deflandre B, Maciejewska M, Tellatin D, Tenconi E, et al. A Single
510 Biosynthetic Gene Cluster Is Responsible for the Production of Bagremycin Antibiotics and
511 Ferroverdin Iron Chelators. *MBio.* 2019;10. doi:10.1128/mBio.01230-19
- 512 12. Martínez-Núñez MA, López VEL y. Nonribosomal peptides synthetases and their
513 applications in industry. *Sustainable Chemical Processes.* 2016;4: 1–8.
- 514 13. Ridley CP, Lee HY, Khosla C. Evolution of polyketide synthases in bacteria. *Proc Natl Acad*
515 *Sci U S A.* 2008;105: 4595–4600.
- 516 14. Tran PN, Yen M-R, Chiang C-Y, Lin H-C, Chen P-Y. Detecting and prioritizing biosynthetic
517 gene clusters for bioactive compounds in bacteria and fungi. *Appl Microbiol Biotechnol.*
518 2019;103: 3277–3287.
- 519 15. Cundliffe E, Bate N, Butler A, Fish S, Gandechea A, Merson-Davies L. The tylosin-
520 biosynthetic genes of *Streptomyces fradiae*. *Antonie Van Leeuwenhoek.* 2001;79: 229–234.
- 521 16. Kwun MJ, Hong H-J. Genome Sequence of *Streptomyces toyocaensis* NRRL 15009,
522 Producer of the Glycopeptide Antibiotic A47934. *Genome Announc.* 2014;2.

doi:10.1128/genomeA.00749-14

17. Busi SB, Bourquin M, Fodelianakis S, Michoud G, Kohler TJ, Peter H, et al. Genomic and metabolic adaptations of biofilms to ecological windows of opportunities in glacier-fed streams. *bioRxiv*. 2021. p. 2021.10.07.463499. doi:10.1101/2021.10.07.463499
18. Battin TJ, Besemer K, Bengtsson MM, Romani AM, Packmann AI. The ecology and biogeochemistry of stream biofilms. *Nat Rev Microbiol*. 2016;14: 251–263.
19. Battin TJ, Wille A, Sattler B, Psenner R. Phylogenetic and functional heterogeneity of sediment biofilms along environmental gradients in a glacial stream. *Appl Environ Microbiol*. 2001;67: 799–807.
20. Gaynes R. The Discovery of Penicillin—New Insights After More Than 75 Years of Clinical Use. *Emerg Infect Dis*. 2017;23: 849.
21. Netzker T, Flak M, Krespach MK, Stroe MC, Weber J, Schroeckh V, et al. Microbial interactions trigger the production of antibiotics. *Curr Opin Microbiol*. 2018;45: 117–123.
22. Busi SB, Pramateftaki P, Brandani J, Fodelianakis S, Peter H, Halder R, et al. Optimised biomolecular extraction for metagenomic analysis of microbial biofilms from high-mountain streams. *PeerJ*. 2020;8: e9973.
23. Narayanasamy S, Jarosz Y, Muller EEL, Heintz-Buschart A, Herold M, Kaysen A, et al. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol*. 2016;17: 260.
24. de Nies L, Lopes S, Busi SB, Galata V, Heintz-Buschart A, Laczny CC, et al. PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. *Microbiome*. 2021;9: 49.
25. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, et al. CARD

2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* 2020;48: D517–D525.

26. Liao Y, Smyth GK, Shi W. featureCounts: An efficient general-purpose program for assigning sequence reads to genomic features. *arXiv [q-bio.GN]*. 2013. Available: <http://arxiv.org/abs/1305.3347>

27. Yoon B-J. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr Genomics*. 2009;10: 402–415.

28. Eddy SR. Accelerated Profile HMM Searches. *PLoS Comput Biol*. 2011;7: e1002195.

29. Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.* 2021;49: W29–W35.

30. Hannigan GD, Prihoda D, Palicka A, Soukup J, Klempir O, Rampula L, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.* 2019;47: e110.

31. Hu Y, Yang X, Qin J, Lu N, Cheng G, Wu N, et al. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nat Commun*. 2013;4: 2151.

32. Computing R, Others. R: A language and environment for statistical computing. Vienna: R Core Team. 2013. Available: <https://www.yumpu.com/en/document/view/6853895/r-a-language-and-environment-for-statistical-computing>

33. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the tidyverse. *J Open Source Softw*. 2019;4: 1686.

34. Brunson J. ggalluvial: Layered Grammar for Alluvial Plots. *J Open Source Softw*. 2020;5: 2017.

35. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*. 2016;32: 2847–2849.
36. Krause KM, Serio AW, Kane TR, Connolly LE. Aminoglycosides: An Overview. *Cold Spring Harb Perspect Med*. 2016;6. doi:10.1101/cshperspect.a027029
37. Tahlan K, Jensen SE. Origins of the β -lactam rings in natural products. *J Antibiot* . 2013;66: 401–410.
38. Borges-Walmsley MI, McKeegan KS, Walmsley AR. Structure and function of efflux pumps that confer resistance to drugs. *Biochem J*. 2003;376: 313–338.
39. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 2000;28: 27–30.
40. Tortorella E, Tedesco P, Palma Esposito F, January GG, Fani R, Jaspars M, et al. Antibiotics from Deep-Sea Microorganisms: Current Discoveries and Perspectives. *Mar Drugs*. 2018;16. doi:10.3390/md16100355
41. McCann CM, Christgen B, Roberts JA, Su J-Q, Arnold KE, Gray ND, et al. Understanding drivers of antibiotic resistance genes in High Arctic soil ecosystems. *Environ Int*. 2019;125: 497–504.
42. Yuan K, Yu K, Yang R, Zhang Q, Yang Y, Chen E, et al. Metagenomic characterization of antibiotic resistance genes in Antarctic soils. *Ecotoxicol Environ Saf*. 2019;176: 300–308.
43. Centurion VB, Delforno TP, Lacerda-Júnior GV, Duarte AWF, Silva LJ, Bellini GB, et al. Unveiling resistome profiles in the sediments of an Antarctic volcanic island. *Environ Pollut*. 2019;255: 113240.
44. Van Goethem MW, Pierneef R, Bezuidt OKI, Van De Peer Y, Cowan DA, Makhalanyane TP. A reservoir of “historical” antibiotic resistance genes in remote pristine Antarctic soils.

Microbiome. 2018;6: 40.

45. Brown JR, Zhang J, Hodgson JE. A bacterial antibiotic resistance gene with eukaryotic origins. *Curr Biol*. 1998;8: R365–7.

46. Fairlamb AH, Gow NAR, Matthews KR, Waters AP. Drug resistance in eukaryotic microorganisms. *Nat Microbiol*. 2016;1: 16092.

47. Silva A, Silva SA, Carpena M, Garcia-Oliveira P, Gullón P, Barroso MF, et al. Macroalgae as a Source of Valuable Antimicrobial Compounds: Extraction and Applications. *Antibiotics* (Basel). 2020;9. doi:10.3390/antibiotics9100642

48. Martins RM, Nedel F, Guimarães VBS, da Silva AF, Colepicolo P, de Pereira CMP, et al. Macroalgae Extracts From Antarctica Have Antimicrobial and Anticancer Potential. *Front Microbiol*. 2018;9: 412.

49. Karkman A, Pärnänen K, Larsson DGJ. Fecal pollution can explain antibiotic resistance gene abundances in anthropogenically impacted environments. *Nat Commun*. 2019;10: 80.

50. Antelo V, Giménez M, Azziz G, Valdespino-Castillo P, Falcón LI, Ruberto LAM, et al. Metagenomic strategies identify diverse integron-integrase and antibiotic resistance genes in the Antarctic environment. *Microbiologyopen*. 2021;10. doi:10.1002/mbo3.1219

51. Hernández F, Calisto-Ulloa N, Gómez-Fuentes C, Gómez M, Ferrer J, González-Rocha G, et al. Occurrence of antibiotics and bacterial resistance in wastewater and sea water from the Antarctic. *J Hazard Mater*. 2019;363: 447–456.

52. Reygaert WC. An overview of the antimicrobial resistance mechanisms of bacteria. *AIMS Microbiol*. 2018;4: 482–501.

53. Granato ET, Meiller-Legrand TA, Foster KR. The Evolution and Ecology of Bacterial Warfare. *Curr Biol*. 2019;29: R521–R537.

54. Waschulin V, Borsetto C, James R, Newsham KK, Donadio S, Corre C, et al. Biosynthetic potential of uncultured Antarctic soil bacteria revealed through long-read metagenomic sequencing. *ISME J.* 2021. doi:10.1038/s41396-021-01052-3
55. Liao L, Su S, Zhao B, Fan C, Zhang J, Li H, et al. Biosynthetic Potential of a Novel Antarctic Actinobacterium *Marisediminicola antarctica* ZS314T Revealed by Genomic Data Mining and Pigment Characterization. *Mar Drugs.* 2019;17. doi:10.3390/md17070388
56. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. *Nat Microbiol.* 2016;1: 16048.
57. Tian R, Ning D, He Z, Zhang P, Spencer SJ, Gao S, et al. Small and mighty: adaptation of superphylum Patescibacteria to groundwater environment drives their genome simplicity. *Microbiome.* 2020;8: 51.
58. Vigneron A, Cruaud P, Langlois V, Lovejoy C, Culley AI, Vincent WF. Ultra-small and abundant: Candidate phyla radiation bacteria are potential catalysts of carbon transformation in a thermokarst lake ecosystem. *Limnol Oceanogr Lett.* 2020;5: 212–220.
59. Maatouk M, Ibrahim A, Rolain J-M, Merhej V, Bittar F. Small and equipped: the rich repertoire of antibiotic resistance genes in Candidate Phyla Radiation genomes. *bioRxiv.* 2021. p. 2021.07.02.450847. doi:10.1101/2021.07.02.450847
60. Bottery MJ, Pitchford JW, Friman V-P. Ecology and evolution of antimicrobial resistance in bacterial communities. *ISME J.* 2021;15: 939–948.
61. Bottery MJ, Passaris I, Dytham C, Wood AJ, van der Woude MW. Spatial Organization of Expanding Bacterial Colonies Is Affected by Contact-Dependent Growth Inhibition. *Curr Biol.* 2019;29: 3622–3634.e5.
62. Schluter J, Nadell CD, Bassler BL, Foster KR. Adhesion as a weapon in microbial

638 competition. ISME J. 2015;9: 139–149.

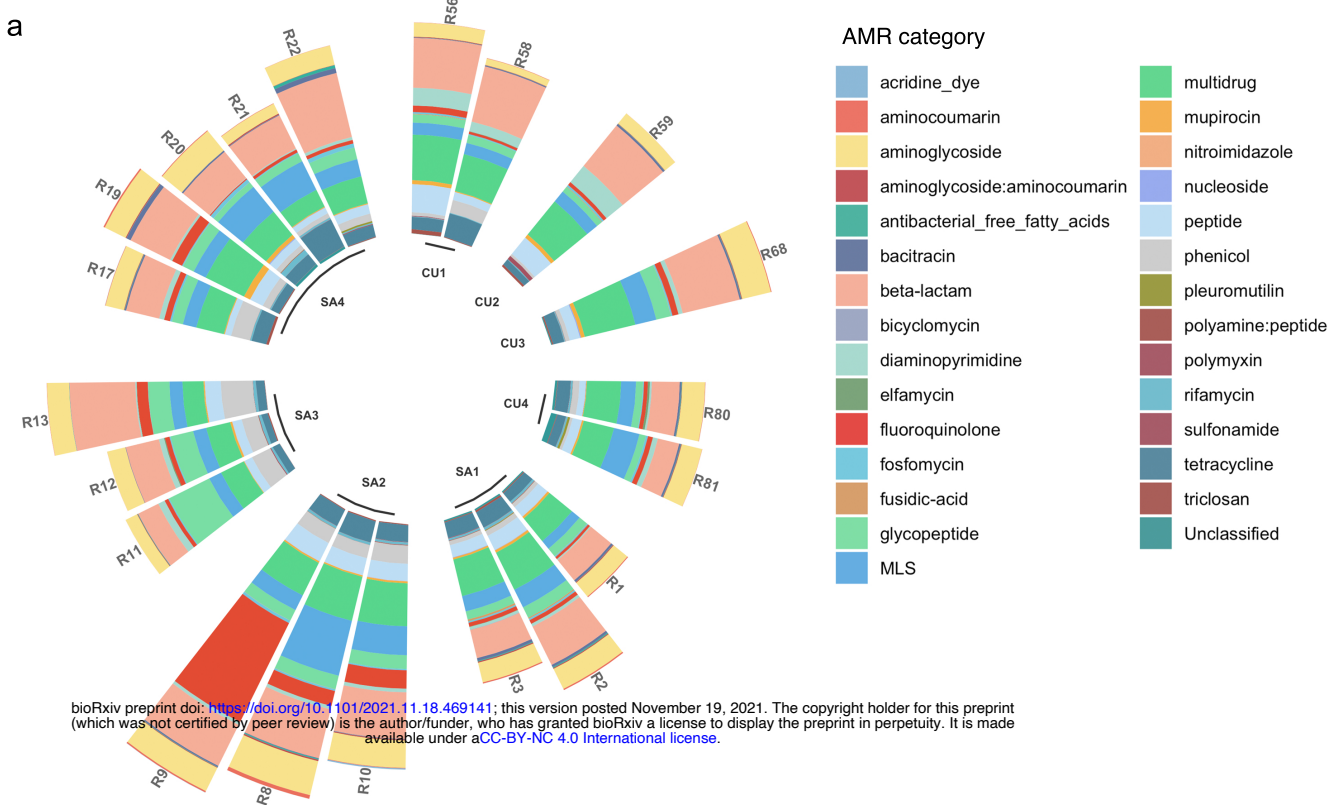
639 63. Stubbendieck RM, Straight PD. Multifaceted Interfaces of Bacterial Competition. J
640 Bacteriol. 2016;198: 2145–2155.

641 64. Estrela S, Brown SP. Community interactions and spatial structure shape selection on
642 antibiotic resistant lineages. PLoS Comput Biol. 2018;14: e1006179.

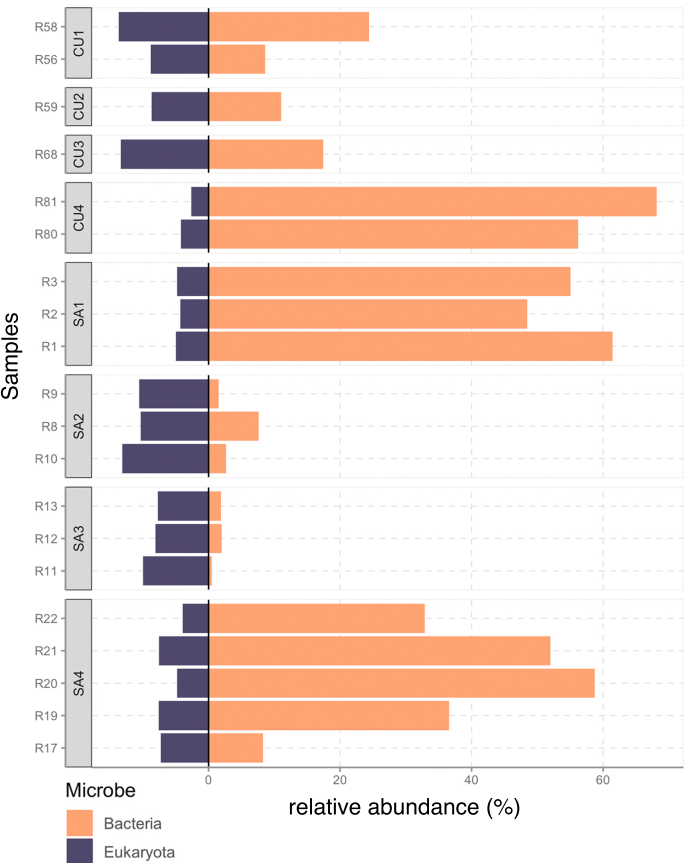
643 65. Varrette S, Bouvry P, Cartiaux H, Georgatos F. Management of an academic HPC cluster:
644 The UL experience. 2014 International Conference on High Performance Computing
645 Simulation (HPCS). 2014. pp. 959–967.

Figure 1

a



b



c

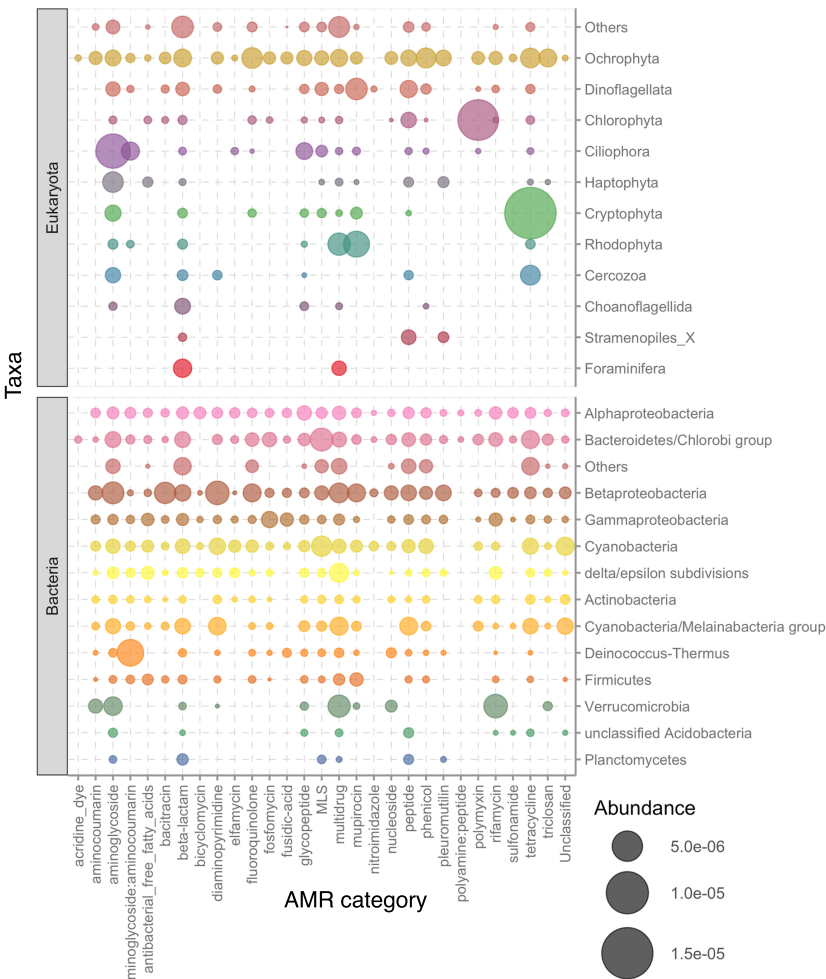
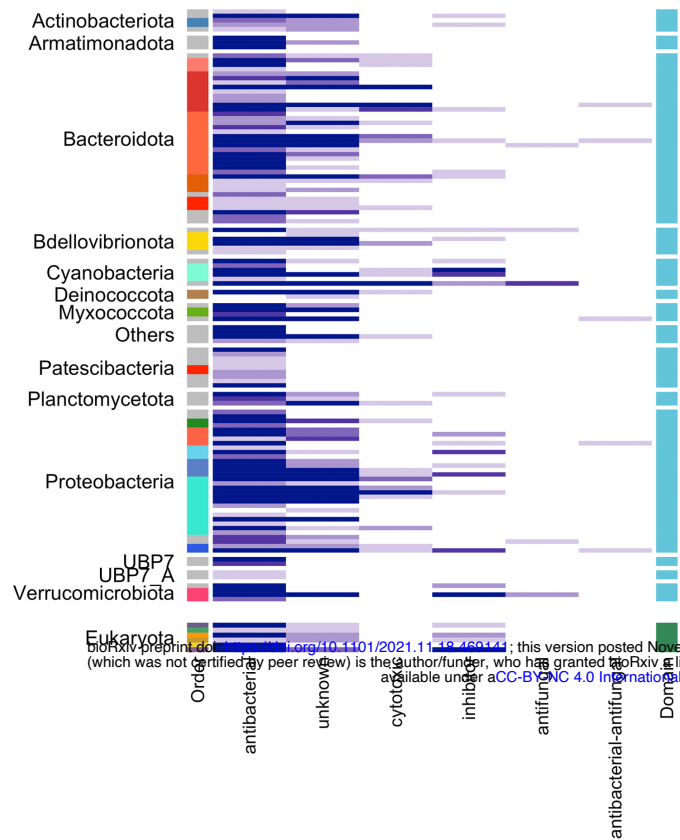
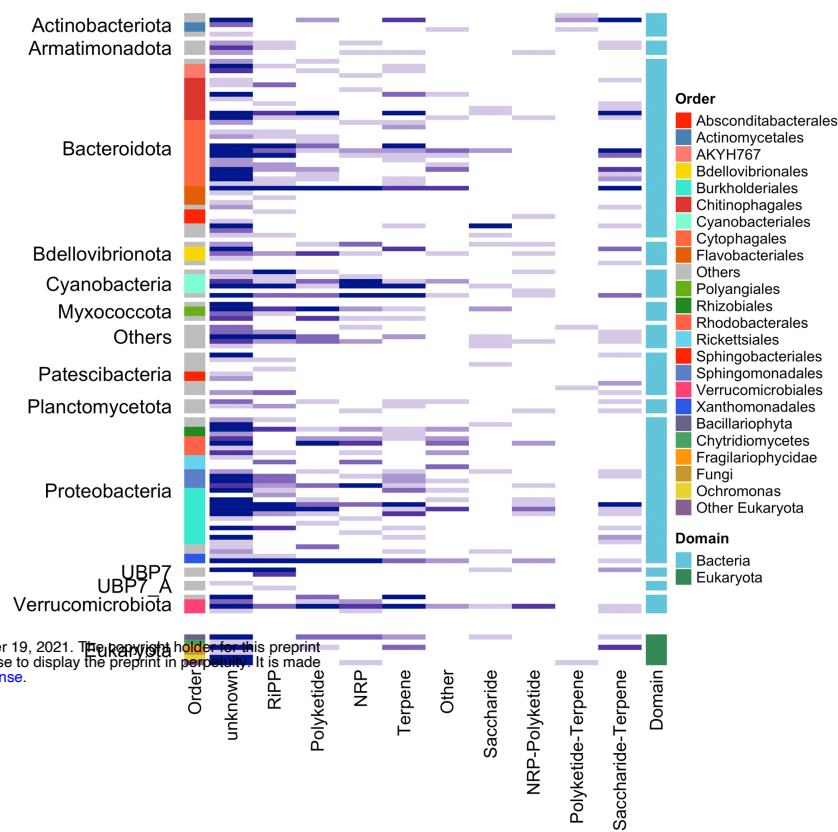


Figure 2

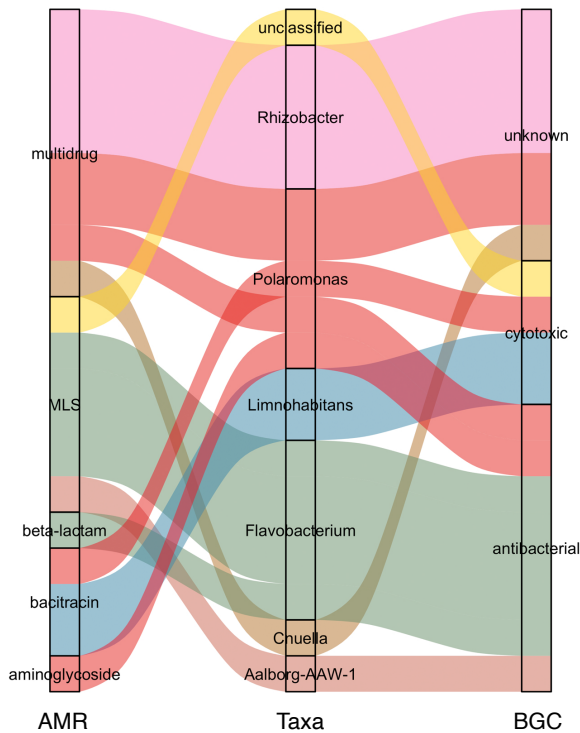
a



b

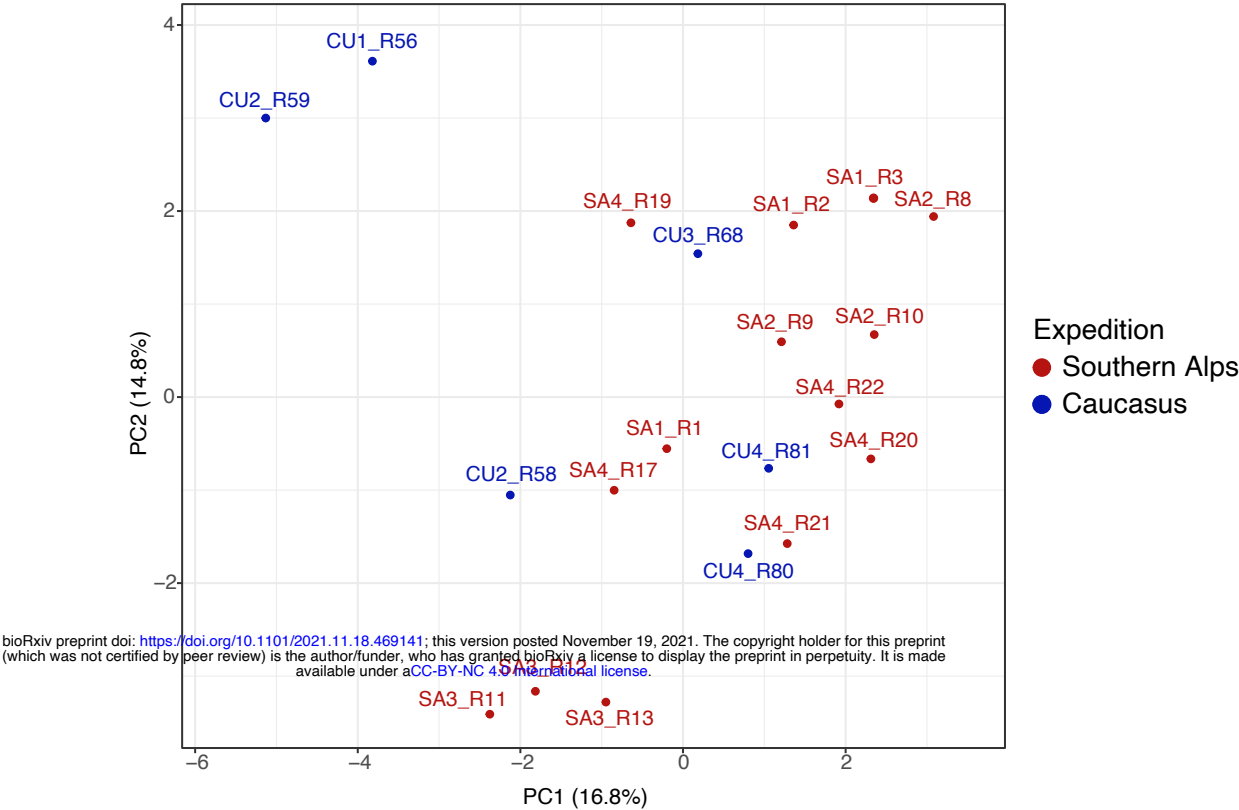


c

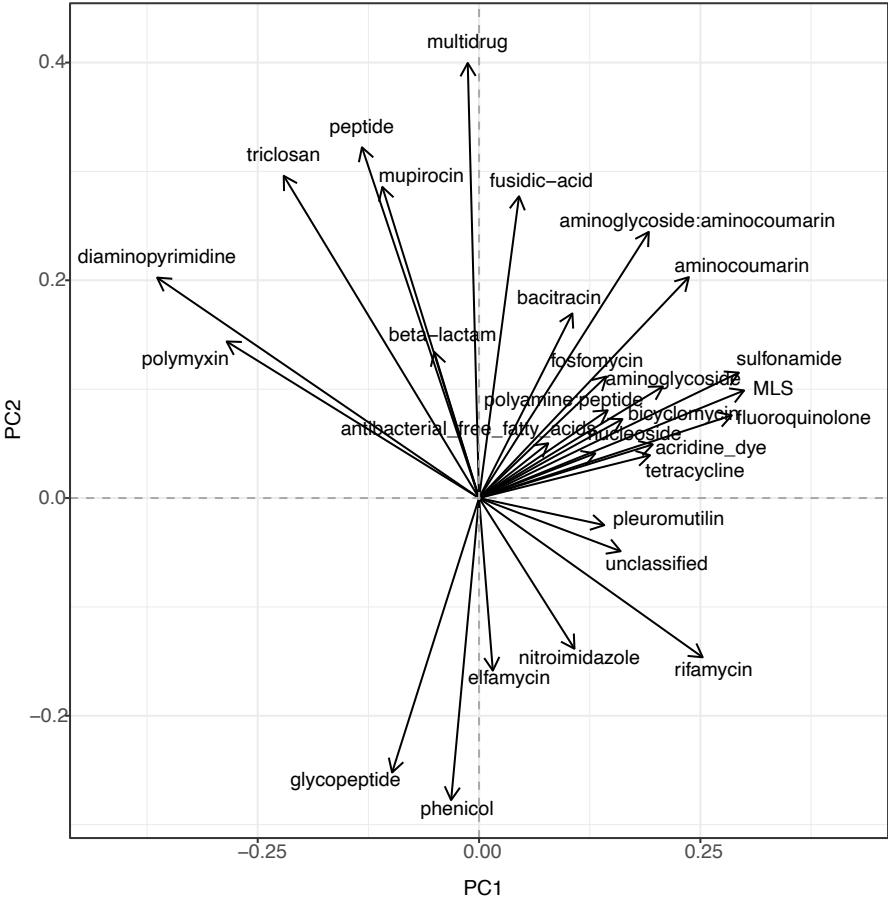


Supplementary Figure 1

a

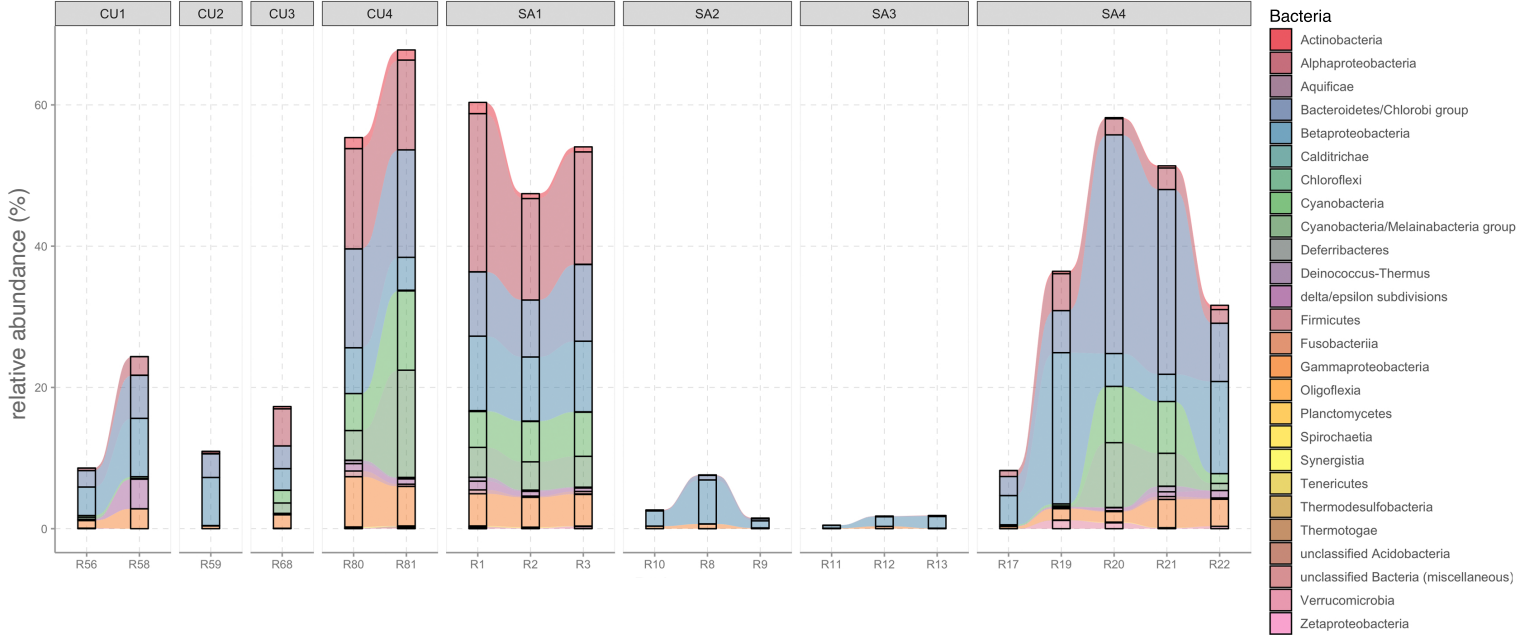


b

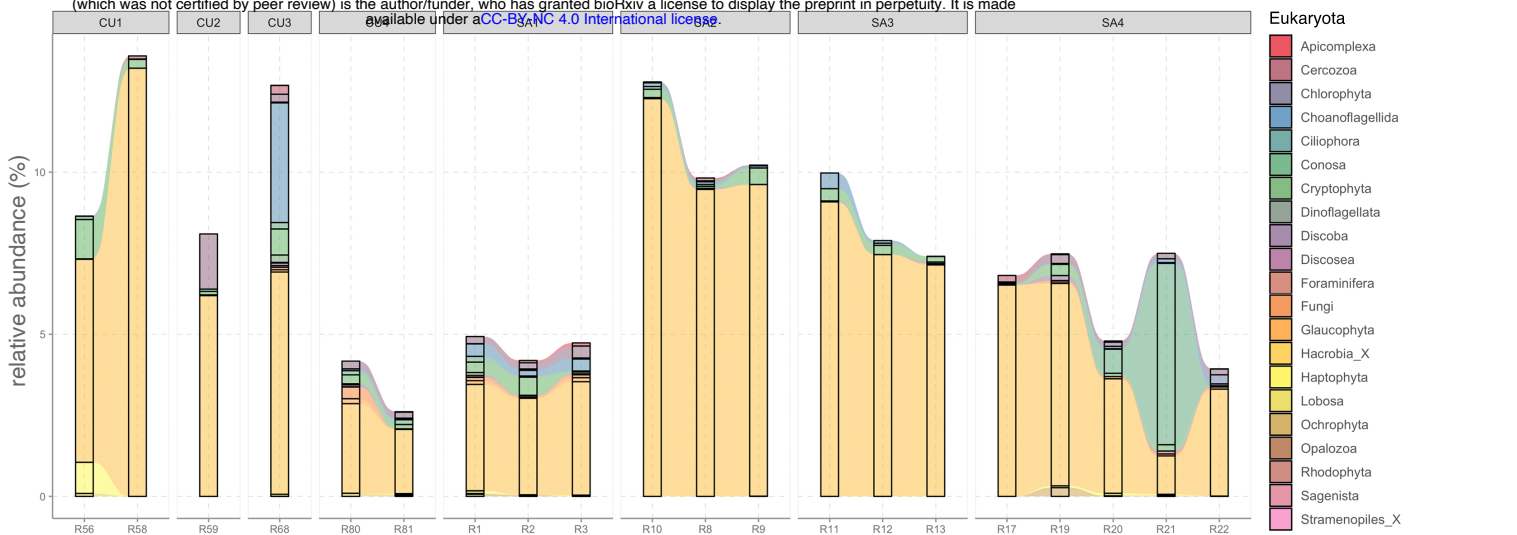


Supplementary Figure 2

a

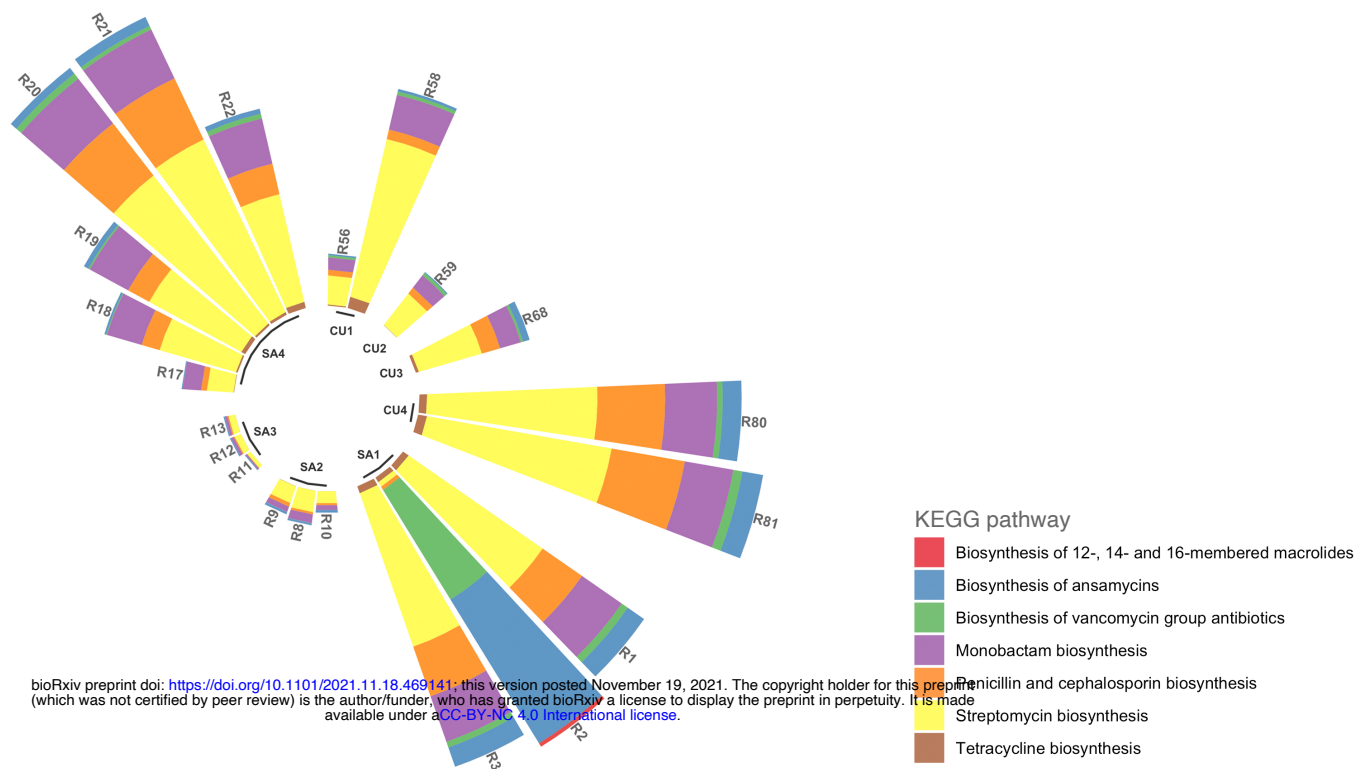


b

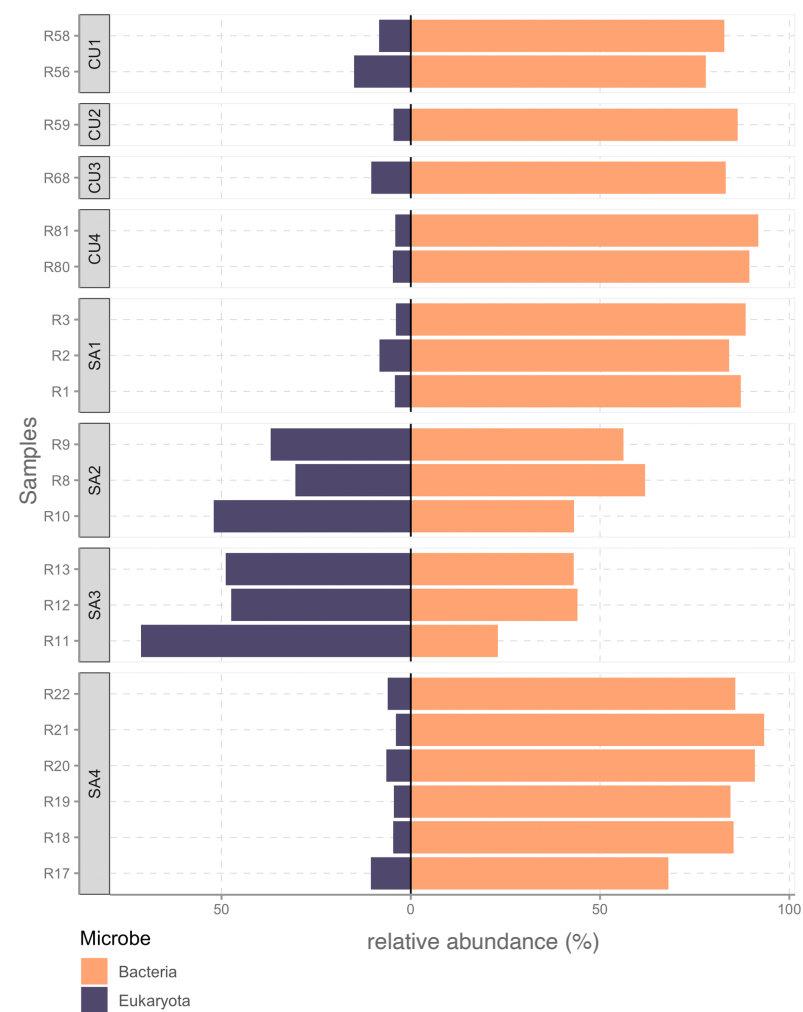


Supplementary Figure 3

a



b



c



Appendix A.6

Functional meta-omics provide critical insights into
long- and short-read assemblies

Functional meta-omics provide critical insights into long- and short-read assemblies

Valentina Galata[†], Susheel Bhanu Busi[†], Benoît Josef Kunath[†], Laura de Nies[†], Magdalena Calusinska[†], Rashi Halder[†], Patrick May[†], Paul Wilmes[†] and Cédric Christian Laczny[†]

Corresponding author: Cédric Christian Laczny, Luxembourg Centre for Systems Biomedicine, 7, avenue des Hauts-Fourneaux, Esch-sur Alzette L-4362, Luxembourg. Tel.: (+352) 46 66 44 9070, Fax: F (+352) 46 66 44 6949; E-mail: cedric.laczny@uni.lu

[†]Equal contribution.

Abstract

Real-world evaluations of metagenomic reconstructions are challenged by distinguishing reconstruction artifacts from genes and proteins present *in situ*. Here, we evaluate short-read-only, long-read-only and hybrid assembly approaches on four different metagenomic samples of varying complexity. We demonstrate how different assembly approaches affect gene and protein inference, which is particularly relevant for downstream functional analyses. For a human gut microbiome sample, we use complementary metatranscriptomic and metaproteomic data to assess the metagenomic data-based protein predictions. Our findings pave the way for critical assessments of metagenomic reconstructions. We propose a reference-independent solution, which exploits the synergistic effects of multi-omic data integration for the *in situ* study of microbiomes using long-read sequencing data.

Key words: third-generation sequencing; long reads; Oxford Nanopore Technologies; hybrid assembly; functional omics; meta-omics

Valentina Galata is a Research Associate in the Systems Ecology Group, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg.

Susheel B. Bhusi is a Research Associate in the Systems Ecology Group, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg.

Benoît J. Kunath is a Research Associate in the Systems Ecology Group, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg.

Laura de Nies is a PhD student in the Systems Ecology Group, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg.

Magdalena Calusinska is a Senior Research & Technology Associate in the BioSystems and Bioprocessing Engineering group, Luxembourg Institute of Science and Technology, Belvaux, Luxembourg.

Rashi Halder is a Research Associate in the Systems Ecology Group and in the LCSB Sequencing Platform, University of Luxembourg, Esch-sur-Alzette, Luxembourg.

Patrick May is a Senior Researcher and the Head of the Genome Analysis group, Bioinformatics Core, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg.

Paul Wilmes is the Head of the Systems Ecology Group, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg, and a Professor in the Department of Life Sciences and Medicine, Faculty of Science, Technology and Medicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg.

Cédric C. Laczny is a Research Scientist in the Systems Ecology Group, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg.

Submitted: 14 June 2021; Received (in revised form): 13 July 2021

© The Author(s) 2021. Published by Oxford University Press.

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

Background

Third-generation, single-molecule, long-read (LR) sequencing is considered to be the next frontier of genomics [1], especially in the context of studying microbial populations [2, 3]. Given the ability to attain read lengths in excess of 10 Kbp [4] and continuous sequence accuracy improvements [5], LR sequencing has been recommended for its ability to resolve GC-rich regions, complex and repetitive loci, and segmental duplications in genomes, among others [4]. However, LR applications to study microbiomes have focused on genome assemblies [6, 7], closing a select few bacterial genomes [8], haplotype and strain resolution [9] as well as mock (low diversity) communities [3]. Stewart et al. [10] recently were among the first to demonstrate the utility of using LRs for improving upon existing protein databases owing to a large collection of novel proteins and enzymes identified, thereby hinting at the benefits of LRs also for functional microbiome studies.

Single base accuracy of raw LRs remains lower—for now—compared with short-read (SR) methodologies [11]; however, Nanopore LR quality is steadily increasing. Several approaches including assembly based and/or including polishing steps have been developed [11–13] to increase the reconstruction accuracy. The impact of remnant errors in LR assemblies on gene calling and thereby protein prediction was recently highlighted by Watson et al. [14]. Hybrid (HY) assembly methods [15, 16] using both SRs and LRs have been proposed to further reduce the error rates compared with LR-only assemblies. Although Watson et al. [14] showed that insertions/deletions (indels) play a critical role in microbial protein identification, the overall impact of assembly methods on understanding the functional potential of microbial communities is lacking.

Here, we demonstrate that metagenomic assembly approaches (SR, LR and HY) not only differ markedly in their overall assembly performance, but also in the inferred functional potential. We reveal the effects of the assembly approach on predicted genes and proteins in samples ranging from low to high diversity, from mock communities to human fecal and rumen metagenomes. We find proteins which are exclusive to respective assemblers and demonstrate using metatranscriptomic and metaproteomic data available for the human fecal sample the synergistic effect on protein verification. Our results indicate that irrespective of sample diversity, the sequencing and assembly strategies impact downstream analyses and that complementary omics are a key for functional analyses of microbiomes.

Results and discussion

To understand how sample diversity, assembly quality and assembly approach are linked, we assembled published metagenomic (metaG) data from a mock community (Zymo), a natural whey starter culture (NWC), a cow rumen sample (Rumen) and a novel metagenomic dataset from a human fecal sample (GDB). The latter was complemented with metatranscriptomic (metaT) and metaproteomic (metaP) data. The samples' diversity ranged from low (Zymo and NWC) to high (GDB and Rumen). As expected [10], the assembly approach strongly affected the quality of the resulting assembly (Supplementary Figure S1). LR and HY approaches generated fewer contigs with a larger N50 value, supporting the added value of these approaches for achieving increased contiguity and decreased redundancy, thereby also improving the recovery of metagenome-assembled

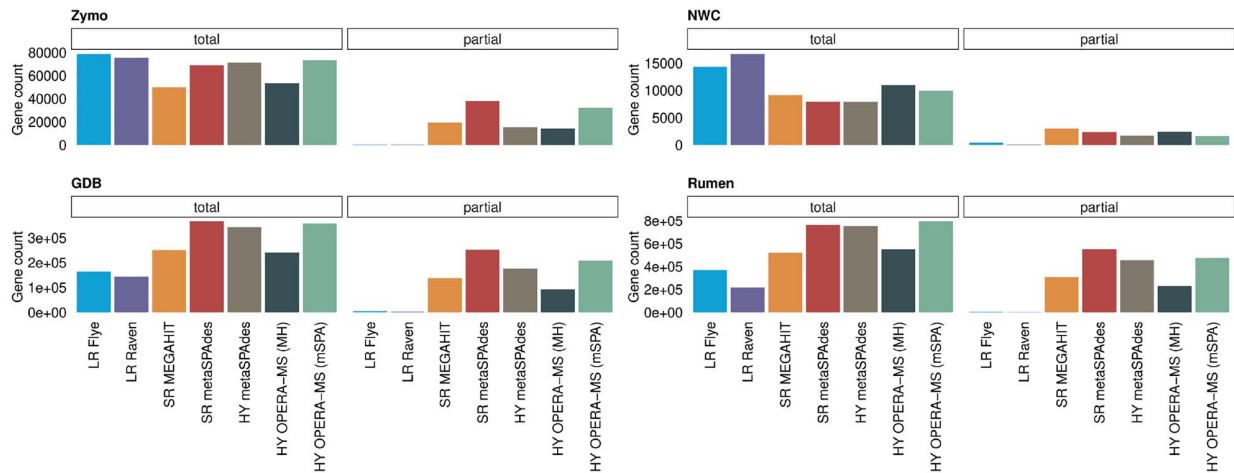
genomes [8]. However, other assembly metrics, e.g. the total assembly length, varied between the samples and assembly types. The metaG read mapping rate (including multi-mapped reads), as a proxy of data usage, was unaffected by the assembler choice when considering all contigs, though the values for the LR assemblies were a bit lower than for SR or HY assemblies of the high-diversity samples (GDB and Rumen). However, the mapping rates dropped markedly in SR assemblies, especially in NWC and Rumen, when filtering out contigs below 5000 bp (Supplementary Figure S2). In GDB, we observed higher metaT read mapping rates in SR and HY assemblies than in LR assemblies. This indicates the complementarity of SR and LR data. The mapping rates decreased considerably in SR assemblies when removing short contigs (Supplementary Figure S3), suggesting the presence of expressed genes located on these contigs. This demonstrates the loss of information when contigs below a certain threshold are removed, which is frequently done in metagenomic studies.

Comparing assemblies pairwise, we observed higher dissimilarities between the LR and SR/HY assemblies than within the latter groups. In addition, OPERA-MS-based HY assemblies clustered together with the SR assemblies on which they were based (Supplementary Figure S4). To assess functional potential overlap between the different assembly approaches, we studied the proteins found in the individual metagenomes. The overall number and quality of predicted proteins was highly influenced by the assembly approach. In highly diverse metagenomes (GDB and Rumen), the total number of proteins in SR and HY assemblies was higher (by a factor of up to 3.67) than in LR assemblies (Figure 1i). However, throughout all samples, the SR and HY approaches produced more partial proteins [incomplete coding sequence (CDS)]. Since SR and HY assemblies may be more fragmented, the polished LR assemblies may have led to an improved recovery of genes. We clustered the predicted protein sequences and found a considerable number of proteins exclusive to individual assemblies. We also found proteins that were shared within a subset of the assemblies only, and that increased sample diversity resulted in an overall increase in the number of exclusive proteins (Figure 1ii).

As reported previously by Watson et al. [14], errors in LR assemblies can have an impact on the predicted proteins. To evaluate how the sample diversity might affect this, we mapped the predicted proteins against the UniProtKB/TrEMBL nonredundant (nr) protein database and computed the query-to-subject length ratio [10]. In all cases, the density distribution of the ratio values had two peaks (below 0.5 and around 1), though the differences between the assembly methods varied across the samples (Supplementary Figure S5). Considering the above findings and despite multiple rounds of polishing, we cannot disregard the impact of (remnant) errors in LRs affecting the results. Furthermore, the results may also be affected by the sequencing depth and gene prediction methods. One also has to account for the microbial composition per sample, given that a large proportion of proteins from the Rumen sample might not have homologs within the UniProtKB/TrEMBL nr database.

Due to the differences in annotations, which we found to be exclusive to individual assembly approaches, we subsequently studied the effect of assembler choice on two well-defined, functionally relevant classes of genes: ribosomal RNA (rRNA) and antimicrobial resistance (AMR) genes. Overall, the total number of rRNA genes recovered by LR and HY approaches was higher across all samples. Within the archaeal and bacterial domains, LR and HY assemblies led to the prediction of more

i.



ii.

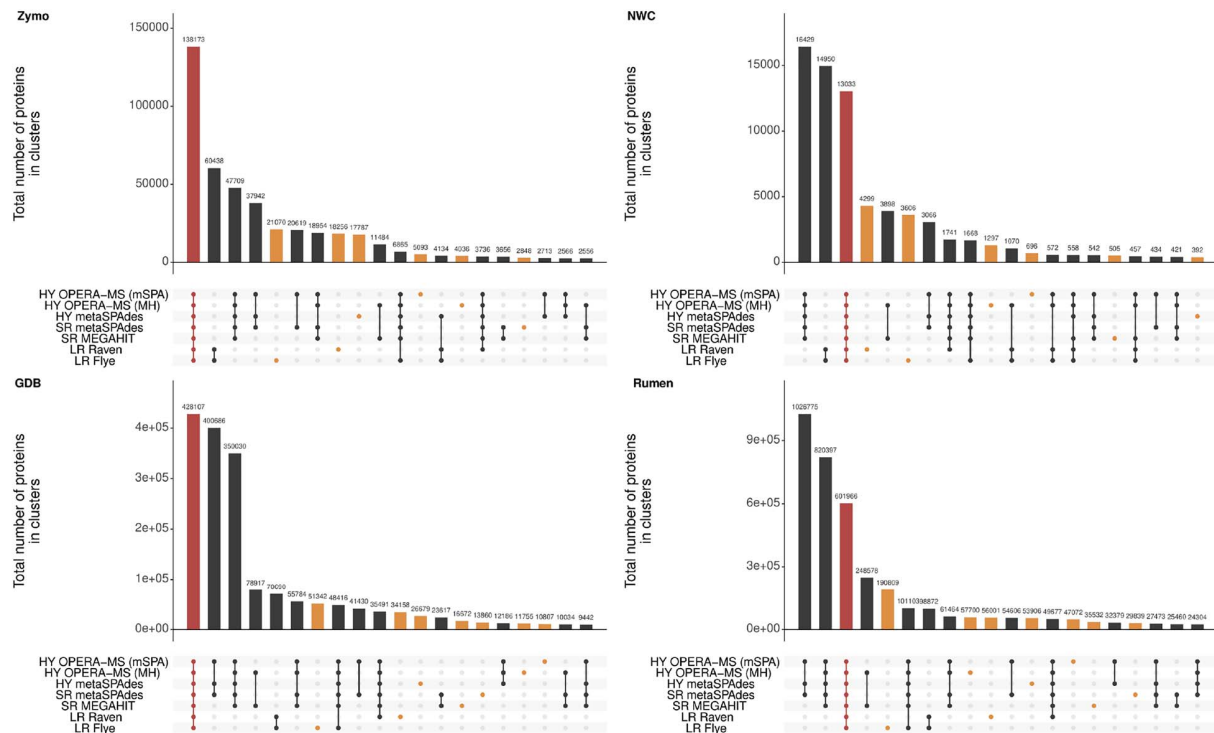
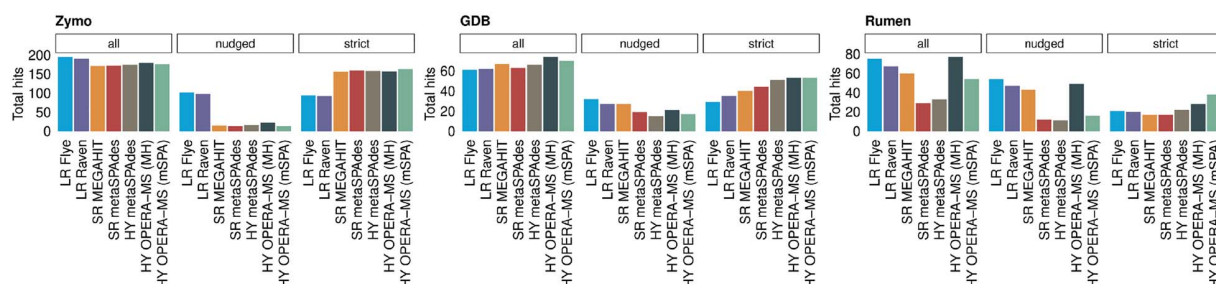


Figure 1. Discrepancy and uniqueness of predicted proteins in assemblies. (i) Number of proteins (total and partial) predicted by Prodigal in each assembly and sample. The color corresponds to the metagenomic assembly approach. (ii) Number of shared predicted proteins which were clustered using MMSseq2 per sample. Each protein cluster was labeled by the combination of assembly tools represented by the clustered proteins (i.e. the assembly where these proteins originated from). The depicted number of shared proteins per assembly tool combination is the total protein count over all associated clusters. Top 20 combinations are shown. The number of proteins found in clusters representing all assembly tools is highlighted in red; the number of proteins exclusive to an assembly is highlighted in orange.

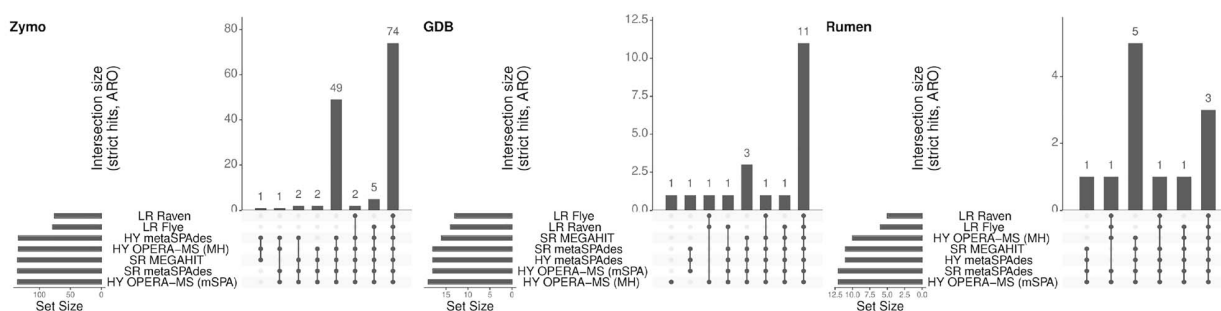
complete genes compared with SR (Supplementary Figure S6). Our findings are in line with Overholt *et al.* [17] and Xie *et al.* [18], who reported improved recovery and contiguity of rRNA genes, and improved gene completeness, respectively. When analyzing AMR proteins and focusing only on 'strict' hits (i.e. excluding loose hits flagged as 'nudged' by the Resistance Gene Identifier (RGI) tool, see Methods), HY assemblers were more adept at

reconstructing these proteins compared with either SR or LR. Moreover, LR assemblies contained more 'nudged' hits than SR or HY assemblies, suggesting that error rates or other factors might have affected the reconstruction of some AMR genes (Figure 2i). Interestingly, we did not identify any AMR hits in the NWC metagenome, possibly due to it being a food-grade additive [19]. When comparing the overlap of the Antibiotic Resistance

i.



ii.



iii.

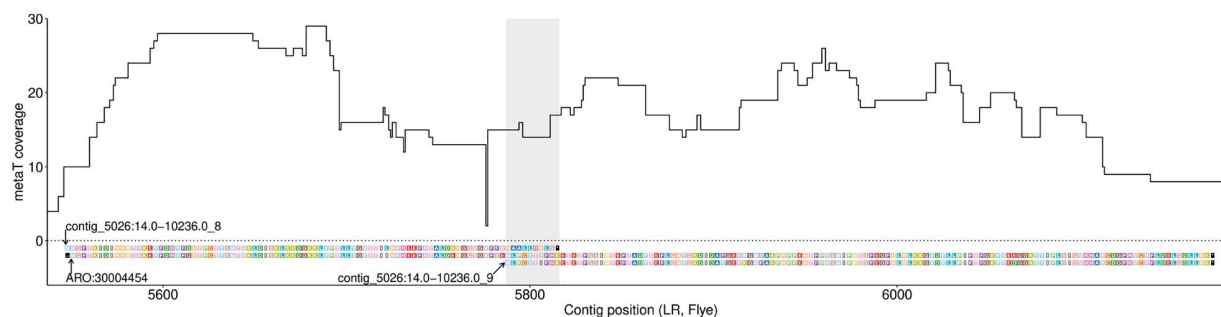


Figure 2. Assembly effects on antimicrobial resistance gene identification. (i) Number of hits ('all', 'strict' and 'nudged') for each assembly and sample when searching the assembly proteins in the CARD database using RGI. The NWC sample is not shown because no hits were found in any of its assemblies. 'Nudged' hits are loose hits (distant/incomplete homolog) flagged as such by RGI; the remaining hits are 'strict' hits. (ii) Number of Antibiotic Resistance Ontologies (AROs), which were covered by 'strict' RGI hits by different assemblies per sample. The bar plot shows the number of shared AROs per assembly tools combination. (iii) Metatranscriptomic (metaT) coverage of the two coding sequences (CDSs) from the long-read (LR) assembly constructed with Flye and having a 'nudged' RGI hit to ARO 3004454 (a chloramphenicol acetyltransferase) in the GDB sample. The x-axis represents the contig coordinates and the y-axis the metaT coverage. The amino acid sequence of the two CDSs and the ARO is included in the plot.

Ontology (ARO) terms covered by 'strict' hits, we found that some AROs were only identified in SR and HY assemblies, but not in LR, whereas no AROs were found in LR assemblies only (Figure 2ii).

To validate the exclusive AROs found in SR and HY assemblies, we assessed metaT and metaP coverage of the corresponding genes and proteins in the GDB sample. The genes mapping to the exclusive AROs had an average metaT coverage above 14× in the SR and HY assemblies, suggesting that these

genes are expressed *in situ*; the few 'nudged' hits were below 6× (Supplementary Table S1). However, we did not identify these genes in the metaP data potentially due to low expression levels, variation in extraction protocols or posttranslational modifications affecting the peptide/proteomic recovery. Though no 'strict' hits were found in LR assemblies, some of their 'nudged' hits had an average metaT coverage above 10×. To understand why these seemingly expressed genes obtained only a partial

hit, we focused on two ‘nudged’ hits assigned to ARO 3004454 (a chloramphenicol acetyltransferase) in the LR assembly constructed with Flye. We found that the CDSs were located on the same contig and had an overlap of 29 bp. The sequence alignments showed that the respective genes represent two fragments of the true CDS (corresponding to ARO 3004454) most likely created by an indel, which introduced a frameshift and also a premature stop codon. This finding was also supported by the metaT coverage extending beyond the stop codon of the first CDS until the end of the second CDS with a single drop in coverage before the putative indel (Figure 2iii). In both the SR and HY assembly approaches, ARO 3004454 obtained a ‘non-nudged’ hit, indicating a potentially complete gene sequence.

To identify high-confidence proteins without the need for a reference, we first considered proteins and protein clusters found in all assemblies, which represented 22.97% of the proteins and 8.54% of the protein clusters. These included genes reconstructed by the different and independent assembly approaches, thus lending mutual support. We then used the complementary metaT data and included all additional proteins with an average metaT coverage $\geq 10\times$ and the corresponding protein clusters. This doubled the number of high-confidence protein clusters (17.63%) and increased the percentage of high-confidence proteins to 30.32%.

Conclusions

We show how the assembler choice but also the assembly strategy or polishing strategy can affect metagenomic reconstruction results when using SR and LR data. The LR assembly approaches studied herein include polishing using SR and LR, whereas the HY assembly approaches construct an initial SR assembly graph, followed by graph traversal using LR and subsequent polishing using SR sequencing data. This distinction is important due to recent developments which leverage LR and SR reads together for the assembly. Furthermore, owing to the existence of established benchmarks [20] for SR/LR/HY assemblers, our analyses specifically address the influence of assemblies on functional discrepancies, which remain understudied thus far. Here, we reveal that sample diversity, along with assembly-mediated effects influence the prediction of genes and proteins. This causes discrepancies between the assemblies, thereby highlighting the potential for complementary means to validate these predictions. The observed discrepancies included conserved and also functionally relevant genes (rRNA and AMR genes, respectively), potentially impacting phylogenetic as well as functional studies. Besides software-driven differences, e.g. assembler choice, the extent of the differences in the metagenomic reconstruction approaches will also depend on associated costs and the thus achievable sequencing depth [21], complexity/composition of the sample and other factors, such as DNA extraction approaches [22], or library preparation methods [23], especially for low-abundance organisms. In addition to our newly generated human-borne multi-omic data (GDB), we used publicly available SR and LR metagenomic data originating from the same respective sample (Zymo, NWC or Rumen). The limited number of samples in this study is due to the limited availability of SR and LR sequence data and do not necessarily represent the extent of diversities across several metagenomic datasets. Although few studies exist to date that have published SR and LR data of the same sample, we expect more datasets to become available in the future due to the advantages that LR data brings for metagenomic reconstructions. To evaluate discrepancies in assembly approaches, we propose a reference-independent approach to identify high-confidence genomic

reconstructions by combining metagenomic, metatranscriptomic and metaproteomic data. The appropriate coverage of metagenomic assemblies via metatranscriptomic reads and the potential presence of peptides mapping to the respective gene and proteins of interest indicate a validated transcription and translation, respectively. Overall, we show that the sequencing approach and assembly strategy can have a significant impact on the characterization of the microbiome’s functional potential and demonstrate the added value of multi-omic strategies for reconstruction quality evaluation, i.e. going beyond their original purpose, to resolve the functional microbiome.

Materials and methods

Sample origin and collection

The datasets (Zymo, NWC, Rumen) used herein were acquired from previously published reports regarding the utility of LR sequencing (Supplementary Table S2), with concomitant SR and LR sequencing data. The human fecal samples were freshly collected from a healthy volunteer (GDB) and immediately flash-frozen in liquid nitrogen. The samples were stored at -80°C until they were processed for biomolecular extraction of DNA, RNA and proteins.

Biomolecular extraction

To obtain high-molecular weight (HMW) DNA, we followed the protocol proposed recently [8], with minor modifications. Frozen stool sample was weighed out in triplicates, to 0.7 g and aliquoted into phase-lock gel tubes (Fisher Scientific, Waltham, MA), along with a 4 mm stainless steel grinding balls (RETSCH 22.455.0003). The sample was subsequently suspended in 500 μl phosphate buffered saline (PBS) (Fisher Scientific, Waltham, MA) with brief gentle vortexing at 10 s intervals repeated five times. Thereafter, 5 μl of lytic enzyme solution (Qiagen, Hilden, Germany) was added and the samples were mixed by gentle inversion six times, and then incubated for 1 h at 37°C ; 12 μl 20% (w/v) sodium dodecyl sulfate (SDS) (Fisher Scientific, Waltham, MA) was added followed by 500 μl phenol:chloroform:isoamyl alcohol at pH 8 (Fisher Scientific, Waltham, MA). The samples were gently vortexed for 5 s, and then centrifuged at 10 000 g for 5 min. The aqueous phase was decanted into a new 2 ml tube. Next, the DNA was precipitated with 90 μl 3 M sodium acetate (Fisher Scientific) and 500 μl isopropanol (Fisher Scientific). After slowly inverting three times, samples were incubated at room temperature for 10 min, followed by centrifugation for 10 min at 10 000 g . The supernatant was removed, and the pellet was washed twice with freshly prepared 80% (v/v) ethanol (Fisher Scientific). Washing was done by adding 1 ml of 80% EtOH, followed by centrifugation for 10 min at 10 000 g . The pellet was then air dried with heating for 10 min at 37°C or until the pellet was matte in appearance, and then resuspended in 100 μl nuclease-free water (Ambion, ThermoFisher Scientific, Waltham, MA). To the pellet, 1 ml Qiagen buffer G2, 4 μl Qiagen RNase A at 100 mg/ml and 25 μl Qiagen Proteinase K were added. The samples were then gently inverted three times and incubated for 90 min at 56°C . After the first 30 min, pellets were dislodged by a single gentle inversion. During the 90 min incubation, one Qiagen Genomic-tip 20/G column per triplicate sample was equilibrated with 1 ml Qiagen buffer QBT and allowed to empty by gravity flow. Samples were gently inverted twice, applied to columns and allowed to flow through. Three stool extractions (triplicates for each sample) were combined per column. Columns were then washed with 3 ml Qiagen buffer QC, where 1 ml of QC buffer was added each time and allowed

to drain the column. Next, the column was placed in a new sterile 1.5 ml Eppendorf tube and the DNA was then eluted with 1 ml of Qiagen buffer QF prewarmed to 56°C. The eluted DNA was then precipitated by addition of 700 µl isopropanol and incubated at room temperature for 10 min, followed by inversion and centrifugation for 15 min at 10 000 g. The supernatant was carefully removed by pipette, and pellets were washed with 1 ml 80% (v/v) ethanol (washing=add 1 ml EtOH, centrifuge for 10 min at 10 000 g). Residual ethanol was removed by air drying 10 min at 37°C, followed by resuspension of the pellet in 100 µl water overnight at 4°C without agitation of any kind. The pooled sample was quantified using the Qubit Broad-Range DNA concentration kit and was estimated at 323.35 ng/µl with an $OD_{260/280} = 1.85$. The extracted HMW DNA was used for both SR and LR sequencing. RNA was extracted from an aliquot of the same fecal sample using PowerMicrobiome RNA isolation kit (cat. no. 26000-50, MoBio) as suggested by the manufacturer. For the protein extractions, a modified protocol based on a previously established sequential extraction method [24] was used. Briefly, proteins were precipitated by adding one volume of All-Prep Protein (APP) Buffer to the flow-through from an independent RNA purification, followed by mixing and incubation for 10 min at room temperature. After incubation, the mixture was centrifuged for 10 min at 12 000 g and the pellet was washed twice in 70% ethanol, with 1 min centrifuge cycles at 12 000 g, and dried at room temperature for 7 min after removing excess ethanol. The pellet was then dissolved in 100 µl ALO buffer and incubated for 5 min at 95°C. After complete dissolution and denaturation of the protein, the sample was cooled to room temperature and centrifuged for 1 min at 12 000 g, from which the supernatant was collected for downstream protein analysis.

Sequencing

SR sequencing

The DNA sample was subjected to random shotgun sequencing. The sequencing library was prepared using KAPA HyperPlus Kit (cat. no. 07962401001, Roche) for the GDB fecal sample using the protocol provided with the kit. Enzymatic fragmentation time was 15 min to aim for 350 bp average size. There was no additional polymerase chain reaction amplification of the prepared library.

RNA sample for metaT analysis was subjected to rRNA depletion using the QIAseq FastSelect 5S/16S/23S kit (cat. no. 335921, Qiagen) for the GDB fecal sample. Library preparation of rRNA-depleted RNA was done using TruSeq Stranded mRNA library preparation kit (cat. no. 20020594, Illumina) according to the protocol provided by the manufacturer with the exception of omitting the initial steps for mRNA pull down.

Both metaG and metaT libraries were quantified using Qubit HS assay (Invitrogen) and their quality was assessed on a Bio-analyzer HS chip (Agilent). We used the NextSeq500 (Illumina) instrument to perform the sequencing using 2 × 150 bp read length at the Luxembourg Centre for Systems Biomedicine (LCSB) Sequencing Platform.

LR sequencing

DNA library for the fecal sample was size selected using AMPure beads for longer fragments. The DNA was sheared using a G-tube (cat. no. 520079, Covaris) aiming for 8 kb average size according to the protocol provided by the manufacturer. Library preparation for LR sequencing was done using the genomic DNA ligation kit

(SQK-LSK109) according to the protocol provided by the manufacturer using a MiniON R9.4.1 flowcell. Once all the library loaded on the flowcell was finished, the library was reloaded after either flowcell wash or nuclease flush. In total, the library was loaded four times to achieve 16 Gbp of sequencing data for this fecal sample (Supplementary Table S3).

Data analysis

Snakemake (v. 5.18.1) [25] was used to implement the analysis workflow. We provide a brief description of the most important steps in the following.

Sequence data preprocessing

Short reads

The raw SRs were trimmed and preprocessed with fastp (v. 0.20.0) [26] with a minimum length of 40 bp. FastQC (v. 0.11.9) [27] reports were generated from the processed FASTQ files. MetaT SRs from the GDB sample were filtered by discarding reads mapping to rRNA gene references included in the repository of SortMeRNA [28] (v4.2.0-10-g1358b9b, <https://github.com/biocore/sortmerna>) using BBDuk from the BBDuk toolkit (v.38.86, kmer length set to 31 bp) [29]. In addition, for the GDB sample, reads mapping to the human genome (GCF_000001405.38_GRCh38.p12) were removed using BBDuk (kmer length set to 31 bp, input and output quality encoding offset set to 33).

Long reads

For each sample except NWC, single-FAST5 files were converted to multi-FAST5 files using `single_to_multi_fast5` from `ont-fast5-api` (v. 3.1.5), the resulting files were basecalled using `guppy` on a GPU node (v. 3.6.0+98ff765, configuration file `dna_r9.4.1_450bps_modbases_dam-dcm-cpg_hac.cfg`, disabled transmission of telemetry pings, chunk size of 1000, 8000 records per FASTQ file) and concatenated into a single FASTQ file. For NWC, no FAST5 were available and, thus, only the provided FASTQ file was used for the analysis. Nanostat (v. 1.1.2) [30] reports were created from the FASTQ files using default parameters. As for the SRs, LR of the GDB sample were filtered to remove reads mapping to the human genome (GCF_000001405.38_GRCh38.p12) using the same parameters.

Metagenomic assembly

Short reads

SR assemblies were done using preprocessed reads and MEGAHIT or metaSPAdes. MEGAHIT (v. 1.2.9) [31] was run using default parameters; metaSPAdes (v. 3.14.1) [32] was run using kmer lengths 21, 33, 55 and 77 bp.

Long reads

LR assemblies were done using Flye and Raven. Flye (v. 2.8.1) [33] was run by providing the (processed) LR in a FASTQ file (input parameter `--nano-raw`) and with the flag `--meta`. Raven (v. 1.2.2) [34] was run with default parameters. Assemblies were polished using LR and SR: one round of Racon (v. 1.4.13) [35] with LR using the flag `--include-unpolished` where reads were mapped to contigs using BWA MEM (v. 0.7.17) [36] with the option `-x ont2d` and processed using samtools (v. 1.9); four rounds of Racon with SRs using the flag `--include-unpolished` where reads were mapped to contigs using Burrows-Wheeler Aligner

(BWA-MEM) and processed using samtools; one round of Medaka (v. 0.8.1) [37] with LR using the model 'r941_min_high'.

Hybrid

HY assemblies, i.e. using SR and LR together, were done using metaSPAdes and OPERA-MS. SPAdes was run with the flag '--meta' and the same k-mer lengths as the SR assemblies by additionally providing the LR using the input parameter flag '--nanopore'. OPERA-MS (v. v0.8.2-63-gc18b4f3) [15] was run using paired SRs, LR and the SR assemblies created by MEGAHIT and metaSPAdes, respectively, using minimap2 [38] as the LR mapper. The assemblies were polished by running five rounds of Racon with SRs as described for the LR assemblies. If not stated otherwise, only polished contigs were used for the LR and HY assemblies in the following analysis steps.

Mapping rate and assembly coverage

For the mapping rate, the used reads were mapped back to the contigs and processed using BWA MEM and samtools in the same way as described above when polishing the LR and HY assemblies using Racon. For HY assemblies, both LR and SR were mapped to the polished contigs and the BAM files were merged using samtools. For the sample GDB, metaT SRs were also separately mapped to the (polished) contigs. Mapping statistics were computed from the BAM files using samtools' options 'flagstat', to determine the number of reads mapping back to the assemblies, and 'idxstats' for per-contig mapping information. For GDB, metaT per-base coverage was computed for each assembly from the BAM files using bedtools (v. 2.29.2) [39] (utility 'genomecov' with the parameter '-d').

Assembly annotation

For each sample and assembly, protein prediction was done using Prodigal (v. 2.6.3) [40] using the option '-p meta'; the keyword 'partial' in the headers of the obtained protein FASTA files was used to distinguish complete and partial proteins. Known antibiotic resistance factors were searched in the predicted proteins (after discarding the stop codon symbol '*' from the FASTA files) by running RGI (v. 5.1.1) [41] together with the CARD database (v. 3.1.0) [42] and DIAMOND (v. 0.8.36) [43] for protein alignments. Loose hits flagged as 'nudged' by the tool were highlighted as such (i.e. as 'nudged') in the downstream analysis.

The tool barrnap (v. 0.9) [44] was run to predict rRNA genes on assembly contigs using the four provided databases of bacterial, archaeal, metazoan mitochondrial, and eukaryotic rRNA genes, respectively. Predictions containing the word 'partial' in their product annotation in the obtained General Feature Format (GFF) files were considered as partial hits.

Analysis

Assembly statistics were computed by running metaQUAST (v. 5.0.2) [45] without using any genome references, setting the minimum contig length to 0 bp and retrieving the statistics for the contig length thresholds of 0, 1000, 2000 and 5000 bp subsequently. Per sample, assemblies were compared using Mash (v. 2.2.2) [46]; sketches were computed per assembly using a k-mer length of 31 bp and a sketch size of 100 000, and pairwise distances were then estimated. Per sample, proteins from all assemblies were clustered using MMseqs2 (v. 12.113e3) [47]. First, a database was created from a concatenated FASTA file

of protein sequences ('--dbtype 1'). Then, option 'linclust' with default parameters was used to perform the clustering and the obtained files were converted to tables using option 'createtsv'. DIAMOND (v. 0.9.25) [43] with the option 'blastp' and default parameters was used to align the predicted proteins against the UniProtKB/TrEMBL database (downloaded and created on 24 August 2019 from http://ftp.uniprot.org/pub/database/uniprot/current_release/knowledgebase/complete/, archive uniprot_trembl.fasta.gz) [48]. The created DIAMOND alignment archive (DAA) files were converted to tables using option 'view' and the parameter '--max-target-seqs 1'. When processing the hits, these were sorted per query and e-value in ascending order and only the first hit was used. For GDB and metaT, using the per-base coverage information computed for each assembly, the average coverage was computed for the corresponding gene sequences of each predicted protein.

MS/MS acquisition and metaproteomic analysis

One microgram of extracted proteins was denatured and loaded on a SDS gel to produce one gel band. The reduction, alkylation and tryptic digestion of the proteins into peptides were performed in-gel. The tryptic peptides were extracted from the gel and desalted prior to mass spectrometry analysis. Peptides were analyzed using a nano Liquid Chromatography-Mass Spectrometry/Mass Spectrometry (nanoLC-MS/MS) system (120 min gradient) connected to a Q-Exactive HF orbitrap mass spectrometer (Thermo Scientific, Germany) equipped with a nano-electrospray ion source. The Q-Exactive mass spectrometer was operated in data-dependent mode and the 10 most intense peptide precursor ions were selected for fragmentation and MS/MS acquisition.

For each assembly separately and for all assemblies together, the FASTA file of predicted proteins was concatenated with a common Repository of Adventitious Proteins (cRAP) database of contaminants [49] and with the human UniProtKB Reference Proteome prior metaproteomic search. In addition, reversed sequences of all protein entries were concatenated to the databases for the estimation of false discovery rates (FDRs). The search was performed using SearchGUI-3.3.20 [50] with the X!Tandem [51], MS-GF+ [52] and Comet [53] search engines and the following parameters: trypsin was used as the digestion enzyme and a maximum of two missed cleavages was allowed. The tolerance levels for matching to the database were 10 ppm for MS1 and 0.02 Da for MS2. Carbamidomethylation of cysteine residues was set as a fixed modification and protein N-terminal acetylation and oxidation of methionines was allowed as variable modification. Peptides with length between 7 and 60 amino acids and with a charge state composed between +2 and +4 were considered for identification. The results from SearchGUI were merged using PeptideShaker-1.16.45 [54] and all identifications were filtered in order to achieve a protein FDR of 1%.

Plots

Figures were generated in R (v. 4.0.2, <https://www.r-project.org/>) using, *inter alia*, Pheatmap (v. 1.0.12, <https://github.com/raivokolde/pheatmap>) for heatmap plots, UpSetR (v. 1.4.0) [55] for intersection plots, ggplot2 (v. 3.3.2) [56] and its various extensions for other plot types, color palettes from the viridis (v. 0.5.1, <https://github.com/sjmgarnier/viridis>) and ggsci (v. 2.9, <https://github.com/road2stat/ggsci>) packages and the patchwork package (v. 1.1.1, <https://github.com/thomasp85/patchwork>) for combining plots.

Key Points

- Sequencing and assembly approach affect gene and protein inference.
- Meta-omics enable critical assessment of metagenome reconstructions.
- Reference-independent solution which exploits synergies of next-generation and third-generation sequencing approaches that results in improved integration of meta-omics data.

Authors' contributions

S.B.B., V.G. and C.C.L. designed the study. S.B.B. and R.H. performed the biomolecular extractions, whereas R.H. performed the metagenomic and metatranscriptomic sequencing. V.G., S.B.B., L.deN. and C.C.L. analyzed the data. B.J.K. performed the metaproteomic analyses. P.M., M.C. and P.W. provided critical feedback and insights. All authors contributed to the writing and revision of the manuscript.

Supplementary Data

Supplementary data are available online at Briefings in Bioinformatics.

Acknowledgments

The authors are thankful for the assistance of Audrey Frachet Bour, Lea Grandmougin, Janine Habier and Laura Lebrun (LCBS) for laboratory support. The experiments presented in this paper were carried out using the High Performance Computing (HPC) facilities of the University of Luxembourg (<https://hpc.uni.lu> [57]).

Abbreviations

SR, short reads; LR, long reads; HY, hybrid (approach/assembly); metaG, metagenomic (data); metaT, metatranscriptomic (data); metaP, metaproteomic (data); AMR, antimicrobial resistance; rRNA, ribosomal RNA; ARO, Antibiotic Resistance Ontology; CDS, coding sequence; HMW, high-molecular weight; FDR, false discovery rate; indels, insertions/deletions; nr, nonredundant

Ethics approval and consent to participate

This study conformed to the Declaration of Helsinki and was approved by the ethics committee of the Physician's Board Hessen, Germany (FF38/2016).

Consent for publication

All authors acknowledge the content of this manuscript and consent to its publication.

Availability of data and materials

Processed sequencing data of the GDB sample is available under BioProject accession PRJNA723028 (Biosamples: metag_sr: SAMN18797629, metat_sr: SAMN18797630 and metag_lr: SAMN18797631). Metaproteomics data of the GDB

sample is available at ProteomeXchange under accession PXD025505. The code used for the analysis is available at <https://doi.org/10.17881/sgzt-ad12> (v1.0) and supplementary data of relevant results is available at <https://doi.org/10.6084/m9.figshare.14447559>.

Funding

Supported by the Luxembourg National Research Fund (PRIDE17/11823097 and C19/BM/13684739 awarded to PW; C17/SR/11687962 awarded to MC). Supported by the Swiss National Science Foundation (Synergia grant CRSII5_180241 awarded to Tom Battin). PW acknowledges "Probiotics in external applications (PBGL)" and the European Research Council (ERC-CoG 863664).

References

1. Burgess DJ. Genomics: next regeneration sequencing for reference genomes. *Nat Rev Genet* 2018;19:125.
2. Amarasinghe SL, Su S, Dong X, et al. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol* 2020;21:30.
3. Nicholls SM, Quick JC, Tang S, et al. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience* 2019;8: giz043. <https://doi.org/10.1093/gigascience/giz043>.
4. Pollard MO, Gurdasani D, Mentzer AJ, et al. Long reads: their purpose and place. *Hum Mol Genet* 2018;27:R234–41.
5. Jain M, Koren S, Miga KH, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* 2018;36:338–45.
6. Goldstein S, Beka L, Graf J, et al. Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. *BMC Genomics* 2019;20:23.
7. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet* 2020;21:597–614.
8. Moss EL, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat Biotechnol* 2020;38:701–7.
9. Nicholls SM, Aubrey W, De Grave K, et al. On the complexity of haplotyping a microbial community. *Bioinformatics* 2020;37(10):1360–6. doi: [10.1093/bioinformatics/btaa977](https://doi.org/10.1093/bioinformatics/btaa977). Epub ahead of print. PMID: 33444437; PMCID: PMC8208737.
10. Stewart RD, Auffret MD, Warr A, et al. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat Biotechnol* 2019;37:953–61.
11. Zhang H, Jain C, Aluru S. A comprehensive evaluation of long read error correction methods. *BMC Genomics* 2020;21:889.
12. Ryan R, Wick KEH. Benchmarking of long-read assemblers for prokaryote whole genome sequencing. *F1000Research* 2021;8:2138.
13. Dohm JC, Peters P, Stralis-Pavese N, et al. Benchmarking of long-read correction methods. *NAR Genom Bioinform* 2020;2:lqaa037.
14. Watson M, Warr A. Errors in long-read assemblies can critically affect protein prediction. *Nat Biotechnol* 2019;37:124–6.
15. Bertrand D, Shaw J, Kalathiyappan M, et al. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat Biotechnol* 2019;37:937–44.
16. Haghshenas E, Asghari H, Stoye J, et al. HASLR: Fast Hybrid Assembly of Long Reads. *iScience* 2020;23:101389.

17. Overholt WA, Hölzer M, Geesink P, et al. Inclusion of Oxford Nanopore long reads improves all microbial and viral metagenome-assembled genomes from a complex aquifer system. *Environ Microbiol* 2020;**22**:4000–13.
18. Xie H, Yang C, Sun Y, et al. PacBio long reads improve metagenomic assemblies, gene catalogs, and genome binning. *Front Genet* 2020;**11**:516269.
19. Somerville V, Lutz S, Schmid M, et al. Long-read based de novo assembly of low-complexity metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. *BMC Microbiol* 2019;**19**:143.
20. Brown CL, Keenum IM, Dai D, et al. Critical evaluation of short, long, and hybrid assembly for contextual analysis of antibiotic resistance genes in complex environmental metagenomes. *Sci Rep* 2021;**11**:3753.
21. Zaheer R, Noyes N, Ortega Polo R, et al. Impact of sequencing depth on the characterization of the microbiome and resistome. *Sci Rep* 2018;**8**:5890.
22. Sui H-Y, Weil AA, Nuwagira E, et al. Impact of DNA extraction method on variation in human and built environment microbial community and functional profiles assessed by shotgun metagenomics sequencing. *Front Microbiol* 2020;**11**:953.
23. Peng Z, Zhu X, Wang Z, et al. Comparative analysis of sample extraction and library construction for shotgun metagenomics. *Bioinform Biol Insights* 2020;**14**:117793220915459.
24. Roume H, Heintz-Buschart A, Muller EEL, et al. Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample. *Methods Enzymol* 2013;**531**:219–36.
25. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 2012;**28**:2520–2.
26. Chen S, Zhou Y, Chen Y, et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**:i884–90.
27. Andrews S. FastQC: a quality control tool for high throughput sequence data [Online]. 2010. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
28. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 2012;**28**:3211–7.
29. Bushnell B. BBMap: A Fast, Accurate, Splice-Aware Aligner. United States: N. p., 2014. Web.
30. De Coster W, D'Hert S, Schultz DT, et al. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* 2018;**34**:2666–9.
31. Li D, Liu C-M, Luo R, et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;**31**:1674–6.
32. Nurk S, Meleshko D, Korobeynikov A, et al. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 2017;**27**:824–34.
33. Kolmogorov M, Bickhart DM, Behsaz B, et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* 2020;**17**:1103–10.
34. Vaser R, Šikić M. Time- and memory-efficient genome assembly with Raven. *Nat Comput Sci* 2021;**1**:332–336.
35. Vaser R, Sović I, Nagarajan N, et al. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* 2017;**27**:737–46.
36. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013. arXiv:Genomics.
37. GitHub - nanoporetech/medaka: Sequence correction provided by ONT Research. Available online at <https://github.com/nanoporetech/medaka>.
38. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100.
39. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2.
40. Hyatt D, Chen G-L, LoCascio PF, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;**11**:119.
41. Jia B, Raphenya AR, Alcock B, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res* 2017;**45**:D566–73.
42. Alcock BP, Raphenya AR, Lau TTY, et al. CARD 2020: antibiotic resistance surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res* 2020;**48**:D517–25.
43. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**:59–60.
44. Seemann T. barnap 0.9: rapid ribosomal RNA prediction. 2013. Available online at <https://github.com/tseemann/barnap>.
45. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 2016;**32**:1088–90.
46. Ondov BD, Treangen TJ, Melsted P, et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 2016;**17**:132.
47. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;**35**:1026–8.
48. UniProt Consortium. The universal protein resource (UniProt). *Nucleic Acids Res.* 2008;**36**(Database issue):D190–D195.
49. cRAP protein sequences. Available online at <https://www.thegpm.org/crap> [common Repository of Adventitious Proteins, v 2012.01.01; The Global Proteome Machine].
50. Barsnes H, Vaudel M. SearchGUI: a highly adaptable common interface for proteomics search and de novo engines. *J Proteome Res* 2018;**17**:2552–5.
51. Langella O, Valot B, Balliau T, et al. X!TandemPipeline: a tool to manage sequence redundancy for protein inference and phosphosite identification. *J Proteome Res* 2017;**16**:494–503.
52. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat Commun* 2014;**5**:5277.
53. Eng JK, Jahan TA, Hoopmann MR. Comet: an open-source MS/MS sequence database search tool. *Proteomics* 2013;**13**:22–4.
54. Vaudel M, Burkhardt JM, Zahedi RP, et al. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat Biotechnol* 2015;**33**:22–4.
55. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 2017;**33**:2938–40.
56. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2016. Springer-Verlag New York. ISBN 978-3-319-24277-4, <https://ggplot2.tidyverse.org>.
57. S. Varrette, P. Bouvry, H. Cartiaux, et al. “Management of an academic HPC cluster: The UL experience,” 2014 International Conference on High Performance Computing & Simulation (HPCS). 2014, pp. 959–967.

Appendix A.7
Challenges, Strategies, and
Perspectives for Reference-
Independent Longitudinal Multi-Omic
Microbiome Studies



Challenges, Strategies, and Perspectives for Reference-Independent Longitudinal Multi-Omic Microbiome Studies

Susana Martínez Arbas^{1*}, Susheel Bhanu Busi¹, Pedro Queirós¹, Laura de Nies¹, Malte Herold², Patrick May¹, Paul Wilmes^{1,3}, Emilie E. L. Muller⁴ and Shaman Narayanasamy¹

¹Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg, ²Department of Environmental Research and Innovation, Luxembourg Institute of Science and Technology, Belvaux, Luxembourg, ³Department of Life Sciences and Medicine, Faculty of Science, Technology and Medicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg, ⁴Université de Strasbourg, UMR 7156 CNRS, Génétique Moléculaire, Génomique, Microbiologie, Strasbourg, France

OPEN ACCESS

Edited by:

Himel Mallick,
Merck, United States

Reviewed by:

Cecilia Noecker,
University of California,
San Francisco, United States
Siyuan Ma,
University of Pennsylvania,
United States

*Correspondence:

Susana Martínez Arbas
susana.martinez@uni.lu

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 09 February 2021

Accepted: 30 April 2021

Published: 14 June 2021

Citation:

Martínez Arbas S, Busi SB, Queirós P, de Nies L, Herold M, May P, Wilmes P, Muller EEL and Narayanasamy S (2021) Challenges, Strategies, and Perspectives for Reference-Independent Longitudinal Multi-Omic Microbiome Studies. *Front. Genet.* 12:666244. doi: 10.3389/fgene.2021.666244

In recent years, multi-omic studies have enabled resolving community structure and interrogating community function of microbial communities. Simultaneous generation of metagenomic, metatranscriptomic, metaproteomic, and (meta) metabolomic data is more feasible than ever before, thus enabling in-depth assessment of community structure, function, and phenotype, thus resulting in a multitude of multi-omic microbiome datasets and the development of innovative methods to integrate and interrogate those multi-omic datasets. Specifically, the application of reference-independent approaches provides opportunities in identifying novel organisms and functions. At present, most of these large-scale multi-omic datasets stem from spatial sampling (e.g., water/soil microbiomes at several depths, microbiomes in/on different parts of the human anatomy) or case-control studies (e.g., cohorts of human microbiomes). We believe that longitudinal multi-omic microbiome datasets are the logical next step in microbiome studies due to their characteristic advantages in providing a better understanding of community dynamics, including: observation of trends, inference of causality, and ultimately, prediction of community behavior. Furthermore, the acquisition of complementary host-derived omics, environmental measurements, and suitable metadata will further enhance the aforementioned advantages of longitudinal data, which will serve as the basis to resolve drivers of community structure and function to understand the biotic and abiotic factors governing communities and specific populations. Carefully setup future experiments hold great potential to further unveil ecological mechanisms to evolution, microbe-microbe interactions, or microbe-host interactions. In this article, we discuss the challenges, emerging strategies, and best-practices applicable to longitudinal microbiome studies ranging from sampling, biomolecular extraction, systematic multi-omic measurements, reference-independent data integration, modeling, and validation.

Keywords: microbiome, metatranscriptomics, metaproteomics, time-series, metagenomics, metabolomics, *de novo* assembly

INTRODUCTION

Advances in the study of microbial communities have highlighted their important role in natural processes, including those considered as ecosystem services for humankind (Bodelier, 2011). Complex dynamics in microbiomes at the level of composition and structure, as well as function (Heintz-Buschart and Wilmes, 2018) stem from constant adaptation of a given community toward fluctuations of abiotic and biotic factors. However, the fate of these microbial consortia in the face of perturbations is often not understood nor predictable (Muller, 2019). Longitudinal approaches are necessary to understand microbial community dynamics, as they may offer valuable insights into temporal trends and consequences of environmental forcings, when used in tandem with host-derived (Heintz-Buschart et al., 2016; Lloyd-Price et al., 2019; Mars et al., 2020) or environmental (Law et al., 2016; Herold et al., 2020) data. Longitudinal studies can be conducted using diachronic or synchronic approaches (Costa Junior et al., 2013). Herein, we discuss the capacity of longitudinal diachronic approaches as a critical tool toward studying microbial communities. We will further focus on multi-omics longitudinal studies, which leverage the power of the entire high-throughput meta-omic spectrum, namely meta-genomics (MG), -transcriptomics (MT), -proteomics (MP), and -metabolomics (MM), as they are now more feasible and affordable than ever before (Narayanasamy et al., 2015).

Overall, longitudinal multi-omics will enhance our understanding of microbial community dynamics, which could potentially bring about positive outcomes in biomedicine, biotechnology, and for the environment. However, various aspects must be considered when conducting longitudinal multi-omic microbiome studies,

ranging from experimental design, bioinformatic processing, modeling, and validation. In this article, we explore challenges, considerations, and potential solutions for such studies, based on recent advances and reports (Law et al., 2016; Lloyd-Price et al., 2019; Herold et al., 2020; Martínez Arbas et al., 2021), which are applicable to both microbe-centric (e.g., soil, water) or host-centric (e.g., human gut) systems. Finally, although this article focuses on specifically longitudinal multi-omic microbiome studies, the content is generally applicable to any large-scale microbiome studies.

MULTI-OMIC CONSIDERATIONS AND EXPERIMENTAL DESIGN FOR LONGITUDINAL STUDIES

Integration of multi-omic microbiome datasets has been routinely performed, with notable instances, including studies on type-1 diabetes (Heintz-Buschart et al., 2016), cancer (Kaysen et al., 2017), healthy human gut (Tanca et al., 2017), Crohn's disease (Erickson et al., 2012), and activated sludge (Muller et al., 2014; Roume et al., 2015; Yu et al., 2019). These studies clearly demonstrate the maturity of the current microbiome multi-omics toolbox. Despite this, and to the best of our knowledge, equivalent multi-omic surveys based on extensive longitudinal microbiome sampling remain rather limited. **Table 1** lists several relevant studies of longitudinal (at least six timepoints) and multi-omic (at least two omic levels, excluding 16S amplicon sequencing) microbiome datasets.

The famous adage “*absence of evidence is not evidence of absence*” (Altman and Bland, 1995) could likely be a prelude to most microbiome studies. Hence, we discuss these studies in the context of reference-independent bioinformatics

TABLE 1 | Longitudinal multi-omic microbiome datasets and studies.

System	Sample type	Duration*	Frequency*	Total of samples	MG	MT	MP	MM	Complementary data	Studies
Human gut microbiome	Stool samples from 132 humans; healthy or with Crohn's disease or ulcerative colitis	1 year	Bi-weekly	2,965	x	x	x	x	Host genomics, transcriptomics bisulfite sequencing, serologic profiles, diet surveys, and fecal calprotectin	Lloyd-Price et al., 2019 Ruiz-Perez et al., 2021
	Stool samples of 77 individuals	6 months	Monthly	474	x			x	Host transcriptome, metabolome, cytokines, methylome, dietary survey, and physiology	Blasche et al., 2021
Activated sludge	Floating sludge islets from a single anoxic tank	1.5 year	Weekly	53	x	x	x	x	Temperature, pH, oxygen concentration, conductivity, inflow, nitrate concentration, and extracellular metabolites	Herold et al., 2020 Martínez Arbas et al., 2021
	Full- and lab-scale activated sludge	2.5 months	Weekly	10	x	x			Temperature, pH, redox potential and dissolved oxygen	Law et al., 2016

Longitudinal multi-omic data must be of least six timepoints and at least two meta-omic readouts excluding 16S amplicon sequencing. Omics data derived from host(s) are considered separate from the microbial meta-omic spectra.

*Approximate values.

approaches, centered around *de novo* assemblies of sequencing data (MG and MT), subsequently complemented by additional omics (MP and MM, depending on their availability; **Figure 1**). Reference-independent approaches offer asymmetric advantages and opportunities in discovering novel microbial taxa and/or functionalities (Celaj et al., 2014; Narayanasamy et al., 2015; Lapidus and Korobeynikov, 2021), compared to reference-dependent methodologies (Sunagawa et al., 2013; Treangen et al., 2013). Moreover, the integration of multi-omics has been shown to yield superior output compared to single omic studies. For instance, the co-assembly of MG and MT sequencing reads was shown to improve the quality of assembled contigs (Narayanasamy et al., 2016), which in turn improves taxonomic annotation, gene calling/annotation, binning, metabolic pathway (re) construction (Muller et al., 2018; Zhou et al., 2020;

Zimmermann et al., 2021), and quantification of features, e.g., taxa/genes (Narayanasamy et al., 2016). Similarly, MP spectra searches are more effective when performed against gene databases derived from MG assemblies of the same sample/environment, compared to generic databases, thus improving the recruitment of measured peptides (Tanca et al., 2016; Heyer et al., 2017; Timmins-Schiffman et al., 2017). Moreover, such a reference-independent approach may be necessary for microbial communities that are not well characterized and lack extensive unified genome or gene catalogues, such as those available for the human gut microbiome (Li et al., 2014; Almeida et al., 2021). However, most microbial communities are heterogeneous, which further complicates downstream multi-omic data processing, integration, curation, transformation, and modeling (Jiang et al., 2019). Therefore, the adherence toward standards

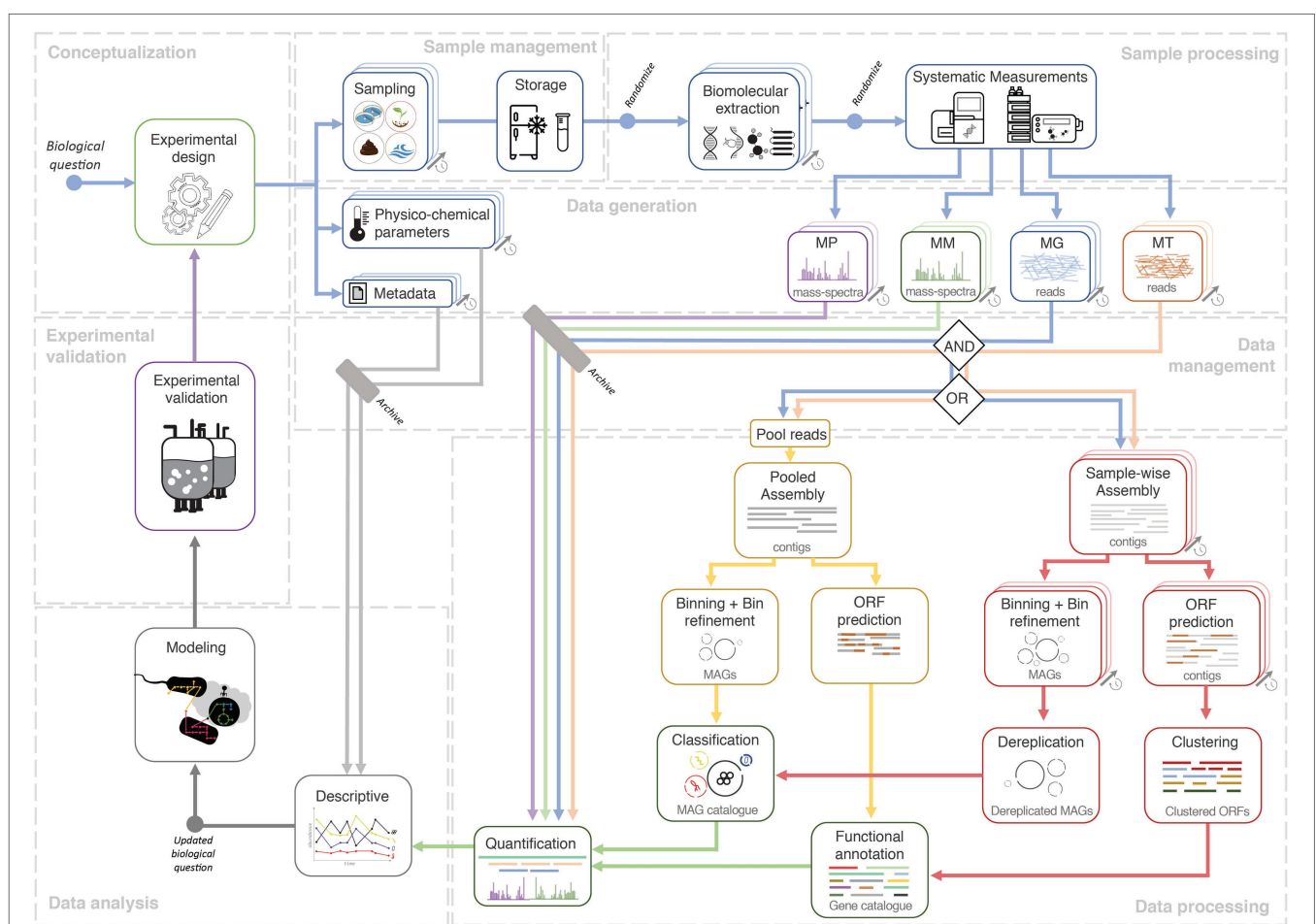


FIGURE 1 | Systems ecology workflow for longitudinal multi-omic microbiome studies. A study conceptualized via an experimental design phase and an initial biological question which is then followed by sample collection, sample management, and systematic high-throughput measurements. The next-generation sequencing (NGS) data could either undergo aggregated processing (yellow track) involving a pooled *de novo* assembly of NGS reads from all longitudinal samples, to eventually yield a metagenome assembled genome (MAG) and/or gene catalogue via binning and gene calling, respectively. In the dereplication approach (red track), data from each sample are first processed in a sample-wise manner, namely the steps of *de novo* assembly, binning, and gene calling. The resulting MAGs and predicted ORFs are then merged through a process called dereplication which generates the catalogue. The availability of a catalogue allows quantification whereby the output could be used for descriptive analyses which could potentially lead to updated or entirely novel biological questions. Quantified values, combined with descriptive analyses, could then be used within dynamic or metabolic models (gray track). Validation of models could lead to further *in situ* longitudinal experimental designs. Finally, all data (raw input, output, metadata) and code (not depicted) should be archived under a data and code management strategy. Free icons were used from <https://www.flaticon.com> (creators: Freepik, Gregor Cresnar, Freepik, and Smashicons).

and best-practices, spanning from sampling to data analyses is important to the outcome of a project. Accordingly, **Figure 1** illustrates the potential lifecycle of a longitudinal multi-omic microbiome study.

Longitudinal multi-omic studies require systematic and thorough study designs that consider sampling parameters (Gerber, 2014; Cao et al., 2017; Liang et al., 2020), metadata, and complementary measurements, such as physico-chemical parameters or questionnaires (Kumar et al., 2014), all of which affect downstream analyses. Sampling parameters, such as duration and frequency, are dictated by the inherent properties of a given microbial system. For instance, the sampling duration when studying gut microbiome development of neonates could span from birth until a “mature” gut microbiome composition is achieved (Stewart et al., 2018), which may vary from subject to subject. Naturally-occurring microbial systems that are exposed to the environment may exhibit annual cyclical behavior based on seasonality and, therefore, could be sampled for at least one complete season-to-season cycle (Johnston et al., 2019). Sampling frequency may be determined by the dynamics and/or generational-timescale of a given system. For instance, the human gut microbiome is known to exhibit daily fluctuations, and therefore could be sampled on a daily basis within a given temporal study (David et al., 2014), while activated sludge systems are known to exhibit (approximately) weekly doubling periods and thus could be sampled on a weekly basis (Herold et al., 2020; Martínez Arbas et al., 2021). Based on the recommendations of Sefer et al. (2016), if biological replicates are either not feasible (i.e., $n = 1$) or limited (i.e., low n) (Herold et al., 2020), one should ideally opt for higher frequency (dense) longitudinal sampling, and less dense sampling if biological replicates were available (i.e., high n), e.g., a cohort of patients (Lloyd-Price et al., 2019). Equidistant sampling is required by many downstream mathematical frameworks, such as cross-correlation or local similarity analysis (Faust et al., 2015), and thus should be strived for, as much as possible. However, the datasets listed in **Table 1**, albeit extensive and resource intensive, are not perfectly equidistant, further highlighting the practical challenges for longitudinal sampling *in situ*, including, but not limited to, accessibility, consistent biomass availability, and cost.

SAMPLE, DATA AND CODE MANAGEMENT

It is crucial to limit potential biases linked to longitudinal data, e.g., in extended time-series; samples are stored for long periods, while multiple personnel may be involved in sample collection, handling, storage, and documentation. Hence, clear guidelines and standardization must be established, as they are key factors that potentially affect downstream processes and overall outcome (Blekhman et al., 2016; Schoenenberger et al., 2016).

Biomolecular extraction from a single sample is ideal over multiple extractions from subsamples (Roume et al., 2013a). Advantageously, commercial kits for concomitant extraction of

multiple biomolecules are available, including reports proposing adapted methods for extracting various biomolecules, such as DNA, total RNA, small RNA, protein, and metabolites (Peña-Llopis and Brugarolas, 2013; Roume et al., 2013b; Thorn et al., 2019). The availability of sufficient biomass (Eisenhofer et al., 2019) lysis-, homogenization- (Machiels et al., 2000; Santiago et al., 2014; Fiedorová et al., 2019) and preservation- (Borén, 2015; Hickl et al., 2019) methods are key factors that determine effectiveness to comprehensively recover all intracellular and/or extracellular biomolecules. Next, biomolecular extraction should be automated, whenever possible. While evaluations have shown that it may not necessarily provide better quality results compared to a human operator (Phillips et al., 2012), the output is more consistent (Fidler et al., 2020). In the same vein, omic readouts should also be generated on a single platform (s) as unique batches to ensure consistent output quality.

Batch effects are often overlooked in omic studies (de Goffau et al., 2021), but can be minimized during stages of sample processing by including randomization, sample tracking, and extensive documentation (Leek et al., 2010). Sample randomization implemented within batches of biomolecular extraction and high-throughput measurements could help discriminate batch effects and temporal variation, i.e., different sets of randomly selected samples from different timepoints could be treated together at each different step (Oh et al., 2019). Additionally, batch effects could be mitigated using downstream analytical (Wang and Cao, 2019) and computational methods (Gibbons et al., 2018; McLaren et al., 2019).

A potential effective experimental measure for minimizing and elucidating batch effects is the inclusion of mock/control samples during both the extraction and high-throughput measurements (Bokulich et al., 2016; Hornung et al., 2019; ATCC Mock Microbial Communities, 2020). Samples with low biomass, e.g., from neonates, glacier-streams, or acid-mine drainage, should include extraction blanks as negative controls, which are extremely valuable to discriminate contaminants arising from kits and reagents (Salter et al., 2014; Heintz-Buschart et al., 2018; Wampach et al., 2018; Weyrich et al., 2019). Furthermore, spike-ins could be helpful for downstream quantification (Zinter et al., 2019). Importantly, replicates can be used within downstream statistical frameworks (Sokal, 1995; Anderson, 2017; Kuznetsova et al., 2017; Mallick et al., 2021) to understand both within- and between-sample heterogeneity, thereby minimizing mischaracterisation of contaminants or findings driven by batch effects (de Goffau et al., 2021).

Longitudinal and multi-omic studies yield large datasets, where data processing and analyses are typically time and resource intensive. These rich datasets may be reused to study multiple aspects of a given microbial system (**Table 1**). Therefore, equal emphasis should be placed on designing bioinformatic workflows and code/data management strategies to improve reproducibility and transparency. For example, peer-review journals have begun mandating “data availability” sections and links to code repositories in adherence to project/coding best practices and standards (Sandve et al., 2013; Bokulich et al., 2020), further improving posterior data integration and analysis in the short-term, while improving scaling-up

and knowledge transfer in the long run (Shahin et al., 2017; Wilson et al., 2017). In addition, format-free archival repositories, such as Zenodo could be used for non-standard data types,¹ for instance simulated raw data, physico-chemical measurements, intermediate data, large tables, and archived Github repositories. Despite this, reports indicate that 26% of bioinformatics tools are no longer available (Mangul et al., 2019), while gaps in available raw data (Jurburg et al., 2020) and metadata (Schriml et al., 2020) still exist.

CONSTRUCTION OF LONGITUDINAL GENE AND GENOME REFERENCE CATALOGUES

Microbiomes may be studied from a gene-centric perspective (Roume et al., 2015), which requires read or contig-level taxonomic classification (Segata et al., 2012; Wood and Salzberg, 2014), ORF prediction (Hyatt et al., 2010; Rho et al., 2010), and gene annotation (Seemann, 2014; Buchfink et al., 2015; Franzosa et al., 2018; Queirós et al., 2020). Metagenome assembled genomes (MAGs) provide genomic context and can be obtained through binning (Chen et al., 2020; Yue et al., 2020) followed by taxonomic classification (Bremges et al., 2020; Chaumeil et al., 2020) and functional annotation. In that regard, several tools exist that improve the binning process by automating the selection of highest-quality MAGs (bins) and/or performing MAG refinement (Broeksema et al., 2017; Sieber et al., 2018; Uritskiy et al., 2018). These tools enable ensemble binning approaches, balancing out the strengths and weaknesses of different binning methods (Chen et al., 2020; Yue et al., 2020).

Features (i.e., taxa or genes) appear in varying quantities, in different timepoints of longitudinal meta-omic studies. It is challenging to link and track features from one timepoint to another without any given point of reference. Therefore, the construction of what we term as “representative longitudinal catalogues” (hereafter referred to as catalogues) of MAGs/genes, provides a non-redundant representative base to link features from the different longitudinal samples (Herold et al., 2020; Martínez Arbas et al., 2021). The outcome of any downstream analysis is highly reliant on the quality of the MAGs and genes within a catalogue, which further depends on the quality of large-scale bioinformatic processing (e.g., *de novo* assembly and binning). **Figure 1** illustrates two methods of constructing such catalogues, which are through aggregated processing of data from all samples or through de-replicating the output from individually processed sample data (i.e., sample-wise processing). A third alternative to these methods could be the representation of non-redundant genes in pangenomes from MAGs annotated at the species-level (Tettelin et al., 2005; Delmont and Eren, 2018), collected across all timepoints. This allows for identifying any varying patterns especially in the context of environmental factors and phylogenetic constraints influencing gene acquisition and/or genome-streamlining (Tettelin et al., 2005). Given that

others have highlighted the catalogue building methodologies (Qin et al., 2010; Nayfach et al., 2020; Almeida et al., 2021); here, we elaborate methods discussed above in the context of both gene- and MAG-centric strategies.

The general advantage of the aggregated processing approach is simplicity, whereby a single run is required for all the large-scale bioinformatic processing steps (**Figure 1**). Moreover, pooled assemblies have been shown to be effective (Magasin and Gerloff, 2015), especially in the advent of highly efficient *de novo* assemblers (Li et al., 2016) and digital normalization (Brown et al., 2012). However, pooling reads from a large number of samples increases the complexity of the *de novo* assembly process, especially for complex communities. It also requires substantial computational resources, while potentially resulting in lower quality contigs, MAGs, and genes (Chen et al., 2020).

The dereplication method (**Figure 1**) is applied after independent sample-wise large-scale bioinformatic processing (Evans and Denef, 2020). Predicted ORFs could be de-replicated through clustering (Li and Godzik, 2006; Edgar, 2010; Mirdita et al., 2019), producing a gene catalogue (Li et al., 2014). On the contrary, the dereplication of MAGs is more complex, requiring several steps: binning from sample-wise *de novo* assemblies to generate MAGs, curation of high-quality MAGs (Parks et al., 2015), and dereplication of MAGs (Olm et al., 2017; Wampach et al., 2018) to select the most representative MAGs of the longitudinal data (Uritskiy et al., 2018; Chen et al., 2020). In general, dereplication methods are particularly advantageous for longitudinal microbiome studies with many deeply sequenced samples (Herold et al., 2020; Martínez Arbas et al., 2021).

Although not systematically evaluated, one caveat worth considering when constructing a catalogue based on *de novo* assemblies, binning, and dereplication is the potential loss of resolution in population-level diversity (Kashtan et al., 2014; Evans and Denef, 2020; Quince et al., 2020), which may include single nucleotide variants, copy number variants, strains, and auxiliary gene content (Evans and Denef, 2020) potentially impacting important downstream steps, such as integration of metaproteomic data (Tanca et al., 2016) or time-resolved strain tracking (Brito and Alm, 2016; Zlitni et al., 2020). To the best of our knowledge, the extent of the impact has yet to be systematically investigated. In our opinion, several strategies can be applied to overcome this issue, including the usage of a comparative genomics methodology, i.e., pangenomes (Delmont and Eren, 2018), even opt for (re) assemblies of read subsets associated to particular taxa or MAGs of interest (Albertsen et al., 2013), or the application of strain-level analysis tools (Anyansi et al., 2020).

Overall, choosing the specific methods for constructing a longitudinal catalogue depends on various factors, including the biological question, complexity of the community (van der Walt et al., 2017), number of samples, and sequencing depth. To the best of our knowledge, a comparison between an aggregated processing approach and a dereplication approach has yet to be conducted. Such a comparison would further help to inform researchers on selecting the best strategy for longitudinal analyses.

¹<https://zenodo.org>

QUANTIFICATION AND NORMALIZATION

Longitudinal catalogues provide compositional information of community taxa and potential functions. However, the relative quantification of community members and functionalities is key in harnessing the power of longitudinal microbiome data, as it allows the observation of community taxa/functional dynamics and could be used in downstream modeling. In that regard, quantifying MG and MT sequencing data is a standard process of aligning reads (Li and Durbin, 2009) to relevant catalogues, and then quantifying features of interest (e.g., population/gene relative genomic abundance, gene expression) based on those alignments, providing information on community structure, functional potential, and gene expression. Complementally, MP data provide functional insights, whereby several methods are available for the quantification of such data (Delogu et al., 2020; Pible et al., 2020), while identification and quantification of metabolites through MM data (Kapoor and Vaidyanathan, 2016; Mallick et al., 2019; Røst et al., 2020) provide insights on the community phenotype (s). However, *in situ* measurements of substrate uptake through labeling-based approaches (Starr et al., 2018) are challenging. Therefore, specific metabolites of interest could be indirectly linked to members of a microbial community by proportionally assigning the relative contribution of a MAG to a given (re) constructed metabolic pathway based on genomic abundance or gene/protein expression (Noecker et al., 2016; Blasche et al., 2021).

Normalization of quantified values is required to enable community structure and function comparisons between timepoint samples. The selection of normalization methods is important as it affects downstream analytical steps. There are several methods to normalize longitudinal MG and MT data, from the generation of compositional data to log-ratios and differential rankings (Chen et al., 2018; Pereira et al., 2018; Morton et al., 2019). Additionally, one should also inspect the data for potential confounding batch effects and take it into consideration when performing normalization (Gibbons et al., 2018; McLaren et al., 2019; Coenen et al., 2020). In summary, effective relative quantification and normalization will serve as a strong basis for downstream modeling approaches, and the development of robust methods for absolute quantification will be decisive in the future.

ANALYSIS OF COMMUNITY CHARACTERISTICS AND DYNAMICS

Generally, microbiome omic data are complex, as it is (i) compositional, e.g., provided as relative abundances, which require specific considerations when selecting statistical analyses (Gloor et al., 2017), (ii) highly sparse, such that the interpretation of zero-values generated from sampling, biological, or technical processes heavily affects data-derived conclusions (Silverman et al., 2020), and (iii) high dimensional, which increases modeling difficulty due to the influence of feature selection that heavily affect potential predictions (Bolón-Canedo et al., 2016). Furthermore, multi-omic studies may contain gaps within the

omic spectrum, such that certain samples may not be represented within a certain omic layer (Lloyd-Price et al., 2019). Despite introducing complexity, the complementary use of different omics could improve analysis outcomes and add predictive power to models (Muller et al., 2013; Fondi and Liò, 2015). Longitudinal data introduce another layer of complexity, i.e., time dependencies, such that one timepoint is dependent on the previous timepoints, rendering conventional statistical analyses unsuitable as they assume samples to be independent (Coenen et al., 2020). This is further compounded by the fact that samples from longitudinal *in situ* studies are often low in number and non-equidistant (Park et al., 2020). Imputation may be used to supplement missing values (i.e., omic measurements or timepoints; Jiang et al., 2020).

Initial exploration of the microbiome dynamics can be assessed through ordination analyses, where high dimensional population structure data are visualized in a two-dimensional space to observe the trajectory of the samples and the behavior of the system, i.e., metastability, cycles, and alternative states (Gonze et al., 2018). Then, community member relationships may be inferred using, e.g., correlation methods (Faust et al., 2012; Friedman and Alm, 2012; Weiss et al., 2016). Unfortunately, correlations may be insufficient to assess complex community interactions, whereby the application of modeling approaches would be necessary to resolve those relationships (Fisher and Mehta, 2014; Trosvik et al., 2015; Ridenhour et al., 2017). Modeling could serve as a means of integrating several layers of omic data (Lloyd-Price et al., 2019; Ruiz-Perez et al., 2021) further elucidating microbial interplay beyond species abundances and functional potential.

Extensive literature of statistical and mathematical frameworks for multi-omic and/or longitudinal microbiome data is currently available. For instance, Noor et al. (2019) review the integration of multi-omics data from data-driven and knowledge-based perspectives. Coenen et al. (2020) discuss approaches to characterize temporal dynamics and to identify periodicity of populations and putative interactions between them, while Faust et al. (2018) propose a classification scheme for better model selection. Bodein et al. (2019) provide a multivariate framework to integrate longitudinal and multi-omics data, while Park et al. (2020) discuss the development of models and software tools for time-series metagenome and metabolome data. Overall, the application of these methodologies should be tailored toward specific hypotheses and studies, for which data exploration is essential to select modeling approaches that fit the type, quality, and quantity of the data.

More recently, the emergence of studies which track microbiome dynamics of cohorts over time, i.e., multiple individuals/sites (Carmody et al., 2019; Lloyd-Price et al., 2019; Mars et al., 2020), necessitates the ability to discriminate variation stemming from the same individual/environment compared to those from different individuals/environments. In such cases, multi-level statistical modeling (also known as mixed-effects/hierarchical models) is able to account for repeated sampling or nested variation across a sample population (Sokal, 1995; Anderson, 2017; Kuznetsova et al., 2017; Mallick et al., 2021). Most notably Lloyd-Price et al. (2019) extensively applied such

methods to associate multi-omic microbiome signatures with host-derived molecular profiles in a cohort of 132 individuals. Other instances include multi-omic longitudinal studies that combine murine and human datasets to unveil the adaptation of gut microbiomes to raw and cooked food (Carmody et al., 2019) and the identification of therapeutic targets for irritable bowel syndrome (Mars et al., 2020). Finally, there are newer methodologies that apply similar/related statistical frameworks to modeling multi-omic data (Mallick et al., 2021).

The validation of the models remains one of the most challenging issues. Mathematical models combined with culture of synthetic microbial communities are commonly utilized to study mechanisms behind host-microbiome interactions (Moejers et al., 2017). It is also possible to validate interactions between microbes by, e.g., applying environmental perturbations in controlled conditions (Law et al., 2016; Herold et al., 2020). These explorations may result in a further understanding of the role of biotic and abiotic factors in shaping microbiomes, in relation to community phenotypes found in nature, biotechnological processes (Law et al., 2016; Herold et al., 2020), or host-associated microbiomes (Moejers et al., 2017; Garza et al., 2018).

CONCLUSION

Longitudinal microbiome studies combined with integrated multi-omic measurements provide unprecedented opportunities to study microbial community dynamics, both structurally and functionally. In tandem with evolving high-throughput technologies, e.g., long-read sequencing (Moss et al., 2020; Wickramarachchi et al., 2020), these studies will become important tools in the exploration and potential exploitation of microbial consortia. We described strategies to mitigate the various challenges associated with such studies, encompassing study design, best practices, practical

considerations, and bioinformatics processing and modeling. While longitudinal multi-omics datasets are currently scarce (Table 1), we are confident that it will increasingly become more common, similar to how we are increasingly transitioning from single omics to multi-omic (Noor et al., 2019). Longitudinal microbiome multi-omics will serve as an important tool for further improving analytical methods, which will in turn lead to relevant biomedical, biotechnological, and environmental outcomes.

AUTHOR CONTRIBUTIONS

SMA and SN outlined the manuscript and coordinated the writing process. LdN, SN, and SMA prepared the figure. All authors contributed to the writing, reviewing, and editing of the manuscript. All authors approved the submitted version.

FUNDING

The Luxembourg National Research Fund (FNR) supported SMA, PQ, LdN, PM, and EELM through the PRIDE doctoral training unit grants (PRIDE15/10907093) and (PRIDE/18/11823097), the CORE Junior grant (C15/SR/10404839), and the CORE grant (CORE/17/SM/11689322). SBB was supported by the Sinergia grant (CRSII5_180241) through the Swiss National Science Foundation. PW was supported by the European Research Council (ERC-CoG 863664).

ACKNOWLEDGMENTS

We would like to thank Oskar Hickl for his input on metaproteomic analysis.

REFERENCES

- Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., and Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* 31, 533–538. doi: 10.1038/nbt.2579
- Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., et al. (2021). A unified catalog of 204, 938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* 39, 105–114. doi: 10.1038/s41587-020-0603-3
- Altman, D. G., and Bland, J. M. (1995). Statistics notes: absence of evidence is not evidence of absence. *BMJ* 311:485. doi: 10.1136/bmj.311.7003.485
- Anderson, M. J. (2017). "Permutational multivariate analysis of variance (PERMANOVA)," in *Wiley Stats Ref: Statistics Reference Online*. eds. N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J. L. Teugels (Chichester, UK: John Wiley & Sons, Ltd.), 1–15.
- Anyansi, C., Straub, T. J., Manson, A. L., Earl, A. M., and Abeel, T. (2020). Computational methods for strain-level microbial detection in colony and metagenome sequencing data. *Front. Microbiol.* 11:1925. doi: 10.3389/fmicb.2020.01925
- ATCC Mock Microbial Communities (2020). Available at: https://www.atcc.org/en/Products/Microbiome_Standards.aspx (Accessed November 30, 2020).
- Blasche, S., Kim, Y., Mars, R. A. T., Machado, D., Maansson, M., Kafkia, E., et al. (2021). Metabolic cooperation and spatiotemporal niche partitioning in a kefir microbial community. *Nat. Microbiol.* 6, 196–208. doi: 10.1038/s41564-020-00816-5
- Blekhman, R., Tang, K., Archie, E. A., Barreiro, L. B., Johnson, Z. P., Wilson, M. E., et al. (2016). Common methods for fecal sample storage in field studies yield consistent signatures of individual identity in microbiome sequencing data. *Sci. Rep.* 6:31519. doi: 10.1038/srep31519
- Bodein, A., Chapleur, O., Droit, A., and Lê Cao, K.-A. (2019). A generic multivariate framework for the integration of microbiome longitudinal studies with other data types. *Front. Genet.* 10:963. doi: 10.3389/fgene.2019.00963
- Bodelier, P. L. E. (2011). Toward understanding, managing, and protecting microbial ecosystems. *Front. Microbiol.* 2:80. doi: 10.3389/fmicb.2011.00080
- Bokulich, N. A., Rideout, J. R., Mercurio, W. G., Shiffer, A., Wolfe, B., Maurice, C. F., et al. (2016). Mockrobiota: a public resource for microbiome bioinformatics benchmarking. *mSystems* 1:e00062-16. doi: 10.1128/mSystems.00062-16
- Bokulich, N. A., Ziemski, M., Robeson, M. S., and Kaehler, B. D. (2020). Measuring the microbiome: best practices for developing and benchmarking microbiomics methods. *Comput. Struct. Biotechnol. J.* 18, 4048–4062. doi: 10.1016/j.csbj.2020.11.049
- Bolón-Canedo, V., Sánchez-Marño, N., and Alonso-Betanzos, A. (2016). Feature selection for high-dimensional data. *Prog. Artif. Intell.* 5, 65–75. doi: 10.1007/s13748-015-0080-y
- Borén, M. (2015). "Sample preservation Through heat stabilization of proteins: principles and examples," in *Proteomic Profiling Methods in Molecular Biology*. ed. A. Posch (New York, NY: Springer), 21–32.

- Bremges, A., Fritz, A., and McHardy, A. C. (2020). CAMITAX: taxon labels for microbial genomes. *Giga Science* 9:giz154. doi: 10.1093/gigascience/giz154
- Brito, I. L., and Alm, E. J. (2016). Tracking strains in the microbiome: insights from metagenomics and models. *Front. Microbiol.* 7:712. doi: 10.3389/fmicb.2016.00712
- Broeksema, B., Calusinska, M., McGee, F., Winter, K., Bongiovanni, F., Goux, X., et al. (2017). ICoVeR – an interactive visualization tool for verification and refinement of metagenomic bins. *BMC Bioinformatics* 18:233. doi: 10.1186/s12859-017-1653-5
- Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B., and Brom, T. H. (2012). A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. arXiv [Preprint].
- Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi: 10.1038/nmeth.3176
- Cao, H.-T., Gibson, T. E., Bashan, A., and Liu, Y.-Y. (2017). Inferring human microbial dynamics from temporal metagenomics data: pitfalls and lessons. *BioEssays* 39:1600188. doi: 10.1002/bies.201600188
- Carmody, R. N., Bisanz, J. E., Bowen, B. P., Maurice, C. F., Lyalina, S., Louie, K. B., et al. (2019). Cooking shapes the structure and function of the gut microbiome. *Nat. Microbiol.* 4, 2052–2063. doi: 10.1038/s41564-019-0569-4
- Celaj, A., Markle, J., Danska, J., and Parkinson, J. (2014). Comparison of assembly algorithms for improving rate of metatranscriptomic functional annotation. *Microbiome* 2:39. doi: 10.1186/2049-2618-2-39
- Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2020). GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* 36, 1925–1927. doi: 10.1093/bioinformatics/btz848
- Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M., and Banfield, J. F. (2020). Accurate and complete genomes from metagenomes. *Genome Res.* 30, 315–333. doi: 10.1101/gr.258640.119
- Chen, L., Reeve, J., Zhang, L., Huang, S., Wang, X., and Chen, J. (2018). GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ* 6:e4600. doi: 10.7717/peerj.4600
- Coenen, A. R., Hu, S. K., Luo, E., Muratore, D., and Weitz, J. S. (2020). A primer for microbiome time-series analysis. *Front. Genet.* 11:310. doi: 10.3389/fgene.2020.00310
- Costa Junior, C., Corbeels, M., Bernoux, M., Piccolo, M. C., Siqueira Neto, M., Feigl, B. J., et al. (2013). Assessing soil carbon storage rates under no-tillage: comparing the synchronic and diachronic approaches. *Soil Tillage Res.* 134, 207–212. doi: 10.1016/j.still.2013.08.010
- David, L. A., Materna, A. C., Friedman, J., Campos-Baptista, M. I., Blackburn, M. C., Perrotta, A., et al. (2014). Host lifestyle affects human microbiota on daily timescales. *Genome Biol.* 15:R89. doi: 10.1186/gb-2014-15-7-r89
- de Goffau, M. C., Charnock-Jones, D. S., Smith, G. C. S., and Parkhill, J. (2021). Batch effects account for the main findings of an in utero human intestinal bacterial colonization study. *Microbiome* 9:6. doi: 10.1186/s40168-020-00949-z
- Delmont, T. O., and Eren, A. M. (2018). Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. *PeerJ* 6:e4320. doi: 10.7717/peerj.4320
- Delogu, F., Kunath, B. J., Evans, P. N., Arntzen, M. Ø., Hvidsten, T. R., and Pope, P. B. (2020). Integration of absolute multi-omics reveals dynamic protein-to-RNA ratios and metabolic interplay within mixed-domain microbiomes. *Nat. Commun.* 11:4708. doi: 10.1038/s41467-020-18543-0
- Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461
- Eisenhofer, R., Minich, J. J., Marotz, C., Cooper, A., Knight, R., and Weyrich, L. S. (2019). Contamination in low microbial biomass microbiome studies: issues and recommendations. *Trends Microbiol.* 2, 105–117. doi: 10.1016/j.tim.2018.11.003
- Erickson, A. R., Cantarel, B. L., Lamendella, R., Darzi, Y., Mongodin, E. F., Pan, C., et al. (2012). Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. *PLoS One* 7:e49138. doi: 10.1371/journal.pone.0049138
- Evans, J. T., and Deneff, V. J. (2020). To dereplicate or not to dereplicate? *mSphere* 5:e00971-19. doi: 10.1128/mSphere.00971-19
- Faust, K., Bauchinger, F., Laroche, B., de Buyl, S., Lahti, L., Washburne, A. D., et al. (2018). Signatures of ecological processes in microbial community time series. *Microbiome* 6:120. doi: 10.1186/s40168-018-0496-2
- Faust, K., Lahti, L., Gonze, D., de Vos, W. M., and Raes, J. (2015). Metagenomics meets time series analysis: unraveling microbial community dynamics. *Curr. Opin. Microbiol.* 25, 56–66. doi: 10.1016/j.mib.2015.04.004
- Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., et al. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.* 8:e1002606. doi: 10.1371/journal.pcbi.1002606
- Fidler, G., Tolnai, E., Stägel, A., Remenyik, J., Stundl, L., Gal, F., et al. (2020). Tendentious effects of automated and manual metagenomic DNA purification protocols on broiler gut microbiome taxonomic profiling. *Sci. Rep.* 10:3419. doi: 10.1038/s41598-020-60304-y
- Fiedorová, K., Radvanský, M., Němcová, E., Grombířková, H., Bosák, J., Černochová, M., et al. (2019). The impact of DNA extraction methods on stool bacterial and fungal microbiota community recovery. *Front. Microbiol.* 10:821. doi: 10.3389/fmicb.2019.00821
- Fisher, C. K., and Mehta, P. (2014). Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *PLoS One* 9:e102451. doi: 10.1371/journal.pone.0102451
- Fondi, M., and Liò, P. (2015). Multi-omics and metabolic modelling pipelines: challenges and tools for systems microbiology. *Microbiol. Res.* 171, 52–64. doi: 10.1016/j.micres.2015.01.003
- Franzosa, E. A., McIver, L. J., Rahnnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., et al. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* 15, 962–968. doi: 10.1038/s41592-018-0176-y
- Friedman, J., and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* 8:e1002687. doi: 10.1371/journal.pcbi.1002687
- Garza, D. R., van Verk, M. C., Huynen, M. A., and Dutilh, B. E. (2018). Towards predicting the environmental metabolome from metagenomics with a mechanistic model. *Nat. Microbiol.* 3, 456–460. doi: 10.1038/s41564-018-0124-8
- Gerber, G. K. (2014). The dynamic microbiome. *FEBS Lett.* 588, 4131–4139. doi: 10.1016/j.febslet.2014.02.037
- Gibbons, S. M., Duvallet, C., and Alm, E. J. (2018). Correcting for batch effects in case-control microbiome studies. *PLoS Comput. Biol.* 14:e1006102. doi: 10.1371/journal.pcbi.1006102
- Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8:2224. doi: 10.3389/fmicb.2017.02224
- Gonze, D., Coyte, K. Z., Lahti, L., and Faust, K. (2018). Microbial communities as dynamical systems. *Curr. Opin. Microbiol.* 44, 41–49. doi: 10.1016/j.mib.2018.07.004
- Heintz-Buschart, A., May, P., Laczny, C. C., Lebrun, L. A., Bellora, C., Krishna, A., et al. (2016). Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat. Microbiol.* 2:16180. doi: 10.1038/nmicrobiol.2016.227
- Heintz-Buschart, A., and Wilmes, P. (2018). Human gut microbiome: function matters. *Trends Microbiol.* 26, 563–574. doi: 10.1016/j.tim.2017.11.002
- Heintz-Buschart, A., Yusuf, D., Kaysen, A., Etheridge, A., Fritz, J. V., May, P., et al. (2018). Small RNA profiling of low biomass samples: identification and removal of contaminants. *BMC Biol.* 16:52. doi: 10.1186/s12915-018-0522-7
- Herold, M., Arbas, S. M., Narayanasamy, S., Sheik, A. R., Kleine-Borgmann, L. A. K., Lebrun, L. A., et al. (2020). Integration of time-series meta-omics data reveals how microbial ecosystems respond to disturbance. *Nat. Commun.* 11:5281. doi: 10.1038/s41467-020-19006-2
- Heyer, R., Schallert, K., Zoun, R., Becher, B., Saake, G., and Benndorf, D. (2017). Challenges and perspectives of metaproteomic data analysis. *J. Biotechnol.* 261, 24–36. doi: 10.1016/j.jbiotec.2017.06.1201
- Hickl, O., Heintz-Buschart, A., Trautwein-Schult, A., Hercog, R., Bork, P., Wilmes, P., et al. (2019). Sample preservation and storage significantly impact taxonomic and functional profiles in metaproteomics studies of the human gut microbiome. *Microorganisms* 7:367. doi: 10.3390/microorganisms7090367
- Hornung, B. V. H., Zwiittink, R. D., and Kuijper, E. J. (2019). Issues and current standards of controls in microbiome research. *FEMS Microbiol. Ecol.* 95:fiz045. doi: 10.1093/femsec/fiz045
- Hyatt, D., Chen, G.-L., LoCascio, P. E., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119. doi: 10.1186/1471-2105-11-119
- Jiang, D., Armour, C. R., Hu, C., Mei, M., Tian, C., Sharpton, T. J., et al. (2019). Microbiome multi-omics network analysis: statistical considerations,

- limitations, and opportunities. *Front. Genet.* 10:995. doi: 10.3389/fgene.2019.00995
- Jiang, R., Li, W. V., and Li, J. J. (2020). mbImpute: an accurate and robust imputation method for microbiome data. *Genomics* [Preprint]. doi: 10.1101/2020.03.07.982314
- Johnston, J., LaPara, T., and Behrens, S. (2019). Composition and dynamics of the activated sludge microbiome during seasonal nitrification failure. *Sci. Rep.* 9:4565. doi: 10.1038/s41598-019-40872-4
- Jurburg, S. D., Konzack, M., Eisenhauer, N., and Heintz-Buschart, A. (2020). The archives are half-empty: an assessment of the availability of microbial community sequencing data. *Commun. Biol.* 3:474. doi: 10.1038/s42003-020-01204-9
- Kapoor, R. V., and Vaidyanathan, S. (2016). Towards quantitative mass spectrometry-based metabolomics in microbial and mammalian systems. *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.* 374:20150363. doi: 10.1098/rsta.2015.0363
- Kashtan, N., Roggensack, S. E., Rodrigue, S., Thompson, J. W., Biller, S. J., Coe, A., et al. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* 344, 416–420. doi: 10.1126/science.1248575
- Kaysen, A., Heintz-Buschart, A., Muller, E. E. L., Narayanasamy, S., Wampach, L., Laczny, C. C., et al. (2017). Integrated meta-omic analyses of the gastrointestinal tract microbiome in patients undergoing allogeneic hematopoietic stem cell transplantation. *Transl. Res.* 186, 79–94. doi: 10.1016/j.trsl.2017.06.008
- Kumar, R., Eipers, P., Little, R. B., Crowley, M., Crossman, D. K., Lefkowitz, E. J., et al. (2014). Getting started with microbiome analysis: sample acquisition to bioinformatics. *Curr. Protoc. Hum. Genet.* 82, 18.8.1–18.8.29. doi: 10.1002/0471142905.hg180882
- Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13
- Lapidus, A. L., and Korobeynikov, A. I. (2021). Metagenomic data assembly – the way of decoding unknown microorganisms. *Front. Microbiol.* 12:613791. doi: 10.3389/fmicb.2021.613791
- Law, Y., Kirkegaard, R. H., Cokro, A. A., Liu, X., Arumugam, K., Xie, C., et al. (2016). Integrative microbial community analysis reveals full-scale enhanced biological phosphorus removal under tropical conditions. *Sci. Rep.* 6:25719. doi: 10.1038/srep25719
- Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11, 733–739. doi: 10.1038/nrg2825
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324
- Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158
- Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., et al. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* 32, 834–841. doi: 10.1038/nbt.2942
- Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., et al. (2016). MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* 102, 3–11. doi: 10.1016/j.jmeth.2016.02.020
- Liang, Y., Dong, T., Chen, M., He, L., Wang, T., Liu, X., et al. (2020). Systematic analysis of impact of sampling regions and storage methods on fecal gut microbiome and metabolome profiles. *mSphere* 5:e00763-19. doi: 10.1128/mSphere.00763-19
- Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., et al. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 569, 655–662. doi: 10.1038/s41586-019-1237-9
- Machiels, B. M., Ruers, T., Lindhout, M., Hardy, K., Hlavaty, T., Bang, D. D., et al. (2000). New protocol for DNA extraction of stool. *Bio Techniques* 28, 286–290. doi: 10.2144/00282st05
- Magasin, J. D., and Gerloff, D. L. (2015). Pooled assembly of marine metagenomic datasets: enriching annotation through chimerism. *Bioinformatics* 31, 311–317. doi: 10.1093/bioinformatics/btu546
- Mallick, H., Franzosa, E. A., McIver, L. J., Banerjee, S., Sirota-Madi, A., Kostic, A. D., et al. (2019). Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. *Nat. Commun.* 10:3136. doi: 10.1038/s41467-019-10927-1
- Mallick, H., Rahnavard, A., McIver, L. J., Ma, S., Zhang, Y., Nguyen, L. H., et al. (2021). Multivariable association discovery in population-scale metagenomics studies. *Microbiology* [Preprint]. doi: 10.1099/mic.0.001031
- Mangul, S., Martin, L. S., Eskin, E., and Blekhan, R. (2019). Improving the usability and archival stability of bioinformatics software. *Genome Biol.* 20:47. doi: 10.1186/s13059-019-1649-8
- Mars, R. A. T., Yang, Y., Ward, T., Houlti, M., Priya, S., Lekatz, H. R., et al. (2020). Longitudinal multi-omics reveals subset-specific mechanisms underlying irritable bowel syndrome. *Cell* 182, 1460–1473. doi: 10.1016/j.cell.2020.08.007
- Martínez Arbas, S. M., Narayanasamy, S., Herold, M., Lebrun, L. A., Hoopmann, M. R., Li, S., et al. (2021). Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics revealed using time-series integrated meta-omics. *Nat. Microbiol.* 6, 123–135. doi: 10.1038/s41564-020-00794-8
- McLaren, M. R., Willis, A. D., and Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing experiments. *elife* 8:e46923. doi: 10.7554/eLife.46923
- Mirdita, M., Steinegger, M., and Söding, J. (2019). MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics* 35, 2856–2858. doi: 10.1093/bioinformatics/bty1057
- Moeves, F., Succurro, A., Popa, O., Maguire, J., and Ebenhöf, O. (2017). Dynamics of the bacterial community associated with *Phaeodactylum tricornutum* cultures. *Processes* 5:77. doi: 10.3390/pr5040077
- Morton, J. T., Marotz, C., Washburne, A., Silverman, J., Zaramela, L. S., Edlund, A., et al. (2019). Establishing microbial composition measurement standards with reference frames. *Nat. Commun.* 10:2719. doi: 10.1038/s41467-019-10656-5
- Moss, E. L., Maghini, D. G., and Bhatt, A. S. (2020). Complete, closed bacterial genomes from microbiomes using nanopore sequencing. *Nat. Biotechnol.* 38, 701–707. doi: 10.1038/s41587-020-0422-6
- Muller, E. E. L. (2019). Determining microbial niche breadth in the environment for better ecosystem fate predictions. *mSystems* 4:e00080-19. doi: 10.1128/mSystems.00080-19
- Muller, E. E. L., Faust, K., Widder, S., Herold, M., Arbas, S. M., and Wilmes, P. (2018). Using metabolic networks to resolve ecological properties of microbiomes. *Curr. Opin. Syst. Biol.* 8, 73–80. doi: 10.1016/j.coisb.2017.12.004
- Muller, E. E. L., Glaab, E., May, P., Vlassis, N., and Wilmes, P. (2013). Condensing the omics fog of microbial communities. *Trends Microbiol.* 21, 325–333. doi: 10.1016/j.tim.2013.04.009
- Muller, E. E. L., Pinel, N., Laczny, C. C., Hoopmann, M. R., Narayanasamy, S., Lebrun, L. A., et al. (2014). Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. *Nat. Commun.* 5:5603. doi: 10.1038/ncomms6603
- Narayanasamy, S., Jarosz, Y., Muller, E. E. L., Heintz-Buschart, A., Herold, M., Kaysen, A., et al. (2016). IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.* 17:260. doi: 10.1186/s13059-016-1116-8
- Narayanasamy, S., Muller, E. E. L., Sheik, A. R., and Wilmes, P. (2015). Integrated omics for the identification of key functionalities in biological wastewater treatment microbial communities. *Microb. Biotechnol.* 8, 363–368. doi: 10.1111/1751-7915.12255
- Nayfach, S., Roux, S., Seshadri, R., Udvar, D., Varghese, N., Schulz, F., et al. (2020). A genomic catalog of earth's microbiomes. *Nat. Biotechnol.* 39, 499–509. doi: 10.1038/s41587-020-0718-6
- Noecker, C., Eng, A., Srinivasan, S., Theriot, C. M., Young, V. B., Jansson, J. K., et al. (2016). Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. *mSystems* 1:e00013-15. doi: 10.1128/mSystems.00013-15
- Noor, E., Cherkaoui, S., and Sauer, U. (2019). Biological insights through omics data integration. *Gene Regul.* 15, 39–47. doi: 10.1016/j.coisb.2019.03.007
- Oh, S., Li, C., Baldwin, R. L., Song, S., Liu, F., and Li, R. W. (2019). Temporal dynamics in meta longitudinal RNA-Seq data. *Sci. Rep.* 9:763. doi: 10.1038/s41598-018-37397-7
- Olm, M. R., Brown, C. T., Brooks, B., and Banfield, J. F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* 11, 2864–2868. doi: 10.1038/ismej.2017.126

- Park, S.-Y., Ufodu, A., Lee, K., and Jayaraman, A. (2020). Emerging computational tools and models for studying gut microbiota composition and function. *Tissue Cell Pathw. Eng.* 66, 301–311. doi: 10.1016/j.copbio.2020.10.005
- Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). Check M: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. doi: 10.1101/gr.186072.114
- Peña-Llopis, S., and Brugarolas, J. (2013). Simultaneous isolation of high-quality DNA, RNA, miRNA and proteins from tissues for genomic applications. *Nat. Protoc.* 8, 2240–2255. doi: 10.1038/nprot.2013.141
- Pereira, M. B., Wallroth, M., Jonsson, V., and Kristiansson, E. (2018). Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics* 19:274. doi: 10.1186/s12864-018-4637-6
- Phillips, K., McCallum, N., and Welch, L. (2012). A comparison of methods for forensic DNA extraction: Chelex-100® and the QIAGEN DNA Investigator Kit (manual and automated). *Forensic Sci. Int. Genet.* 6, 282–285. doi: 10.1016/j.fsigen.2011.04.018
- Pible, O., Allain, F., Jouffret, V., Culotta, K., Miotello, G., and Armengaud, J. (2020). Estimating relative biomasses of organisms in microbiota using “phyloepitomics”. *Microbiome* 8:30. doi: 10.1186/s40168-020-00797-x
- Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821
- Queirós, P., Delogu, F., Hickl, O., May, P., and Wilmes, P. (2020). Mantis: flexible and consensus-driven genome annotation. *Bioinformatics* [Preprint]. doi: 10.1101/2020.11.02.360933
- Quince, C., Nurk, S., Raguideau, S., James, R., Soyer, O. S., Summers, J. K., et al. (2020). Metagenomics strain resolution on assembly graphs. *Bioinformatics* [Preprint]. doi: 10.1101/2020.09.06.284828
- Rho, M., Tang, H., and Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* 38:e191. doi: 10.1093/nar/gkq747
- Ridenhour, B. J., Brooker, S. L., Williams, J. E., Van Leuven, J. T., Miller, A. W., Dearing, M. D., et al. (2017). Modeling time-series data from microbial communities. *ISME J.* 11, 2526–2537. doi: 10.1038/ismej.2017.107
- Røst, L. M., Brekke Thorfinnsdottir, L., Kumar, K., Fuchino, K., Eide Langørgen, I., Bartosova, Z., et al. (2020). Absolute quantification of the central carbon metabolome in eight commonly applied prokaryotic and eukaryotic model systems. *Metabolites* 10:74. doi: 10.3390/metabo10020074
- Roume, H., Heintz-Buschart, A., Müller, E. E. L., May, P., Satagopam, V. P., Laczny, C. C., et al. (2015). Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. *Npj Biofilms Microbiomes* 1:15007. doi: 10.1038/npjbiofilms.2015.7
- Roume, H., Heintz-Buschart, A., Müller, E. E. L., and Wilmes, P. (2013b). “Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample,” in *Methods in Enzymology*. ed. E. F. DeLong (Cambridge, Massachusetts, United States: Elsevier), 219–236.
- Roume, H., Müller, E. E., Cordes, T., Renaut, J., Hiller, K., and Wilmes, P. (2013a). A biomolecular isolation framework for eco-systems biology. *ISME J.* 7, 110–121. doi: 10.1038/ismej.2012.72
- Ruiz-Perez, D., Lugo-Martínez, J., Bourguignon, N., Mathee, K., Lerner, B., Bar-Joseph, Z., et al. (2021). Dynamic Bayesian networks for integrating multi-omics time series microbiome data. *mSystems* 6:e01105-20. doi: 10.1128/mSystems.01105-20
- Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., et al. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 12:87. doi: 10.1186/s12915-014-0087-z
- Sandve, G. K., Nekrutenko, A., Taylor, J., and Hovig, E. (2013). Ten simple rules for reproducible computational research. *PLoS Comput. Biol.* 9:e1003285. doi: 10.1371/journal.pcbi.1003285
- Santiago, A., Panda, S., Mengels, G., Martínez, X., Azpiroz, F., Dore, J., et al. (2014). Processing faecal samples: a step forward for standards in microbial community analysis. *BMC Microbiol.* 14:112. doi: 10.1186/1471-2180-14-112
- Schoenenberger, A. W., Muggli, F., Parati, G., Gallino, A., Ehret, G., Suter, P. M., et al. (2016). Protocol of the Swiss Longitudinal Cohort Study (SWICOS) in rural Switzerland. *BMJ Open* 6:e013280. doi: 10.1136/bmjopen-2016-013280
- Schriml, L. M., Chuvochina, M., Davies, N., Elie-Fadrosh, E. A., Finn, R. D., Hugenholtz, P., et al. (2020). COVID-19 pandemic reveals the peril of ignoring metadata standards. *Sci. Data* 7:188. doi: 10.1038/s41597-020-0524-5
- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi: 10.1093/bioinformatics/btu153
- Sefer, E., Kleyman, M., and Bar-Joseph, Z. (2016). Tradeoffs between dense and replicate sampling strategies for high-throughput time series experiments. *Cell Syst.* 3, 35–42. doi: 10.1016/j.cels.2016.06.007
- Segata, N., Waldron, L., Ballarín, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* 9, 811–814. doi: 10.1038/nmeth.2066
- Shahin, M., Ali Babar, M., and Zhu, L. (2017). Continuous integration, delivery and deployment: a systematic review on approaches, tools, challenges and practices. *IEEE Access* 5, 3909–3943. doi: 10.1109/ACCESS.2017.2685629
- Siebert, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., et al. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* 3, 836–843. doi: 10.1038/s41564-018-0171-1
- Silverman, J. D., Roche, K., Mukherjee, S., and David, L. A. (2020). Naught all zeros in sequence count data are the same. *Comput. Struct. Biotechnol. J.* 18, 2789–2798. doi: 10.1016/j.csbj.2020.09.014
- Sokal, R. R. (1995). *Biometry: The Principles and Practice of Statistics in Biological Research*. 3rd Edn. New York: W.H. Freeman.
- Starr, E. P., Shi, S., Blazewicz, S. J., Probst, A. J., Herman, D. J., Firestone, M. K., et al. (2018). Stable isotope informed genome-resolved metagenomics reveals that Saccharibacteria utilize microbially-processed plant-derived carbon. *Microbiome* 6:122. doi: 10.1186/s40168-018-0499-z
- Stewart, C. J., Ajami, N. J., O’Brien, J. L., Hutchinson, D. S., Smith, D. P., Wong, M. C., et al. (2018). Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* 562, 583–588. doi: 10.1038/s41586-018-0617-x
- Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. A., Kultima, J. R., et al. (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* 10, 1196–1199. doi: 10.1038/nmeth.2693
- Tanca, A., Abbondio, M., Palomba, A., Fraumene, C., Manghina, V., Cucca, F., et al. (2017). Potential and active functions in the gut microbiota of a healthy human cohort. *Microbiome* 5:79. doi: 10.1186/s40168-017-0293-3
- Tanca, A., Palomba, A., Fraumene, C., Pagnozzi, D., Manghina, V., Deligios, M., et al. (2016). The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome* 4:51. doi: 10.1186/s40168-016-0196-8
- Tettelin, H., Masiagnani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc. Natl. Acad. Sci. U. S. A.* 102, 13950–13955. doi: 10.1073/pnas.0506758102
- Thorn, C. E., Bergesch, C., Joyce, A., Sambrano, G., McDonnell, K., Brennan, F., et al. (2019). A robust, cost-effective method for DNA, RNA and protein co-extraction from soil, other complex microbiomes and pure cultures. *Mol. Ecol. Resour.* 19, 439–455. doi: 10.1111/1755-0998.12979
- Timmins-Schiffman, E., May, D. H., Mikan, M., Riffle, M., Frazar, C., Harvey, H. R., et al. (2017). Critical decisions in metaproteomics: achieving high confidence protein annotations in a sea of unknowns. *ISME J.* 11, 309–314. doi: 10.1038/ismej.2016.132
- Treangen, T. J., Koren, S., Sommer, D. D., Liu, B., Astrovska, I., Ondov, B., et al. (2013). MetaAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol.* 14:R2. doi: 10.1186/gb-2013-14-1-r2
- Trosvik, P., de Muinck, E. J., and Stenseth, N. C. (2015). Biotic interactions and temporal dynamics of the human gastrointestinal microbiota. *ISME J.* 9, 533–541. doi: 10.1038/ismej.2014.147
- Uritskiy, G. V., DiRuggiero, J., and Taylor, J. (2018). MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 6:158. doi: 10.1186/s40168-018-0541-1
- van der Walt, A. J., van Goethem, M. W., Ramond, J.-B., Makhallanyane, T. P., Reva, O., and Cowan, D. A. (2017). Assembling metagenomes, one community at a time. *BMC Genomics* 18:521. doi: 10.1186/s12864-017-3918-9
- Wampach, L., Heintz-Buschart, A., Fritz, J. V., Ramiro-García, J., Habier, J., Herold, M., et al. (2018). Birth mode is associated with earliest strain-conferred gut microbiome functions and immunostimulatory potential. *Nat. Commun.* 9:5091. doi: 10.1038/s41467-018-07631-x
- Wang, Y., and Cao, K.-A. L. (2019). Managing batch effects in microbiome data. *Brief. Bioinform.* 21, 1954–1970. doi: 10.1093/bib/bbz105
- Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., et al. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* 10, 1669–1681. doi: 10.1038/ismej.2015.235

- Weyrich, L. S., Farrer, A. G., Eisenhofer, R., Arriola, L. A., Young, J., Selway, C. A., et al. (2019). Laboratory contamination over time during low-biomass sample analysis. *Mol. Ecol. Resour.* 19, 982–996. doi: 10.1111/1755-0998.13011
- Wickramarachchi, A., Mallawaarachchi, V., Rajan, V., and Lin, Y. (2020). MetaBCC-LR: metagenomics binning by coverage and composition for long reads. *Bioinformatics* 36, i3–i11. doi: 10.1093/bioinformatics/btaa441
- Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., and Teal, T. K. (2017). Good enough practices in scientific computing. *PLoS Comput. Biol.* 13:e1005510. doi: 10.1371/journal.pcbi.1005510
- Wood, D. E., and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15:R46. doi: 10.1186/gb-2014-15-3-r46
- Yu, K., Yi, S., Li, B., Guo, F., Peng, X., Wang, Z., et al. (2019). An integrated meta-omics approach reveals substrates involved in synergistic interactions in a bisphenol A (BPA)-degrading microbial community. *Microbiome* 7:16. doi: 10.1186/s40168-019-0634-5
- Yue, Y., Huang, H., Qi, Z., Dou, H.-M., Liu, X.-Y., Han, T.-F., et al. (2020). Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. *BMC Bioinformatics* 21:334. doi: 10.1186/s12859-020-03667-3
- Zhou, Z., Tran, P. Q., Breiser, A. M., Liu, Y., Kieft, K., Cowley, E. S., et al. (2020). METABOLIC: high-throughput profiling of microbial genomes for functional traits, biogeochemistry, and community-scale metabolic networks. *bioRxiv* [Preprint]. doi: 10.1101/2020.10.27.357558
- Zimmermann, J., Kaleta, C., and Waschina, S. (2021). gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome Biol.* 22:81. doi: 10.1186/s13059-021-02295-1
- Zinter, M. S., Mayday, M. Y., Ryckman, K. K., Jelliffe-Pawlowski, L. L., and DeRisi, J. L. (2019). Towards precision quantification of contamination in metagenomic sequencing experiments. *Microbiome* 7, 62. doi: 10.1186/s40168-019-0678-6
- Zlitni, S., Bishara, A., Moss, E. L., Tkachenko, E., Kang, J. B., Culver, R. N., et al. (2020). Strain-resolved microbiome sequencing reveals mobile elements that drive bacterial competition on a clinical timescale. *Genome Med.* 12:50. doi: 10.1186/s13073-020-00747-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Martínez Arbas, Busi, Queirós, de Nies, Herold, May, Wilmes, Muller and Narayanasamy. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Appendix A.8

Genomic and metabolic adaptations of biofilms to ecological windows of opportunities in glacier-fed streams

Genomic and metabolic adaptations of biofilms to ecological windows of opportunities in glacier-fed streams

Susheel Bhanu Busi^{1,#}, Massimo Bourquin^{2,#}, Stilianos Fodelianakis^{2,#}, Grégoire Michoud², Tyler J. Kohler², Hannes Peter², Paraskevi Pramateftaki², Michail Styllas², Matteo Tolosano², Vincent De Staercke², Martina Schön², Laura de Nies¹, Ramona Marasco³, Daniele Daffonchio³, Leïla Ezzat², Paul Wilmes^{1,*}, & Tom J. Battin^{2,*}

¹Systems Ecology Group, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg

²Stream Biofilm & Ecosystem Research Lab, ENAC, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

³Biological and Environmental Sciences and Engineering Division (BESE), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

[#]Contributed equally to this work

*Corresponding author(s): Tom Battin (tom.battin@epfl.ch), Paul Wilmes (paul.wilmes@uni.lu)

Running title

Genomic insights into glacier-fed stream biofilms

Keywords

glacier-fed streams, biofilms, metagenomics, cross-domain interactions, adaptive traits

Abstract

Microorganisms dominate life in cryospheric ecosystems. In glacier-fed streams (GFSs), ecological windows of opportunities allow complex microbial biofilms to develop and transiently form the basis of the food web, thereby controlling key ecosystem processes. Here, using high-resolution metagenomics, we unravel strategies that allow biofilms to seize this opportunity in an ecosystem otherwise characterized by harsh environmental conditions. We found a diverse

microbiome spanning the entire tree of life and including a rich virome. Various and co-existing energy acquisition pathways point to diverse niches and the simultaneous exploitation of available resources, likely fostering the establishment of complex biofilms in GFSs during windows of opportunity. The wide occurrence of rhodopsins across metagenome-assembled genomes (MAGs), besides chlorophyll, highlights the role of solar energy capture in these biofilms. Concomitantly, internal carbon and nutrient cycling between photoautotrophs and heterotrophs may help overcome constraints imposed by the high oligotrophy in GFSs. MAGs also revealed mechanisms potentially protecting bacteria against low temperatures and high UV-radiation. The selective pressure of the GFS environment is further highlighted by the phylogenomic analysis, differentiating the representatives of the genus *Polaromonas*, an important component of the GFS microbiome, from those found in other ecosystems. Our findings reveal key genomic underpinnings of adaptive traits that contribute to the success of complex biofilms to exploit environmental opportunities in GFSs, now rapidly changing owing to global warming.

Introduction

Ecosystems and their constituent biota are finely tuned to the seasonal variations of their environment. This phenology is particularly pronounced in glacier-fed streams (hereafter GFSs), which are commonly enveloped by snow cover and darkness in winter, and subject to high flow and sediment mobilization in summer. Yet, ecological ‘windows of opportunity’ arise in spring and autumn^{1,2}, and are characterized by elevated light and nutrient (N, P) availability along with moderate flow, allowing algae and cyanobacteria to rapidly develop ‘green oases’ of phototrophic biofilms. Partially due to the absence of terrestrial organic matter subsidies from the catchment, this punctuated exploitation of solar energy in an otherwise energy-limited ecosystem, transiently forms the base of the food web and ecosystem energetics^{1,3}. Such windows of opportunity may therefore function as ‘ecosystem control points’⁴ with disproportionately high ecological processing rates affecting ecosystem dynamics relative to longer intervening time periods. These ecosystem control points are widely distributed across ecosystems and vary across spatial and temporal scales⁴. However, our understanding on the microbiology of the communities that facilitate ecosystem control points remains limited to date.

Owing to climate change, the mass balance and melting dynamics of mountain glaciers are rapidly changing worldwide, altering the annual distribution of runoff in GFSs⁵. Invigorated

glacial melt increases discharge and sediment delivery, but after glaciers shrink past a certain point (*i.e.*, ‘peak water’), GFSs are likely to become warmer, less turbid, and less hydrologically dynamic⁶. These changes are almost certain to have substantial impacts on GFS ecosystem structure and function by either contracting or extending the duration of these windows of opportunity. It is therefore critical to understand how benthic biofilms operate during these times in order to predict how these ecosystems are likely to change and operate in the future⁶.

In streams, and GFSs in particular, biofilms closely interact with the sedimentary environment. For instance, the extracellular polymeric substances (EPS) produced by biofilms can stabilize fine sediments⁷. On the other hand, larger sediments (*e.g.*, boulders) resist flow-induced disturbance, thereby conferring stability to biofilms⁸. This appears particularly important in GFSs characterized by notoriously unstable fine sediments. Furthermore, light is minimally attenuated within a thin layer of water flowing over protruding boulders, which facilitates photosynthesis and biofilm growth. Therefore, it is particularly advantageous for phototrophic biofilms to colonize boulders, which form islands of stability in an otherwise highly unstable ecosystem. This may allow them to persist even during unfavorable periods, and potentially provide a model for microbial life during the windows of opportunity.

The relationship between photoautotrophs, such as algae and cyanobacteria, and microorganisms, primarily other bacteria, regulates nutrient and carbon cycling, and therefore represents a fundamental ecological interface in aquatic ecosystems. This interface (*i.e.*, the phycosphere) has received substantial attention in pelagic ecosystems over the last decades^{9–12}, but less so in stream ecosystems. While early work on phototrophic biofilms colonizing the benthic zone in streams has highlighted the role of algal-bacterial interactions for carbon and nutrient fluxes^{13,14}, we do not currently understand the fine-scale mechanisms of such interactions. For example, cyanobacteria produce pigments that protect biofilms against harmful UV-radiation¹⁵, while mucilage-rich algal colonies (*e.g.*, *Hydrurus* spp.) provide labile organic matter to heterotrophic microorganisms and facilitate their attachment. Such interactions may foster facultative interactions between photoautotrophs and other microorganisms, which, similar to the phycosphere, may be particularly beneficial to microbial life in oligotrophic and harsh ecosystems such as GFSs. Unraveling the genomic and metabolic underpinnings of algal-bacterial relationships in biofilms helps to better understand a most successful mode of microbial life in an extreme ecosystem.

Here we dissect the microbiome of GFSs and describe unprecedented genomic underpinnings of the adaptive mechanisms that contribute to the success of complex biofilms. Using 16S rRNA and 18S rRNA genes amplicon sequencing, we assess the microbiome structure of biofilms associated with two sedimentary habitats that are common in GFSs, namely sandy sediments (*i.e.*, epipsammic biofilms) and boulders (*i.e.*, epilithic biofilms). Furthermore, using high, genome-resolved metagenomics, we screen twenty-one epilithic biofilm microbiomes for energy pathways and cross-domain metabolic interactions. Our findings suggest the diversification of energy-acquiring pathways and metabolic interactions as relevant for epilithic biofilms to adapt to these ecological windows of opportunity, which are likely to become more prevalent as glaciers worldwide recede.

Results and Discussion

Sedimentary habitats affect microbiome structure and assembly. We used 16S rRNA and 18S rRNA gene amplicon sequencing to compare the microbiome structure of 48 epipsammic and epilithic biofilm samples from GFSs in the New Zealand Southern Alps (NZ) and in Caucasus (CC) (*Methods*) (Fig. 1a; Supp. Fig. 1a-b). These geographically distant streams, transcending hemispheres, were selected in order to draw more generalisable conclusions about microbiome structure and assembly. Moreover, to have comparable samples, the collection was largely constrained to the vernal and autumnal windows of opportunity. We found that both the prokaryotic and eukaryotic communities differed between the two habitat types in terms of community structure and alpha diversity (Fig. 1b-c). Overall, taxonomic differences were even apparent at the phylum level, despite high inter-sample variability within the categories (Supp. Fig. 1c-d). Geography explained 11.5% and 12.9% of the variability in the prokaryotic and eukaryotic datasets (db-RDA, $p < 0.05$ for both datasets), while sedimentary habitats explained an additional 10% and 8.3% of the variability (db-RDA, $p < 0.05$ for prokaryotes and eukaryotes).

The estimated α -diversity (*i.e.*, species richness of amplicon sequence variants; ASVs) was higher for both prokaryotes and eukaryotes in epipsammic biofilms when compared to epilithic biofilms (2-3 fold differences, non-parametric t-tests, $p < 0.001$) (Fig. 1d-e). It is plausible that dispersal facilitated by the transport of fine sediments from various upstream sources (*e.g.*, subglacial environment, bare rock and soils) leads to the greater diversity of the epipsammic biofilms. Overall, our results unravel distinct microbiome structures for both habitats within the

same GFS reaches; our results thereby agree with previous studies demonstrating the relationship between streambed physical variation and spatial biodiversity dynamics^{16,17}. Streambeds, including their biofilms, are understood as landscapes where dispersal among patches can shape biodiversity and resilience^{18–20}. Therefore, we hypothesized that epilithic communities are partially structured by dispersal from epipsammic communities that typically dominate the GFS streambeds by area. Using Sloan’s neutral community model²¹, we instead found that the composition of the epilithic biofilms is not dictated by a source-sink relationship with the epipsammic communities (*Supplementary text*). In other words, the epilithic biofilm communities are not determined by epipsammic communities that typically surround the boulders within the complex landscape of the GFS streambed.

Metagenomics unveils the complexity of epilithic biofilms. To unveil the full complexity of the epilithic biofilms, we performed whole genome shotgun metagenomics on 21 epilithic samples from four GFSs each in NZ and CC (Supp. Fig. 1a-b); low biomass associated with sandy sediments precluded epipsammic biofilms from metagenomic analysis. High-resolution sequencing, after quality filtering yielded on average 1.2×10^8 ($\pm 1.4 \times 10^7$ s.d.) reads per sample which were assembled to obtain an average of 8.7×10^5 contigs per sample, that were subsequently binned. Bacteria and eukaryotes dominated the biofilm communities across all samples (Supp. Fig. 2a). Seventy-three (70 bacteria and 3 archaea) medium-to-high quality (>70% completion, <5% contamination) metagenome-assembled genomes (MAGs) from a total of 662 MAGs formed the pool of the prokaryotes. As seen from the phylogenomic analysis, the high-quality MAGs span the bacterial tree of life and based on this, along with the taxonomic information, many of these plausibly represent novel species (Fig. 2a). Aggregated at the genus level, *Polaromonas* was both abundant and prevalent in the biofilms along with representatives of *Flavobacterium*, *Cyanobacteria*, and unclassified MAGs from the Bacteroidota and Candidate Phyla Radiation (CPR; *Patescibacteria*) (Fig. 2b). These taxa were found in over half of the samples, irrespective of geographic origin. The CPR bacteria have only recently been identified based on genomic data²², and *Patescibacteria* specifically have been reported from oligotrophic ecosystems, including groundwater²³ and thermokarst lakes²⁴. Their apparently minimal biosynthetic and metabolic pathways may help them dwell in these ecosystems, which is of equal relevance in GFSs.

Alongside these bacteria, archaea contributed less than 1% to the microbiome of epilithic biofilms, with representatives of Asgardarchaeota, Crenarchaeota and Nanoarchaeota. Intriguingly, the recently discovered lineages of Asgardarchaeota^{25,26} have been reported from freshwater sediments, yet not from cryospheric environments. Algae, mostly diatoms and *Hydrurus* (Ochrophyta phylum), as well as dinoflagellata were the most important photoautotrophs of the eukaryotic domain (Fig. 2c). The prevalence of *Hydrurus* (~87% relative abundance) underscores the function of this filamentous alga as a resource to higher trophic levels in GFS²⁷. Our metagenomic insights further support the notion that phototrophic biofilms are highly diverse with representatives from all three domains of life²⁸.

In addition to the archaeal, bacterial and eukaryotic community members, we also found a diverse viral community associated with epilithic biofilms (Supp. Fig. 2b). Most of the viruses were bacteriophages targeting abundant MAGs such as *Flavobacterium*, *Pseudomonas*, and *Bacillus* genera, but we also identified eukaryotic phages (*i.e.*, *Paramecium bursaria* Chlorella virus). Few have studied viruses in stream biofilms to date²⁹, potentially because it was common wisdom that the biofilm mode of life protects bacteria from viral infection. While viruses have previously been shown to be abundant in glaciers^{30,31}, our findings are the first to provide evidence for a diverse and likely active viral community in GFS biofilms where they may influence bacterial growth and both carbon and nutrient cycling as on the glacier surface³⁰.

Epilithic biofilms form the basis for a ‘green’ food web. Cyanobacteria and eukaryotic algae figured among the most important photoautotrophs in the epilithic biofilms that form the basis of the ‘green’ food web during the window of opportunity. While these photoautotrophs are well known to use chlorophyll to capture solar energy, little is known on retinal-based phototrophy using rhodopsins in GFSs. Intriguingly, we found that MAGs from sixteen out of twenty phyla in the epilithic biofilms, including the abundant groups, such as Proteobacteria (*Polaromonas*) and Bacteroidota (*Flavobacterium*), encoded for (bacterio-)rhodopsins (Fig. 3a). These also included genes encoding for light-harvesting complex 1 (LH1), reaction centre (RC) subunits (*pufBALM*), and transcriptional regulators (*ppsR*) required for aerobic anoxygenic phototrophs along with rhodopsins as a signature of energy-limitation adaptations (Fig. 3a). Recently, rhodopsins were also reported to serve as a photoprotectant in *Flavobacterium* from glaciers³². Collectively, our

findings unveil multiple strategies of photoautotrophy, which may help cyanobacteria and algae in maximising the exploitation of solar energy and to thrive during windows of opportunity.

Rapid growth should be advantageous for primary producers, such as cyanobacteria, to best exploit the short windows of opportunity in GFSs. Moreover, functional independence from other microorganisms could thereby allow them to efficiently react to a window of opportunity. To test this hypothesis, we assessed the relationship between projected times of growth (doubling time in hours), with the median KEGG pathway completion within each MAG. Strikingly, most cyanobacterial MAGs ($n = 38/44$, 86%) exhibited decreased projected times of growth with respect to median KEGG module completion (Spearman's correlation: $r = -0.41$, adj. $p < 0.05$). These observations suggest that when encoding all genes to form a complete KEGG pathway, phototrophic taxa within these epilithic biofilms may indeed be self-sufficient, thereby reducing their dependency from other (micro)organisms and fostering growth.

Given the energetic constraints in GFSs, it would be beneficial for bacterial heterotrophs to interact with these photoautotroph (micro)organisms for meeting their energy and nutrient demands. To investigate such cross-domain relationships, we used network analyses and identified key interacting taxa based on positively co-occurring nodes, using all prokaryotic and eukaryotic MAGs (see *Methods*). Based on a null model assessment (see *Methods*), our interaction networks showed preferential attachment within the nodes, along with increased centralities (*i.e.*, degree and edge-betweenness, Supp. Fig. 3a-b), suggesting that the interactions within these networks were not random. More importantly, the largest connected component (based on degree and betweenness centralities) of the interaction network contained taxa spanning archaea, bacteria and eukaryotic domains (Fig. 3b and Supp. Fig. 3b). Though *Acidobacteria* had a high degree of centrality, both *Polaromonas* and *Methylobacter* demonstrated strong interactions (> 0.6 betweenness centrality) with primary producers (including eukaryotic algae) and fungi. Specifically, *Polaromonas* had a strong interaction with algae, while *Methylobacter* co-occurred with *Chytridiomycetes* (Fig. 3b). These results support our hypothesis of heterotrophic bacteria co-occurring with eukaryotes, primarily algae, for metabolic cross-feeding, similar to those occurring in the phycosphere¹⁰.

Furthermore, our results hint at the existence of a more cryptic interaction in epilithic biofilms between parasitic fungi *Chytridiomycetes* and algae (*i.e.*, *Ochrophyta*). Fungal parasitism on pelagic algae has been recently reported to be more important than expected, even with

consequences for carbon and nutrient cycling as mediated by the fungal shunt^{33,34}. The possibility of fungal parasitism on algae in epilithic biofilms further underlines the role of photoautotrophs as the foundation of a complex food web in GFSs as a typically energy-limited ecosystem.

Genomic underpinnings of algae-bacteria metabolic interactions. As photoautotrophs grow and senesce, they increasingly exude intracellular material into their ambient environment, where it can be metabolized by heterotrophic bacteria through the action of extracellular enzyme activity (EEA). To explore this metabolic cross-feeding between bacterial heterotrophs and algae, we assessed the MAGs for genes encoding five common EEAs required for cleaving complex polysaccharides, phosphomonoesters and proteins³⁵. Not unexpectedly, these genes were predominantly associated with bacterial heterotrophs, rather than with the photoautotrophs (Supp. Fig. 4), which suggests adapted genomic traits to meet specific metabolic needs of the heterotrophs. However, based on the presence of the EEA genes, especially among Cyanobacteria, we cannot discount the possibility of mixotrophy in the epilithic biofilms (Supp. Fig. 4a), including in other abundant members of the epilithic microbiome (Supp. Fig. 1c-d). The widespread occurrence of mixotrophy in planktonic communities, including Cyanobacteria, and the ensuing food web dichotomy is considered as an adaptive strategy to oligotrophic and cold ecosystems (*e.g.*, the polar sea). Therefore, we argue that mixotrophy may also be an important trait of Cyanobacteria within GFS biofilms.

Carbohydrate-active enzymes (CAZymes) are the prime tools used by heterotrophic bacteria to initiate the degradation of polysaccharides, largely algae-derived in the GFS epilithic biofilms. To shed light on this potential trophic interaction identified through specific EEAs, we tested if all the CAZymes in the metagenomes covaried with the abundance of eukaryotes. Overall, we found positive correlations between eukaryote abundances and CAZymes, particularly carbohydrate-binding modules (CBM) and glycoside hydrolases (GH) (Supp. Fig. 4d). More specifically, these correlations were particularly pronounced for GH and some of the algal groups (*e.g.*, Ochrophyta, Haptophyta, Cryptophyta) that we found at relatively high abundances in the epilithic biofilms (Fig. 4c). As some of these algae are known to copiously produce sulfated carbohydrates³⁶, we suggest a similar involvement of CAZymes (Supp. Table 1) in relation to polysaccharide degradation in GFS epilithic biofilms as recently reported from *Verrucomicrobia* isolates³⁷. Given that sulfated carbohydrates are more resistant to bacterial degradation than other

carbohydrates³⁷, our findings suggest that they are still relevant to carbon turnover in an ecosystem that is inherently carbon limited.

In order to understand whether functions potentially geared towards cross-domain interactions were enriched solely in epilithic biofilms, we compared the KEGG orthology (KO) annotations from our metagenomes to 105 metagenomes from a wide range of ecosystems (Supp. Table 2). Strikingly, we found that KOs associated with quorum sensing, vitamin B12 (cobalamin) transporters and thiamine biosynthesis were enriched in epilithic biofilms compared to other ecosystems (Supp. Table 3). The associated pathways and their completion levels were evaluated using KEGGDecoder (Fig. 4d; Supp. Fig. 5) indicating a high completion of pathways associated with cross-domain interactions. These findings are in line with previous genomic insights into algal-bacterial interactions^{38,39}, specifically with the observed upregulation of vitamin biosynthesis in bacteria (*Halomonas*) growing in the presence of algal extracts.

Furthermore, several MAGs were found to encode genes (*e.g.*, quorum sensing, cobalamin metabolism, tryptophan synthesis) potentially facilitating algal-bacterial interactions (Fig. 4a). Particularly, cobalamin metabolism may be relevant for nutrient acquisition in algal-bacterial relationships⁴⁰, whereas tryptophan was reported as a key signalling molecule involved in interactions between bacteria and associated phytoplankton^{11,41}. Collectively these genomic insights stress cross-domain interactions as an adaptive potential that the epilithic microorganisms have developed to exploit the window of opportunity in GFSs.

Energy acquisition and biogeochemical pathways in epilithic biofilm MAGs. The dominance (~88%) of MAGs encoding for organic carbon metabolism highlights the relevance of a ‘green food web’ during the windows of opportunity, potentially sustaining metabolic interactions between primary producers and heterotrophs. Further exploring the gene repertoire of the epilithic biofilms, we found that Cyanobacteria were one of the largest bacterial contributors to carbon fixation along with Bacteroidota and few Gammaproteobacteria (Fig. 4a). An in-depth analysis across the 662 MAGs revealed that 583 MAGs encoded genes involved in organic carbon oxidation, while 120 MAGs encoded genes involved in CO₂ fixation. In line with the above findings, the majority of these MAGs was identified as Cyanobacteria along with few other phyla such as Proteobacteria, Asgardarchaeota, Crenarchaeota and Huberarchaeota. We also note that

351 MAGs encoded genes for fermentation (Fig. 4b) spanning several phyla, including Actinobacteriota, Bacteroidota, Patescibacteria, Planctomycetota and Verrucomicrobiota.

For biofilms to thrive in GFSs, even during the windows of opportunity, it appears opportune to diversify the exploitation of energy sources. Therefore, we performed an in-depth characterisation of chemolithotrophic pathways to disclose the potential role of minerals derived from the glacial comminution of bedrock as an energy source for microorganisms⁴². The prevalence of the *sox* gene cluster in representatives of the Bacteriodota (UBA7662) and Bdellovibrionota reveals the potential importance of inorganic sulfur oxidation in epilithic biofilms. This notion is supported by the broad occurrence of sulfur dioxygenases (SDOs) across the various phyla that facilitate sulfur oxidation (Fig. 4c). Interestingly, Tranter and Raiswell suggested that sulfates are derived from sulfide oxidation in comminuted bedrock⁴³ potentially increasing sulfur availability and acquisition in glacial meltwaters⁴⁴. Sulfide oxidation can stimulate carbonate weathering with the resulting CO₂ potentially being fixed by algae and cyanobacteria in the epilithic biofilms — a link that appears relevant given that GFSs are often undersaturated in CO₂⁴⁵. Furthermore, we found that almost all MAGs encoded for group IV hydrogen dehydrogenases (NiFe_Gp4; Fig. 4c), which potentially serve as an alternate energy acquisition pathway. Hydrogen dehydrogenases have recently been reported to support primary production in various glacial and other extreme environments^{46,47}. This suggests that lithogenic hydrogen may also contribute energy to bacteria within the epilithic biofilms.

Further genomic insights into the nitrogen cycle revealed the Dissimilatory Nitrate Reduction to Ammonium (DNRA, or nitrite ammonification) and, to a lesser extent, by denitrification, as major pathways (Fig. 4d). Relatively little is known regarding these two competing pathways in stream biofilms or sediments⁴⁸. However, our insights into the cross-domain metabolic interactions suggest that epilithic algae provide significant amounts of organic carbon (*i.e.*, electron donors), which may favour bacteria to grow using DNRA. This is in line with other ecosystems where DNRA is favoured over denitrification when alternate electron donors prevail over nitrate⁴⁹. Our analyses revealed Burkholderiales (Gammaproteobacteria) as the largest contributor to nitrate assimilation and ammonia-oxidation genes (Fig. 5a). DNRA, if not conducive to N₂O production, would enhance nitrogen recycling within epilithic biofilms through ammonia assimilation by algae and cyanobacteria, for instance. Our genomic evidence for nitrogen recycling that potentially overwhelms nitrogen losses through denitrification is corroborated by flux

measurements from microbial mats in Antarctic GFSs⁵⁰, and highlights recycling as a strategy to cope with nutrient limitation in glacier ecosystems^{50–52}.

Strikingly, we found only few MAGs, mostly belonging to Deinococcota, Gammaproteobacteria, Beijerinckiaceae and Crenarchaeota, involved in the oxidation of ammonia and nitrite potentially leading to the accumulation of nitrate. The involvement of archaea would be in line with recent studies showing ammonia oxidation by archaea in Arctic soils⁵³ and with the observation that archaea couple ammonia oxidation with biomass formation (*i.e.*, via CO₂ fixation)⁵⁴. Our finding that archaeal MAGs encode for carbon fixation genes (Fig. 4b) further highlight their role in ammonia oxidation and biomass accrual in epilithic biofilms. Overall, the overlap of metabolic capacities within the MAGs suggests that the epilithic biofilms may typify a ‘closed system’, where both carbon and nutrients are efficiently recycled.

Genomic underpinnings of adaptation to the extreme GFS environment. The GFS environment is characterized by near-freezing temperatures, high UV-radiation, and high flow velocities. To assess potential adaptive traits of bacteria dwelling in epilithic biofilms, we first performed a phylogenomic analysis of *Polaromonas* spp., one of the most abundant and prevalent genera in the studied GFSs. Our analysis revealed that a few of the GFS *Polaromonas* formed clades that are distinct from *Polaromonas* identified in other environments (*Methods*), thus potentially comprising novel species’ (Fig. 5a). This phylogenomic pattern indicates that *Polaromonas* has evolved traits that facilitates its success in GFS, both in NZ and CC. To identify such traits, we created a pangenome and performed an enrichment analysis for clusters of orthologous genes. We found three categories that were significantly enriched in GFS *Polaromonas* compared to those from other environments (Supp. Table 4). Two categories are related to defense mechanisms, both general and transcription, and one to energy production (Fig. 5b). It is plausible that these mechanisms are related to high UV-radiation^{55,56} and oxidative stress⁵⁷, as well as to cold stress responses as previously reported from other bacteria^{58–60}. Furthermore, the presence of CRISPR-Cas proteins in the enriched clusters of orthologous genes (COGs) hint at defense mechanisms against phages (Supp. Table 4), which we showed to be present in the epilithic biofilms. This is in accordance with reports demonstrating that cryospheric bacteria (*Janthinobacterium* spp.) develop defense strategies, including biofilm formation⁶¹ and extracellular vesicle formation⁶² to escape viruses. On the other hand, the transcription of ‘defense

mechanisms' genes have been linked to cold adaptation in psychrophiles⁵⁸. Cold-shock proteins regulate transcription at low temperature, while genes involved in membrane biogenesis⁶³ and membrane transport proteins⁶⁴, several of which are also enriched in the GFS *Polaromonas* genomes, are up-regulated. For example, in the psychrophilic *Colwellia psychrerythraea* 34H, adaptation to cold includes the maintenance of the cell membrane liquid-crystalline state via the expression of genes involved in polyunsaturated fatty acid synthesis⁶⁵. Similarly, ATP-driven or proton motive secondary transport systems have been associated with solute transfers across membranes in bacteria and archaea as an adaptation to the cold⁶⁴.

Our insights into the adaptive potential of *Polaromonas* to the GFS environment prompted us to expand our search for adaptive traits across all MAGs from the epilithic biofilms. Querying for 76 genetic traits spanning nine categories related to cold adaptation⁵⁹, we found indeed distinct patterns of genomic adaptation across MAGs (Fig. 5c). Several MAGs encoded for genes associated with membrane and peptidoglycan alterations, cold and heat shock proteins, oxidative stress, and transcription/translation factors alongside DNA replication and repair. While all major phyla encoded for adaptive traits related to the outer membrane and cell wall, Proteobacteria were the predominant group with an overall higher copy number of genes involved in counteracting osmotic and oxidative stress. This is in line with metagenomic studies reporting an enrichment of sigma B genes in Antarctic mats, allowing for surviving severe osmotic stress during freezing⁶⁶. Similarly, *Psychrobacter arcticus*⁶⁷ and *Planococcus halocryophilus* Or1⁶⁸ were shown to have specific genomic modifications, particularly with genes involved in putrescine and spermidine accumulation, both of which are associated with alleviating oxidative stress. Furthermore, MAGs from Proteobacteria were characterized by high prevalence of genes potentially expressed in response to stressors, such as UV and reactive oxygen species (Fig. 5c).

In conclusion, our genome-resolved metagenomics analyses have set the stage for a mechanistic understanding of how the diversification of energy and matter acquisition pathways as well as metabolic interactions allow biofilms to thrive during windows of opportunity in GFSs. We acknowledge that a metagenomic time series outside and throughout windows of opportunity would be required to substantiate some of our observations. Nevertheless, our findings shed light on boulders as important habitats that confer stability of biofilms even outside the typical windows of opportunity. GFSs count among the ecosystems that are most vulnerable to climate change.

Therefore, our findings open a window into the future of how microbial life, with a strong photoautotrophic component, may look like in GFSs as glaciers shrink.

Material and methods

Sample collection. We sampled a total of eight GFSs from the New Zealand Southern Alps and the Russian Caucasus in early- and mid-2019, respectively, for a total of 27 epipsammic samples taken from sandy sediments and 21 epilithic biofilm samples from boulders adjacent to the epipsammic samples (Supp. Table 5). Epipsammic samples were collected from each GFS by first identifying three patches within a reach of ~5-10 m. From each patch, sandy sediments were taken from the <5 cm surface of the streambed with a flame-sterilized metal scoop and sieved to retain the 250 µm to 3.15 mm size fraction. While three epipsammic samples were taken from each stream, epilithic samples were taken opportunistically from up to three boulders per reach (Supp. Table 5). Epilithic biofilms were sampled using a sterilized metal spatula. All samples were immediately flash-frozen in liquid nitrogen in the field and transported and stored frozen pending DNA extraction. Streamwater turbidity, conductivity, temperature, and pH were measured *in situ* during the sampling (Supplementary Table 2).

DNA extraction and purification. A previously established protocol⁶⁹ was used to extract DNA from all samples. Briefly, 5 g of epipsammic and 0.05-0.1 g of epilithic biofilm were subjected to a phenol:chloroform-based extraction and purification method. The differential input volume for the DNA extractions were established to account for the differences in biomass between the epipsammic and epilithic biofilms. The samples were treated with a lysis buffer containing SDS along with 0.1 M Tris-HCl pH 7.5, 0.05 M EDTA pH 8, 1.25% SDS and RNase A (10 µl: 100 mg/ml). The samples were vortexed and incubated at 37 °C for 1 h. Proteinase K (100 µl; 20 mg/ml) was subsequently added and further incubated at 70 °C for 10 min. Samples were purified once with phenol/chloroform/isoamyl alcohol (ratio 25:24:1, pH 8) and the supernatant was subsequently extracted with a 24:1 ratio chloroform/isoamyl alcohol. Linear polyacrylamide (LPA) was used along with sodium acetate and ice-cold isopropanol for precipitating that DNA overnight at -20 °C. For epilithic biofilms, the entire protocol was adapted to a smaller scale due to the availability of higher DNA concentrations compared to sediment. The former was treated with 0.75 ml of lysis buffer (instead of 5 ml for sediment) and all subsequent volumes of reagents

were adapted accordingly (see supplementary material). Furthermore, a mechanical lysis step of bead-beating was necessary along with a lysis buffer to facilitate DNA release from the more developed epilithic biofilms. Due to the higher DNA yields, the addition of LPA was omitted from the DNA precipitation step. DNA quantification was performed for all samples with the Qubit dsDNA HS kit (Invitrogen).

Metabarcoding library preparation and sequencing. The prokaryotic 16S rRNA gene metabarcoding library preparation was performed as described in Fodelianakis *et al.*⁷⁰, targeting the V3-V4 hypervariable region of the 16S rRNA gene with the 341F/785R primers and following Illumina guidelines for 16S metagenomic library preparation for the MiSeq system. The eukaryotic 18S rRNA gene metabarcoding library preparation was performed likewise but using the TAREuk454F-TAREukREV3 primers to target the 18S rRNA gene V4 loop⁷¹. Samples were sequenced using a 300-bp paired-end protocol partly in the Genomic Technologies Facility of the University of Lausanne (27 epipsammic samples) and partly at the Biological Core Lab of the King Abdullah University of Science and Technology (21 epilithic samples).

Metabarcoding analyses. The 16S rRNA gene metabarcoding data were analysed using a combination of Trimmomatic⁷² and QIIME2⁷³ as described in Fodelianakis *et al.*⁷⁰, with the exception that here the latest SILVA database⁷⁴ v138.1 was used for taxonomic classification of 16S rRNA and 18S rRNA gene amplicons. Non-bacterial ASVs including those affiliated to archaea, chloroplasts and mitochondria were discarded from the 16S rRNA amplicon dataset in all downstream analyses. ASVs observed only once were removed from both 16S rRNA and 18S rRNA amplicon datasets. Diversity analyses were performed in R using the *vegan*⁷⁵ and *metacoder*⁷⁶ packages. To test for a source-sink hypothesis from sediments to rocks, the Sloan's Neutral Community Model²¹ was used based on the R implementation developed by Burns *et al.*⁷⁷.

Whole-genome shotgun libraries and sequencing. All epilithic biofilm DNA samples underwent random shotgun sequencing following library preparation using the NEBNext Ultra II FS library kit. Briefly, 50 ng of DNA was used for constructing metagenomic libraries under 6 PCR amplification cycles, following enzymatic fragmentation of the input DNA for 12.5 mins. The average insert size of the libraries was 450 bp. Qubit (Invitrogen) was used to quantify the libraries followed by quality assessment using the Bioanalyzer from Agilent. Sequencing was performed at

the Functional Genomics Centre Zurich on a NovaSeq (Illumina) using a S4 flowcell.

Metagenomic preprocessing, assembly, binning, and analyses. For processing metagenomic sequence data, we used the Integrated Meta-omic Pipeline (IMP)⁷⁸ workflow to process paired forward and reverse reads using version 3.0 (commit# 9672c874; available at <https://git-r3lab.uni.lu/IMP/imp3>), as previously described⁷⁹. IMP's workflow includes pre-processing, assembly, genome reconstructions and additional functional analysis of genes based on custom databases in a reproducible manner. Briefly, adapter trimming is followed by an iterative assembly using MEGAHIT v1.2.9⁸⁰. Concurrently, MetaBAT2 v2.12.1⁸¹ and MaxBin2 v2.2.7⁸² are used for binning in addition to an in-house method established previously⁷⁹ for reconstructing metagenome-assembled genomes (MAGs). Binning was completed by selecting a non-redundant set of MAGs using DASTool⁸³ based on a score threshold of 0.7. The quality of the MAGs was assessed using CheckM v1.1.3⁸⁴, while taxonomy was assigned using the GTDB-toolkit v1.4.1⁸⁵.

For the downstream analyses including identification of viruses, VIBRANT v1.2.1⁸⁶ was used on the metagenomic assemblies. The output from this was used to identify the viral taxa using vConTACT2 v0.9.22⁸⁷. Independently, the viral contigs were also validated using CheckV v0.7.0⁸⁸. To estimate the overall abundances of eukaryotes along with prokaryotes including archaea, we used EUKulele v1.0.5⁸⁹ with both the MMETSP and the PhyloDB databases, run separately, to confirm the detected eukaryotic profiles. To understand the overall metabolic and functional potential of the metagenome and reconstructed MAGs we used MANTIS⁹⁰. Additionally, we used METABOLIC v4.0⁹¹, metabolisHMM v2.21⁹², and Lithogenic from MagicLamp v1.0 (<https://github.com/Arkadiy-Garber/MagicLamp>) to identify metabolic and biogeochemical pathways relevant for determining nutritional phenotypes of all MAGs along with the '*anvi-estimate-metabolism*' function from *anvi'o*⁹³. This information was manually validated based on the different tools to identify which MAGs encode for the respective pathways. Subsequently, to determine the growth rates of prokaryotes, we used codon usage statistics for detecting optimization of genes that are highly expressed, as an indicator of maximal growth rates with gRodon v1.0⁹⁴. All the parameters, databases, and relevant code for the analyses described

above are openly available at https://git-r3lab.uni.lu/susheel.busi/nomis_pipeline and included in the Code availability section.

Eukaryote assembly and binning. To obtain eukaryotic MAGs, an alternate, custom pipeline (https://github.com/Mass23/NOMIS_ENSEMBLE/tree/coassembly) was established for coassembling the twenty-one epilithic biofilm sequence data with subsequent binning. Individual samples were first preprocessed similar to the workflow used in IMP, i.e., using FastP v0.20.0⁹⁵. Subsequently, the reads were deduplicated to avoid overlap and enhance computation efficiency using *clumpify.sh* from the BBmap suite v38.79⁹⁶. Thereafter, any reads mapping to bacteria or viruses were removed by filtering the reads against a Kraken2 v2.0.9beta⁹⁷ maxikraken database available at https://lomanlab.github.io/mockcommunity/mc_databases.html. Only reads that were unknown or mapping to eukaryotes were retained and concatenated. This was followed by another round of deduplication using *clumpify.sh*. The concatenated reads were assembled using MEGAHIT v1.2.7 with the following options: `--kmin-1pass -m 0.9 --k-list 27,37,47,57,67,77,87 -min-contig-len 1000`. Following assembly, EukRep v0.6.7⁹⁸ was used for retrieving eukaryotic contigs with a minimum length of 2000 bp and the ‘-m strict’ flag. These contigs were used for binning into MAGs as described herein.

Eukaryotic MAGs were binned using CONCOCT v1.1.0⁹⁹. To do this, coverages were estimated for the contigs by mapping the reads of all samples against the contigs using the coverm v0.6.1 (<https://github.com/wwood/CoverM>) to generate bam files. These files were then used to generate a table with coverage depth information per sample. The protein coding genes of the MAGs was predicted with MetaEuk v4.a0f584d¹⁰⁰ with their in-house database made with MERC, MMETSP and Uniclust50 (<http://wwwuser.gwdg.de/~compbiol/metaeuk/>). The annotation was then subsequently done with eggNOG-mapper v2.1.0¹⁰¹. The completeness and contamination of the MAGs were assessed with Busco v5.0.0¹⁰² and the eukaryotic lineage (255 genes). We determined their taxonomy by comparing the results of the EUKulele v1.0.3⁸⁹ and EukCC v0.3¹⁰³ along with homology comparisons with publicly available genomes not included in the previous tools by protein BLAST v2.10.0¹⁰⁴.

Co-occurrence interaction networks. Co-occurrence networks between the pro- and eukaryotic MAGs were constructed using an average of the distance matrices created from SparCC¹⁰⁵,

Spearman's correlation and SpiecEasi¹⁰⁶ where the networks were constructed using the 'Meinshausen and Bühlmann (mb)' method. Nodes with fewer than two degrees were discarded to identify cliques with three or more interactions, while negative edges were removed to visualize only mutualistic relationships. The matrix was visualised using the *igraph*¹⁰⁷ R package. The largest component from the overall co-occurrence network was determined using the *components* module of the *igraph* package. Null model hypothesis was tested by assessing the distribution of the node degree and the respective probabilities of the occurrence network against those simulating the Erdos-Renyi, Barabasi-Albert, Stochastic-block null models¹⁰⁸.

Phylogenomics and pangenomes. For the pangenome analyses, we collected all the bins taxonomically identified as *Polaromonas* spp. and used the pangenome workflow described by Meren *et al.* (<http://merenlab.org/2016/11/08/pangenomics-v2/>) using anvi'o⁹³, along with NCBI¹⁰⁹ refseq genomes for comparison and an outgroup from the closely related *Rhodoferrax* genus. The choice of *Polaromonas* spp. was based on its high abundance and prevalence within the epilithic biofilms. The accession IDs from the reference genomes obtained from NCBI are provided in the supplementary material. The pangenome was run using the *--min-bit 0.5*, *--mcl-inflation 10* and *--min-occurrence 2* parameters, excluding the partial gene calls. A phylogenomic tree was built using MUSCLE v3.8.155¹¹⁰ and FastTree2 v2.1.10¹¹¹ on all single-copy gene clusters in the pangenome that were present in at least 30 genomes and had a functional homogeneity index below 0.9, and geometric homogeneity index above 0.9. The phylogenomic tree was used to order the genomes, the frequency of gene clusters (GC) to order the GC dendrogram. A phylogenomic bacterial tree of life containing the 47 high-quality MAGs along with 264 NCBI bacterial genomes was built based on a set of 74 single-copy genes using the GToTree v1.5.51¹¹² pipeline with the *-D* parameter, allowing to retrieve taxonomic information for the NCBI accessions. Briefly, HMMER3 v3.3.2¹¹³ was used to retrieve the single-copy genes after gene-calling with Prodigal v2.6.3¹¹⁴ and aligned using TrimAl v1.4.rev15¹¹⁵. The entire workflow is based on GNU Parallel v20210222¹¹⁶.

Data analyses and figures. Figures for the study including visualizations derived from the taxonomic and functional components, were created using version 3.6 of the R statistical software package¹¹⁷. The maps indicating the collection sites were generated using the *ggmap*¹¹⁸ package

in R. KEGGDecoder¹¹⁹ was used to assess enriched KEGG orthology (KO) IDs in comparison to 105 publicly available metagenome sampled in various ecosystems at a global scale (Supp. Tables 3 and 6), which were processed using the IMP workflow. *DESeq2*¹²⁰ with FDR-adjustments for multiple testing were used to assess KOs significantly enriched in the GFS metagenomes compared to this comparison dataset. The volcano plot highlighting the significant KOs was generated using the *EnhancedVolcano*¹²¹ R package. Figures from metabarcoding data were also generated in Rv3.6 using the *ggplot2*¹²² package and were further annotated graphically using Inkscape.

Data availability

Raw sequencing data samples and the MAGs are available at NCBI's sequence read archive under BioProject accession **PRJNA733707**. The Biosample accession IDs and the metadata associated with each sample are listed under Supp. Table 5.

Competing interests

The authors declare no conflicts of interest.

Code availability

The detailed code used for the for the downstream functional and growth analyses is available at https://git-r3lab.uni.lu/susheel.busi/nomis_pipeline. The custom pipeline for eukaryote analyses can be found here: https://github.com/Mass23/NOMIS_ENSEMBLE/tree/coassembly. Subsequent binning and manual refinement of eukaryotic MAGs was done as described here: https://git-r3lab.uni.lu/susheel.busi/nomis_pipeline/-/blob/master/workflow/notes/MiscEUKMAGs.md A snippet of the relevant results have been uploaded to Zenodo at <https://doi.org/10.5281/zenodo.5545722>.

Funding

This research was funded by The NOMIS Foundation to TJB. SBB was supported by the Synergia grant (CRSII5_180241: Swiss National Science Foundation) to TJB. LdN and PW are supported by the Luxembourg National Research Fund (FNR; PRIDE17/11823097). RM and DD are supported by King Abdullah University of Science and Technology through the baseline research funds to DD.

Acknowledgements

We are thankful for the assistance of Audrey Frachet Bour, Lea Grandmougin, Janine Habier, Laura Lebrun (LCSB) and Emmy Marie Oppliger (EPFL) for laboratory support. We are grateful to Alex Washburne for his feedback on the draft, and we also acknowledge the valuable input from Rashi Halder at the LCSB Sequencing Platform with respect to library preparation. We are equally grateful for the valuable insights into metagenomic processing from Patrick May and Cedric Christian Laczny, and especially Valentina Galata with the python scripts and Snakemake workflows. The computational analyses presented in this paper were carried out using the HPC facilities at the University of Luxembourg (<https://hpc.uni.lu>)¹²³.

Author contributions

SBB, MB, SF, HP, PW, and TJB conceived of the project. MiST, MT, VDS, MaSc, and HP conducted the fieldwork. PP and EMO extracted DNA, SBB and PP prepared the metagenomic and metabarcoding libraries, and RM and DD performed the sequencing. SBB conceptualized the data analyses, while SBB, MB, SF, GM performed the analyses. LdN contributed to the python scripts and Snakemake workflows for the analyses. SBB, MB, TJK, PW and TJB wrote the manuscript with significant input and editing from all coauthors.

Figure legends

Figure 1. Sedimentary habitats affect microbiome structure and assembly

(a) Representative images of sample collection indicating GFS and adjacent epilithic biofilm (left) with images of epilithic biofilms (right). Photo credits: Martina Schön. Ordination analyses of the epipsammic and epilithic biofilm based on prokaryote (b) and eukaryote (c) metabarcoding profiles from Southern Alps and Caucasus. Microbial richness across geographic locations and sample types in (e) prokaryotes and (f) eukaryotes.

Figure 2. Metagenomics unveils the complexity of epilithic biofilms

(a) Bacterial phylogenetic tree constructed using high-quality (>90% completion and <2% contamination) MAGs reconstructed from the epilithic biofilms. The numbers beside the phylum names indicate the number of high-quality MAGs assigned to the respective phylum. (b)

Normalized abundance of reconstructed prokaryotic genomes, *i.e.*, MAGs, from the epilithic biofilms. Taxonomy at phylum and genus levels is depicted. NA: unclassified genus. Samples from the Southern Alps are indicated in red, while those from Caucasus are shown in blue. (c) Eukaryotic relative abundance profile obtained from metagenomic sequencing across all epilithic biofilms samples.

Figure 3. Epilithic biofilms are the basis for a ‘green food chain’

(a) Abundance of genes involved in energy production (*light-harvesting complex*, *transcriptional regulator for phototrophy*, and *rhodopsin*) and photo-heterotrophic interactions (*cobalamin metabolism* and *tryptophan synthesis*), across all prokaryotic phyla are represented in the heatmap. Values indicate the \log_{10} abundance per gene within the phyla. (b) Largest component of the co-occurrence network between pro- and eukaryotic MAGs. Each node corresponds to a MAG (pro- or eukaryote). Size of the node corresponds to degree centrality and the edges represent the positive coefficients of correlation between each node. Colour of each node represents the phylum annotation. NA: unclassified genus. (c) Spearman’s correlation analyses of relative abundances of eukaryotic primary producers with the CAZyme abundances. FDR-adjusted *p*-values are indicated by *, *i.e.*, * < 0.05, ** < 0.01, *** < 0.001. (d) KEGG orthology (KO) pathways enriched in epilithic biofilms compared to publicly available cryospheric metagenomes were further assessed via KEGGDecoder for pathway completion and are displayed. The completeness of the pathways is indicated in the heatmap, per sample.

Figure 4. Functional redundancies across MAGs enable diverse energy acquisition and biogeochemical pathways

(a) The alluvial plot represents the metabolic pathways identified within all prokaryotic MAGs, with the respective taxonomic classification and category of nutrients. (b) Total number of MAGs encoding genes for and involved in the Carbon cycle (*see Methods*) are depicted in the flow gram created using a modified script from METABOLIC⁹¹. Each sub-pathway is indicated as a step with the corresponding number of genomes encoding the respective genes. (c) Phylum and order-level distributions of chemolithotrophic (hydrogen, nitrogen and sulfur) pathways with the respective gene copies per pathway are depicted in the heatmap. (d) Flow diagram indicating the MAGs

encoding for pathways in the nitrogen cycle (*Methods*). Each sub-pathway is indicated as a step with the corresponding number of genomes encoding the respective genes.

Figure 5. Genomic underpinnings of adaptation to the extreme GFS environment

(a) Phylogenomic tree based on *Polaromonas* genomes recovered from Southern Alps (red) and Caucasus (blue) along with publicly available genomes (grey) and an outgroup (*Rhodospirillum rubrum*, dark grey). (b) Clusters of orthologous (COG20) group pathways enriched in epilithic biofilms MAGs compared to the reference genomes are depicted in the barplot. (c) Heatmap representing the abundance of genes involved in cold adaptation. Taxonomy at phylum and order levels is depicted. Columns indicate clusters of orthologous groups associated with adaptive genes.

Supplementary Figure 1. Sediment and epilithic biofilm sites

Regions indicating the collection sites for the epilithic and epipsammic biofilms from (a) Caucasus and (b) Southern Alps. Relative abundance of prokaryotes (c) and eukaryotes (d) at the phylum and subdomain levels based on the sequencing of the 16S and 18S rRNA genes, respectively.

Supplementary Figure 2. Epilithic biofilm metagenomic profiles

(a) Relative abundance profiles across the three domains of life: archaea, bacteria and eukaryotes in the epilithic biofilms, obtained from the sample metagenomes. Samples from the Southern Alps are indicated in red, while those from Caucasus are shown in blue. (b) Virome profile indicating the top 50 viruses. Scaled abundance from low (-2) to high (2) are indicated in the heatmap.

Supplementary Figure 3. Cross-domain interactions and adaptations of epilithic biofilms

(a) Corplot based on Spearman's correlation between pro- and eukaryotic MAGs aggregated at the phylum level. (b) Co-occurrence network of all MAGs. Each node represents a MAG, while the size represents the degree centrality. The edges represent the positive coefficient of co-occurrence along with the corresponding betweenness centrality between the MAGs. Unconnected nodes represent MAGs with lower betweenness (< 0.5) compared to other MAGs. The color of the nodes represents the individual taxa, while the lines represent the edges connecting the nodes. The thickness of the lines indicates those edges with a betweenness greater than 0.5.

Supplementary Figure 4. Extracellular enzyme genes based on lifestyle

The classification at phylum and genus levels of MAGs identified as (a) heterotrophs, (b) phototrophs, or (c) those with ‘unknown’ trophic metabolisms are depicted, showing the abundance of genes encoding for extracellular enzymes. NA: unclassified genus; AG: α -1,4-glucosidase; BG: β -1,4-glucosidase; LAP: leucine aminopeptidase; NAG: β -1,4-N-acetylglucosaminidase; AP: acid (alkaline) phosphatase. (d) (c) Spearman’s correlation analyses of overall eukaryote relative abundances with the CAZyme abundances. FDR-adjusted p -values are indicated by *, *i.e.*, * < 0.05, ** < 0.01, *** < 0.001.

Supplementary Figure 5. Comparison to public metagenomes reveals differential gene abundances

Volcano plot indicating the total number of KOs (n = 9,335; total = 17,406) enriched in epilithic biofilms compared to 105 publicly available metagenomes.

References

1. Uehlinger, U., Robinson, C. T., Hieber, M. & Zah, R. The physico-chemical habitat template for periphyton in alpine glacial streams under a changing climate. in *Global Change and River Ecosystems—Implications for Structure, Function and Ecosystem Services* (eds. Stevenson, R. J. & Sabater, S.) 107–121 (Springer Netherlands, 2010).
2. Battin, T. J., Wille, A., Psenner, R. & Richter, A. Large-scale environmental controls on microbial biofilms in high-alpine streams. *Biogeosciences* **1**, 159–171 (2004).
3. Boix Canadell, M. *et al.* Regimes of primary production and their drivers in Alpine streams. *Freshw. Biol.* **66**, 1449–1463 (2021).
4. Bernhardt, E. S. *et al.* Control Points in Ecosystems: Moving Beyond the Hot Spot Hot Moment Concept. *Ecosystems* **20**, 665–682 (2017).
5. Huss, M. & Hock, R. Global-scale hydrological response to future glacier mass loss. *Nat. Clim. Chang.* **8**, 135–140 (2018).

6. Milner, A. M. *et al.* Glacier shrinkage driving global changes in downstream systems. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 9770–9778 (2017).
7. Roncoroni, M., Brandani, J., Battin, T. I. & Lane, S. N. Ecosystem engineers: Biofilms and the ontogeny of glacier floodplain ecosystems. *WIREs Water* **6**, e1390 (2019).
8. Hoyle, J. T., Kilroy, C., Hicks, D. M. & Brown, L. The influence of sediment mobility and channel geomorphology on periphyton abundance. *Freshw. Biol.* **62**, 258–273 (2017).
9. Cole, J. J. Interactions Between Bacteria and Algae in Aquatic Ecosystems. *Annu. Rev. Ecol. Syst.* **13**, 291–314 (1982).
10. Seymour, J. R., Amin, S. A., Raina, J.-B. & Stocker, R. Zooming in on the phycosphere: the ecological interface for phytoplankton–bacteria relationships. *Nature Microbiology* vol. 2 (2017).
11. Amin, S. A. *et al.* Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria. *Nature* **522**, 98–101 (2015).
12. Christie-Oleza, J. A., Sousoni, D., Lloyd, M., Armengaud, J. & Scanlan, D. J. Nutrient recycling facilitates long-term stability of marine microbial phototroph–heterotroph interactions. *Nature Microbiology* **2**, 1–10 (2017).
13. Haack, T. K. & McFeters, G. A. Nutritional relationships among microorganisms in an epilithic biofilm community. *Microb. Ecol.* **8**, 115–126 (1982).
14. Kaplan, L. A. & Bott, T. L. Diel fluctuations in bacterial activity on streambed substrata during vernal algal blooms: Effects of temperature, water chemistry, and habitat. *Limnol. Oceanogr.* **34**, 718–733 (1989).
15. Vincent, W. F., Downes, M. T., Castenholz, R. W. & Howard-Williams, C. Community structure and pigment organisation of cyanobacteria-dominated microbial mats in

- Antarctica. *European Journal of Phycology* vol. 28 213–221 (1993).
16. Besemer, K., Singer, G., Hödl, I. & Battin, T. J. Bacterial community composition of stream biofilms in spatially variable-flow environments. *Appl. Environ. Microbiol.* **75**, 7189–7195 (2009).
17. Risse-Buhl, U. *et al.* Near streambed flow shapes microbial guilds within and across trophic levels in fluvial biofilms. *Limnol. Oceanogr.* **65**, 2261–2277 (2020).
18. Palmer, M. A., Swan, C. M., Nelson, K., Silver, P. & Alvestad, R. Streambed landscapes: evidence that stream invertebrates respond to the type and spatial arrangement of patches. *Landsc. Ecol.* **15**, 563–576 (2000).
19. Battin, T. J. *et al.* Microbial landscapes: new paths to biofilm research. *Nat. Rev. Microbiol.* **5**, 76–81 (2007).
20. Dzubakova, K. *et al.* Environmental heterogeneity promotes spatial resilience of phototrophic biofilms in streambeds. *Biol. Lett.* **14**, (2018).
21. Sloan, W. T. *et al.* Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environ. Microbiol.* **8**, 732–740 (2006).
22. Hug, L. A. *et al.* A new view of the tree of life. *Nat Microbiol* **1**, 16048 (2016).
23. Chaudhari, N. M., Overholt, W. A. & Figueroa-Gonzalez, P. A. The economical lifestyle of CPR bacteria in groundwater allows little preference for environmental drivers. *bioRxiv* (2021).
24. Vigneron, A. *et al.* Ultra-small and abundant: Candidate phyla radiation bacteria are potential catalysts of carbon transformation in a thermokarst lake ecosystem. *Limnol. Oceanogr. Lett.* **5**, 212–220 (2020).
25. Liu, Y. *et al.* Expanded diversity of Asgard archaea and their relationships with eukaryotes.

Nature **593**, 553–557 (2021).

26. Cai, M. *et al.* Ecological features and global distribution of Asgard archaea. *Sci. Total Environ.* **758**, 143581 (2021).

27. Niedrist, G. H. & Füreder, L. When the going gets tough, the tough get going: The enigma of survival strategies in harsh glacial stream environments. *Freshwater Biology* vol. 63 1260–1272 (2018).

28. Bengtsson, M. M., Wagner, K., Schwab, C., Urich, T. & Battin, T. J. Light availability impacts structure and function of phototrophic stream biofilms across domains and trophic levels. *Mol. Ecol.* **27**, 2913–2925 (2018).

29. Payne, A. T. *et al.* Widespread cryptic viral infections in lotic biofilms. *Biofilms* **2**, 100016 (2020).

30. Anesio, A. M., Mindl, B., Laybourn-Parry, J., Hodson, A. J. & Sattler, B. Viral dynamics in cryoconite holes on a high Arctic glacier (Svalbard). *J. Geophys. Res.* **112**, (2007).

31. Bellas, C. M., Schroeder, D. C., Edwards, A., Barker, G. & Anesio, A. M. Flexible genes establish widespread bacteriophage pan-genomes in cryoconite hole ecosystems. *Nat. Commun.* **11**, 4403 (2020).

32. Liu, Q. *et al.* Light stimulates anoxic and oligotrophic growth of glacial *Flavobacterium* strains that produce zeaxanthin. *ISME J.* **15**, 1844–1857 (2021).

33. Sánchez Barranco, V. *et al.* Trophic position, elemental ratios and nitrogen transfer in a planktonic host-parasite-consumer food chain including a fungal parasite. *Oecologia* **194**, 541–554 (2020).

34. Klawonn, I. *et al.* Characterizing the ‘fungal shunt’: Parasitic fungi on diatoms affect carbon flow and bacterial communities in aquatic microbial food webs. *Proc. Natl. Acad. Sci. U. S.*

A. **118**, (2021).

35. Sinsabaugh, R. L., Hill, B. H. & Follstad Shah, J. J. Ecoenzymatic stoichiometry of microbial organic nutrient acquisition in soil and sediment. *Nature* **462**, 795–798 (2009).
36. Avcı, B., Krüger, K., Fuchs, B. M., Teeling, H. & Amann, R. I. Polysaccharide niche partitioning of distinct *Polaribacter* clades during North Sea spring algal blooms. *ISME J.* **14**, 1369–1383 (2020).
37. Sichert, A. *et al.* Verrucomicrobia use hundreds of enzymes to digest the algal polysaccharide fucoidan. *Nat Microbiol* **5**, 1026–1039 (2020).
38. Zhou, J., Lyu, Y., Richlen, M., Anderson, D. M. & Cai, Z. Quorum sensing is a language of chemical signals and plays an ecological role in algal-bacterial interactions. *CRC Crit. Rev. Plant Sci.* **35**, 81–105 (2016).
39. Croft, M. T., Lawrence, A. D., Raux-Deery, E., Warren, M. J. & Smith, A. G. Algae acquire vitamin B12 through a symbiotic relationship with bacteria. *Nature* **438**, 90–93 (2005).
40. Grossman, A. Nutrient Acquisition: The Generation of Bioactive Vitamin B12 by Microalgae. *Current biology: CB* vol. 26 R319–21 (2016).
41. Segev, E. *et al.* Dynamic metabolic exchange governs a marine algal-bacterial interaction. *Elife* **5**, (2016).
42. Anesio, A. M., Lutz, S., Christmas, N. A. M. & Benning, L. G. The microbiome of glaciers and ice sheets. *NPJ Biofilms Microbiomes* **3**, 10 (2017).
43. Tranter, M., Mills, R. & Raiswell, R. Chemical weathering reactions in Alpine glacial meltwaters. in *International symposium on water-rock interaction* 687–690 (1989).
44. Tranter, M., Brown, G., Raiswell, R., Sharp, M. & Gurnell, A. A conceptual model of solute acquisition by Alpine glacial meltwaters. *J. Glaciol.* **39**, 573–581 (1993).

45. St Pierre, K. A. *et al.* Proglacial freshwaters are significant and previously unrecognized sinks of atmospheric CO₂. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 17690–17695 (2019).
46. Dunham, E. C., Dore, J. E., Skidmore, M. L., Roden, E. E. & Boyd, E. S. Lithogenic hydrogen supports microbial primary production in subglacial and proglacial environments. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
47. Hernández, M. *et al.* Reconstructing Genomes of Carbon Monoxide Oxidisers in Volcanic Deposits Including Members of the Class Ktedonobacteria. *Microorganisms* **8**, (2020).
48. Quick, A. M. *et al.* Nitrous oxide from streams and rivers: A review of primary biogeochemical pathways and environmental variables. *Earth-Sci. Rev.* **191**, 224–262 (2019).
49. Kuypers, M. M. M., Marchant, H. K. & Kartal, B. The microbial nitrogen-cycling network. *Nat. Rev. Microbiol.* **16**, 263–276 (2018).
50. Gooseff, M. N., McKnight, D. M., Runkel, R. L. & Duff, J. H. Denitrification and hydrologic transient storage in a glacial meltwater stream, McMurdo Dry Valleys, Antarctica. *Limnol. Oceanogr.* **49**, 1884–1895 (2004).
51. Varin, T., Lovejoy, C., Jungblut, A. D., Vincent, W. F. & Corbeil, J. Metagenomic profiling of Arctic microbial mat communities as nutrient scavenging and recycling systems. *Limnol. Oceanogr.* **55**, 1901–1911 (2010).
52. Kohler, T. J. *et al.* Patterns and Drivers of Extracellular Enzyme Activity in New Zealand Glacier-Fed Streams. *Front. Microbiol.* **11**, 591465 (2020).
53. Alves, R. J. E. *et al.* Ammonia Oxidation by the Arctic Terrestrial Thaumarchaeote *Candidatus Nitrosocosmicus arcticus* Is Stimulated by Increasing Temperatures. *Front. Microbiol.* **10**, 1571 (2019).

54. Könneke, M. *et al.* Ammonia-oxidizing archaea use the most energy-efficient aerobic pathway for CO₂ fixation. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 8239–8244 (2014).
55. Cockell, C. S. *et al.* Influence of ice and snow covers on the UV exposure of terrestrial microbial communities: dosimetric studies. *J. Photochem. Photobiol. B* **68**, 23–32 (2002).
56. Sommaruga, R. The role of solar UV radiation in the ecology of alpine lakes. *J. Photochem. Photobiol. B* **62**, 35–42 (2001).
57. Margesin, R. & Collins, T. Microbial ecology of the cryosphere (glacial and permafrost habitats): current knowledge. *Appl. Microbiol. Biotechnol.* **103**, 2537–2549 (2019).
58. De Maayer, P., Anderson, D., Cary, C. & Cowan, D. A. Some like it cold: understanding the survival strategies of psychrophiles. *EMBO Rep.* **15**, 508–517 (2014).
59. Tribelli, P. M. & López, N. I. Reporting Key Features in Cold-Adapted Bacteria. *Life* **8**, (2018).
60. Varin, T., Lovejoy, C., Jungblut, A. D., Vincent, W. F. & Corbeil, J. Metagenomic analysis of stress genes in microbial mat communities from Antarctica and the High Arctic. *Appl. Environ. Microbiol.* **78**, 549–559 (2012).
61. Alonso-Sáez, L. *et al.* Winter bloom of a rare betaproteobacterium in the Arctic Ocean. *Front. Microbiol.* **5**, 425 (2014).
62. Hornung, C. *et al.* The *Janthinobacterium* sp. HH01 genome encodes a homologue of the *V. cholerae* CqsA and *L. pneumophila* LqsA autoinducer synthases. *PLoS One* **8**, e55045 (2013).
63. Maillot, N. J., Honoré, F. A., Byrne, D., Méjean, V. & Genest, O. Cold adaptation in the environmental bacterium *Shewanella oneidensis* is controlled by a J-domain co-chaperone protein network. *Commun Biol* **2**, 323 (2019).

64. Konings, W. N., Albers, S.-V., Koning, S. & Driessen, A. J. M. The cell membrane plays a crucial role in survival of bacteria and archaea in extreme environments. *Antonie Van Leeuwenhoek* **81**, 61–72 (2002).
65. Methé, B. A. *et al.* The psychrophilic lifestyle as revealed by the genome sequence of *Colwellia psychrerythraea* 34H through genomic and proteomic analyses. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 10913–10918 (2005).
66. Metagenomic Analysis of Stress Genes in Microbial Mat Communities from Antarctica and the High Arctic. <https://journals.asm.org/doi/abs/10.1128/AEM.06354-11>.
67. Ayala-del-Río, H. L. *et al.* The genome sequence of *Psychrobacter arcticus* 273-4, a psychroactive Siberian permafrost bacterium, reveals mechanisms for adaptation to low-temperature growth. *Appl. Environ. Microbiol.* **76**, 2304–2312 (2010).
68. Mykytczuk, N. C. S. *et al.* Bacterial growth at –15 °C; molecular insights from the permafrost bacterium *Planococcus halocryophilus* Or1. *The ISME Journal* vol. 7 1211–1226 (2013).
69. Busi, S. B. *et al.* Optimised biomolecular extraction for metagenomic analysis of microbial biofilms from high-mountain streams. *PeerJ* **8**, e9973 (2020).
70. Fodelianakis, S. *et al.* Microdiversity characterizes prevalent phylogenetic clades in the glacier-fed stream microbiome. *ISME J.* (2021) doi:10.1038/s41396-021-01106-6.
71. Stoeck, T. *et al.* Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol. Ecol.* **19 Suppl 1**, 21–31 (2010).
72. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

73. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
74. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–6 (2013).
75. Dixon, P. VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science* vol. 14 927–930 (2003).
76. Foster, Z. S. L., Sharpton, T. J. & Grünwald, N. J. Metacoder: An R package for visualization and manipulation of community taxonomic diversity data. *PLoS Comput. Biol.* **13**, e1005404 (2017).
77. Burns, A. R. *et al.* Contribution of neutral processes to the assembly of gut microbial communities in the zebrafish over host development. *ISME J.* **10**, 655–664 (2016).
78. Narayanasamy, S. *et al.* IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol.* **17**, 260 (2016).
79. Heintz-Buschart, A. *et al.* Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat Microbiol* **2**, 16180 (2016).
80. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
81. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).
82. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).

83. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* **3**, 836–843 (2018).
84. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
85. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz848.
86. Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90 (2020).
87. Zablocki, O., Jang, H. B., Bolduc, B. & Sullivan, M. B. vConTACT 2: A Tool to Automate Genome-Based Prokaryotic Viral Taxonomy. in *Plant and Animal Genome XXVII Conference (January 12-16, 2019)* (PAG, 2019).
88. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).
89. Krinos, A. I., Hu, S. K., Cohen, N. R. & Alexander, H. EUKulele: Taxonomic annotation of the unsung eukaryotic microbes. *arXiv [q-bio.PE]* (2020).
90. Queirós, P., Delogu, F., Hickl, O., May, P. & Wilmes, P. Mantis: flexible and consensus-driven genome annotation. *bioRxiv* (2020).
91. Zhou, Z. *et al.* METABOLIC: High-throughput profiling of microbial genomes for functional traits, biogeochemistry, and community-scale metabolic networks. *bioRxiv* 761643 (2020) doi:10.1101/761643.

92. McDaniel, E. A., Anantharaman, K. & McMahon, K. D. metabolisHMM: Phylogenomic analysis for exploration of microbial phylogenies and metabolic pathways. *bioRxiv* 2019.12.20.884627 (2019) doi:10.1101/2019.12.20.884627.
93. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
94. Weissman, J. L., Hou, S. & Fuhrman, J. A. Estimating maximal microbial growth rates from cultures, metagenomes, and single cells via codon usage patterns. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
95. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
96. Bushnell, B. *BBMap: A fast, accurate, splice-aware aligner*. <https://www.osti.gov/biblio/1241166> (2014).
97. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
98. West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C. & Banfield, J. F. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.* **28**, 569–580 (04 2018).
99. Alneberg, J. *et al.* CONCOCT: Clustering cONTigs on COverage and ComposiTion. *arXiv [q-bio.GN]* (2013).
100. Levy Karin, E., Mirdita, M. & Söding, J. MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* **8**, 48 (2020).
101. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically

- 903 annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids*
904 *Res.* **47**, D309–D314 (2019).
- 905 102. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and
906 Annotation Completeness. *Methods Mol. Biol.* **1962**, 227–245 (2019).
- 907 103. Saary, P., Mitchell, A. L. & Finn, R. D. Estimating the quality of eukaryotic genomes
908 recovered from metagenomic analysis with EukCC. *Genome Biol.* **21**, 244 (2020).
- 909 104. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment
910 search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
- 911 105. Inferring Correlation Networks from Genomic Survey Data.
912 <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002687>.
- 913 106. Kurtz, Z. D. *et al.* Sparse and compositionally robust inference of microbial ecological
914 networks. *PLoS Comput. Biol.* **11**, e1004226 (2015).
- 915 107. Csardi, G. & Nepusz, T. The igraph software package for complex network research.
916 *InterJournal, Complex Systems* **1695**, 1–9 (2006).
- 917 108. Dormann, C. F., Frund, J., Bluthgen, N. & Gruber, B. Indices, graphs and null models:
918 Analyzing bipartite ecological networks. *Open Ecol. J.* **2**, 7–24 (2009).
- 919 109. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated
920 non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*
921 **35**, D61–5 (2007).
- 922 110. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high
923 throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
- 924 111. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood
925 trees for large alignments. *PLoS One* **5**, e9490 (2010).

112. Lee, M. D. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics* **35**, 4162–4164 (2019).
113. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
114. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
115. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
116. Tange, O. *GNU Parallel 2018*. (Lulu.com, 2018).
117. Team, R. C. & Others. R: A language and environment for statistical computing. (2013).
118. Kahle, D. & Wickham, H. Ggmap: Spatial visualization with ggplot2. *R J.* **5**, 144 (2013).
119. Graham, E. D., Heidelberg, J. F. & Tully, B. J. Potential for primary productivity in a globally-distributed bacterial phototroph. *ISME J.* **12**, 1861–1866 (2018).
120. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, (2014).
121. kevinblighe/EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. <https://github.com/kevinblighe/EnhancedVolcano>.
122. Wickham, H. ggplot2: ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* **3**, 180–185 (2011).
123. Varrette, S., Bouvry, P., Cartiaux, H. & Georgatos, F. Management of an academic HPC cluster: The UL experience. in *2014 International Conference on High Performance Computing Simulation (HPCS)* 959–967 (2014).

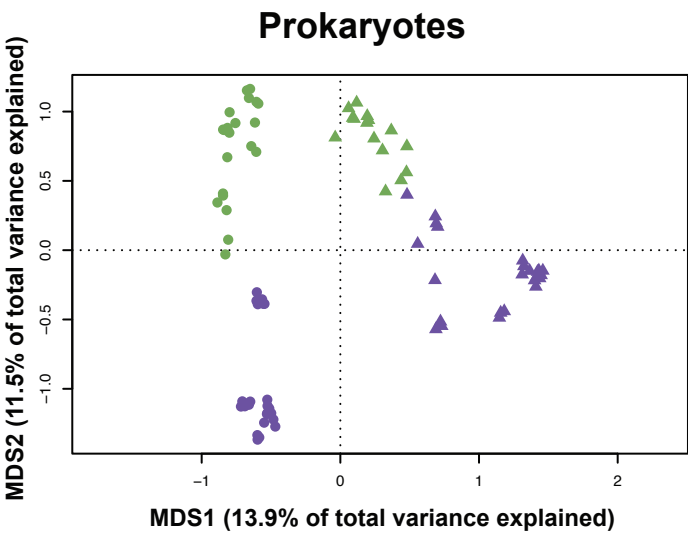
Fig.1. Sedimentary habitats affect microbiome structure and assembly

a

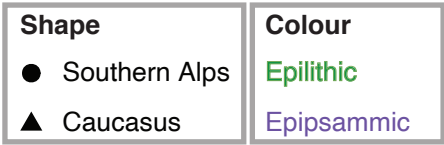
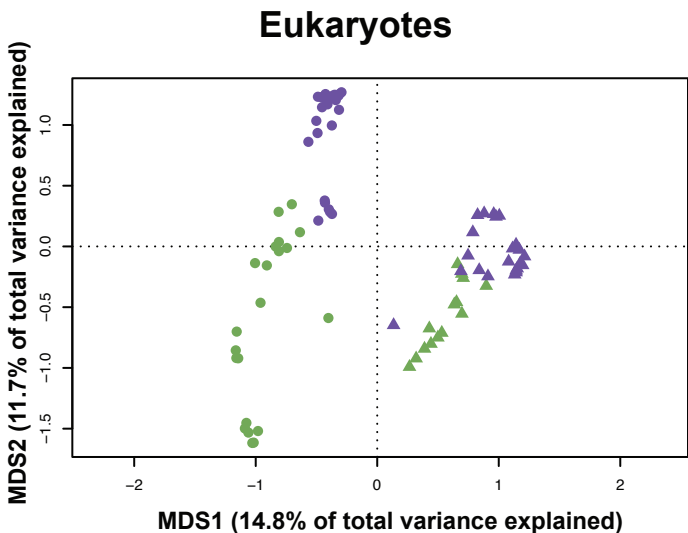
bioRxiv preprint doi: <https://doi.org/10.1101/2021.10.07.463499>; this version posted October 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



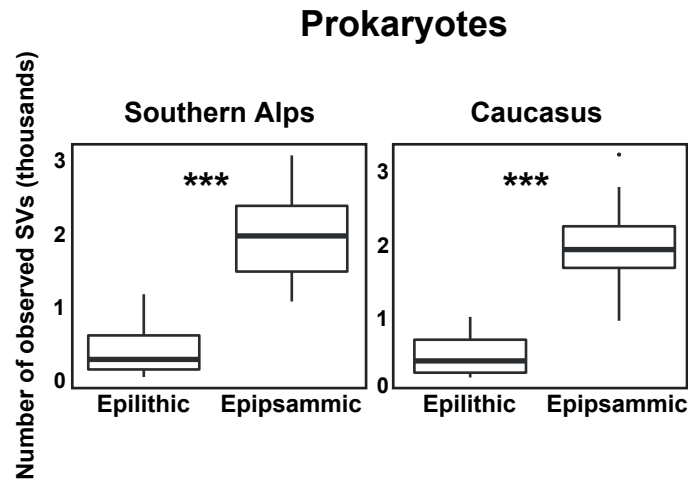
b



c



d



e

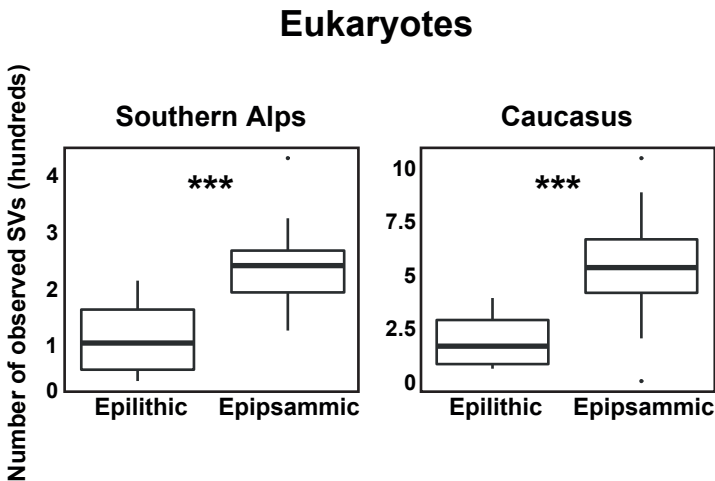
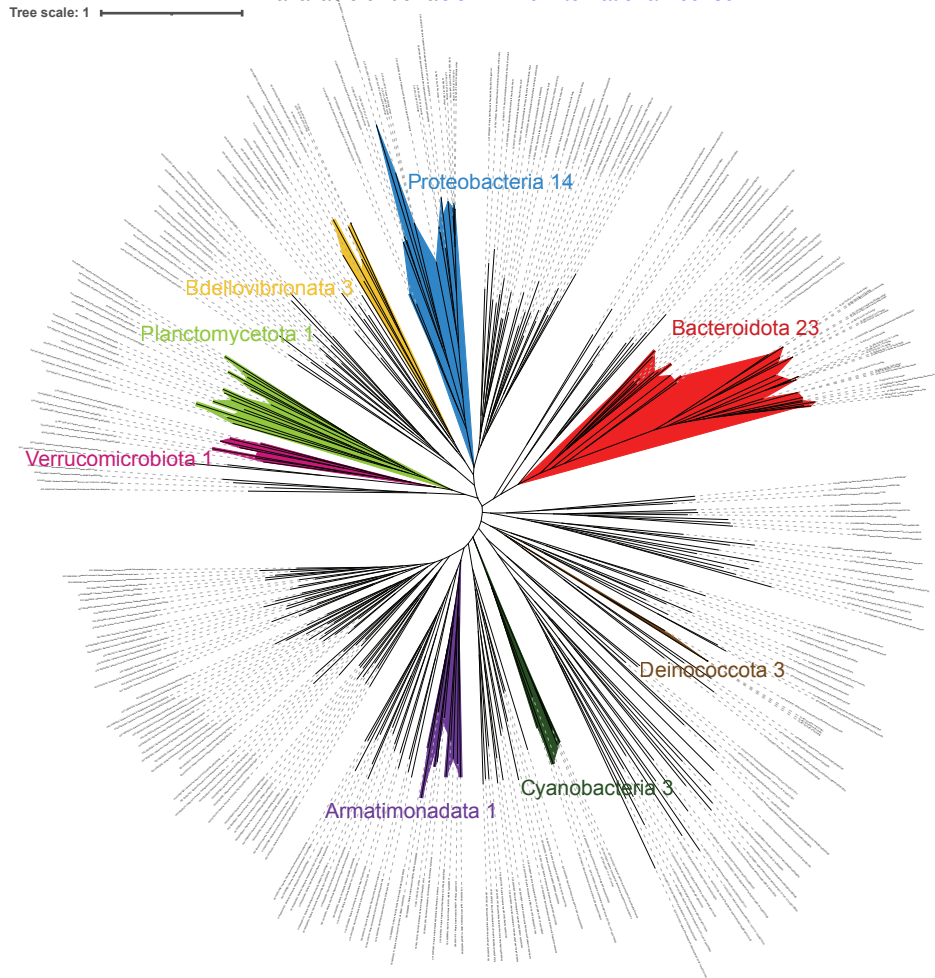


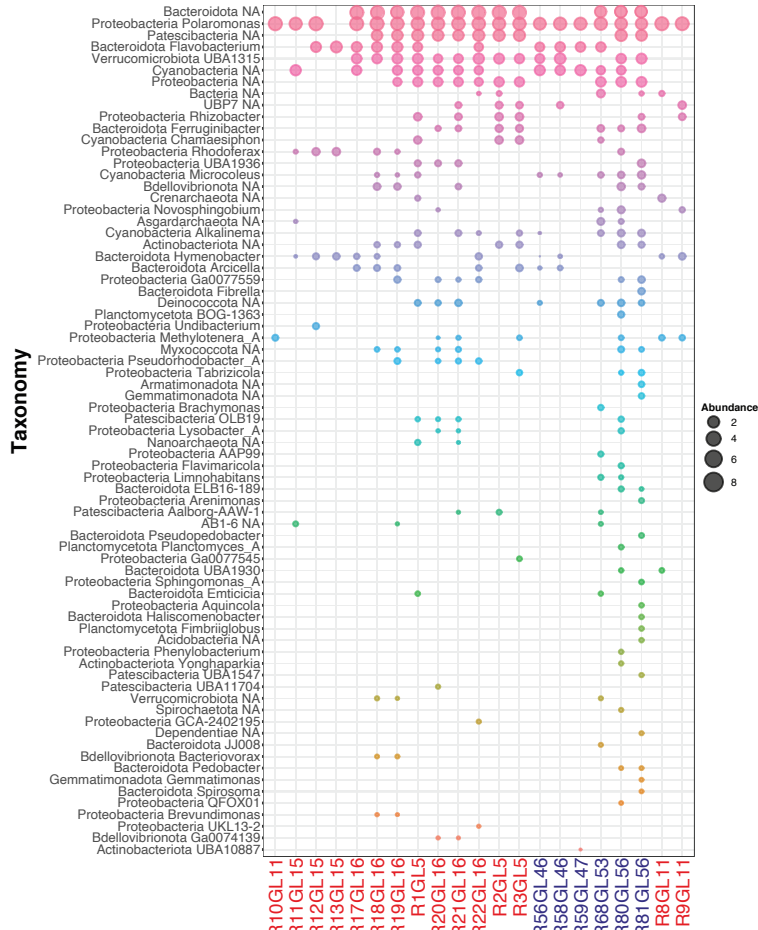
Fig.2. Metagenomics unveils the complexity of epilithic biofilms

bioRxiv preprint doi: <https://doi.org/10.1101/2021.10.07.463499>; this version posted October 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

a



b



c

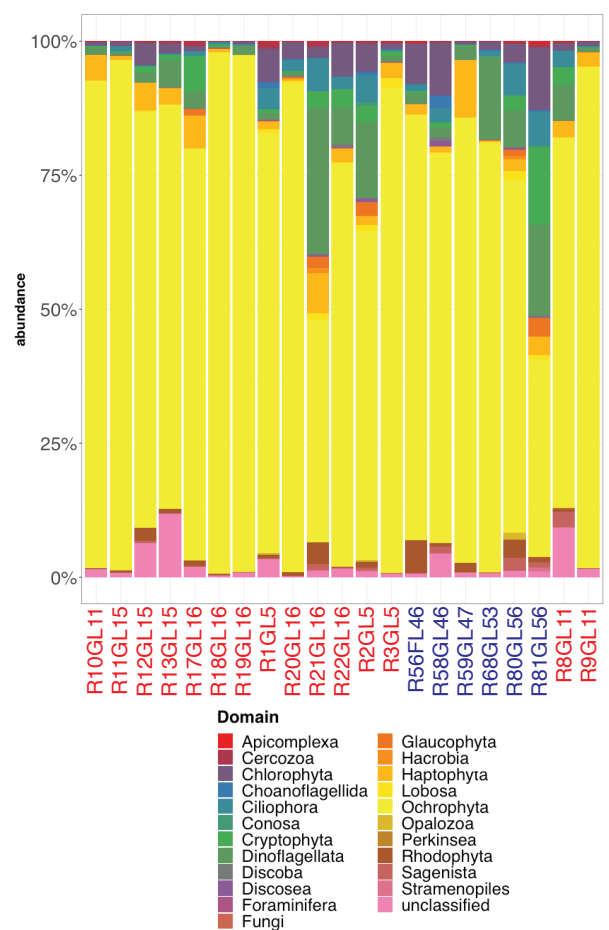


Fig.3. Epilithic biofilms are the basis for a ‘green food chain’

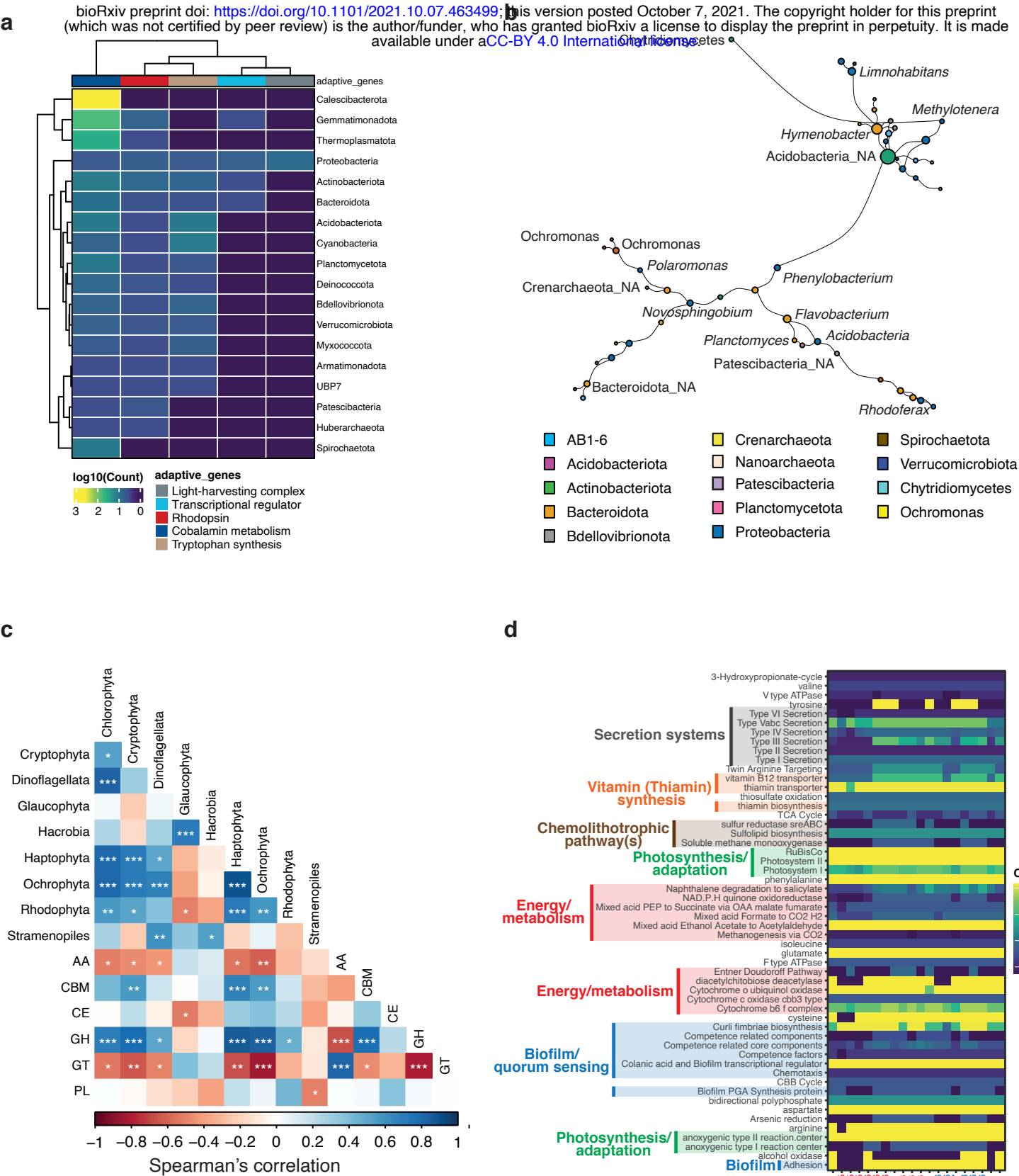
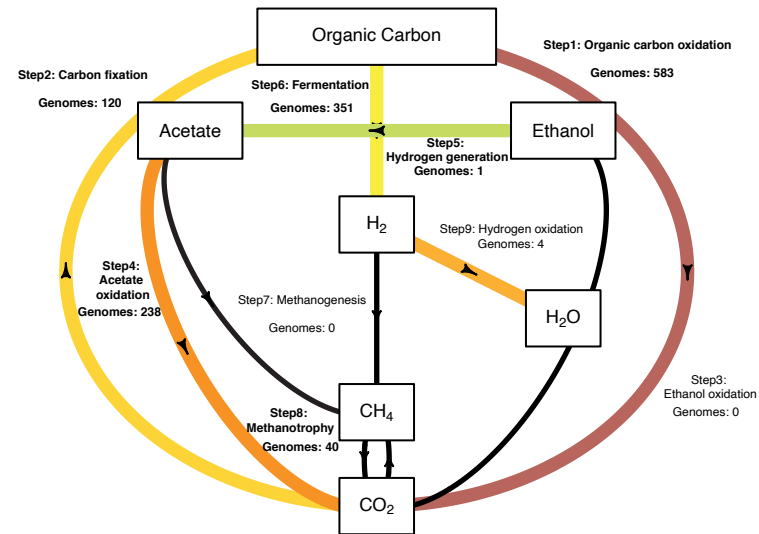
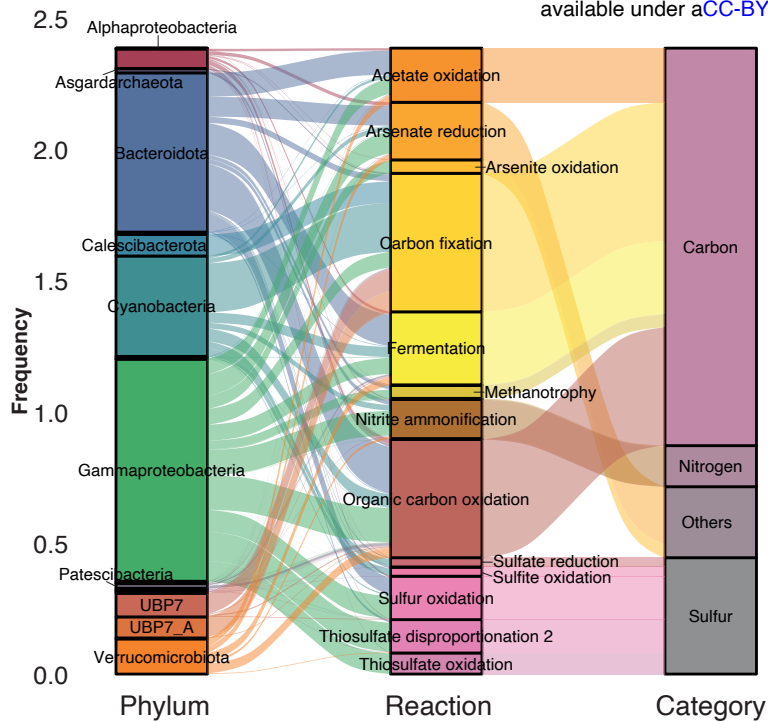


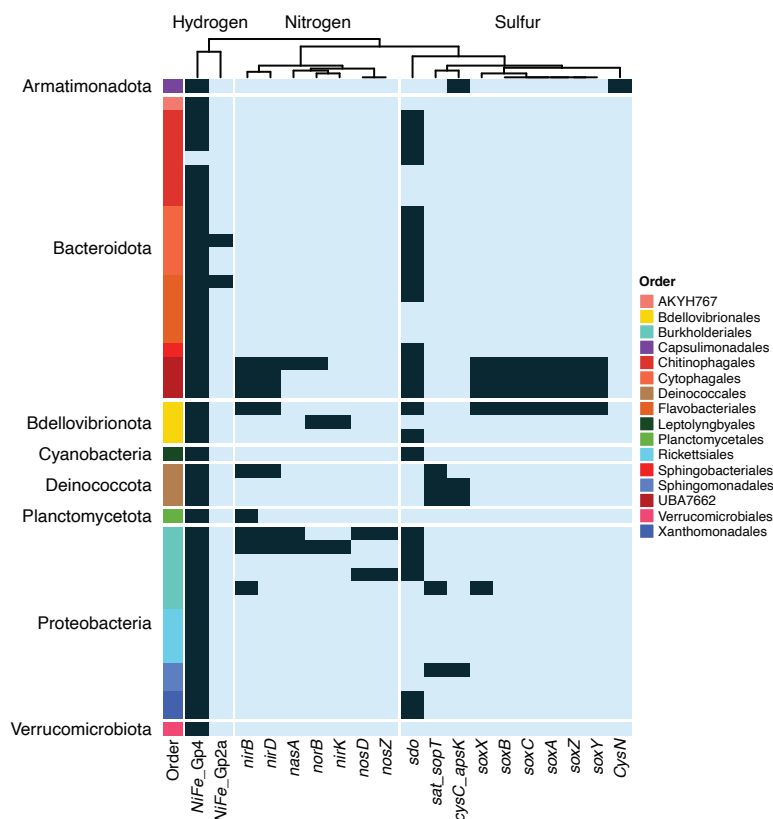
Fig.4. Functional redundancies across MAGs enable diverse energy acquisition and biogeochemical pathways

a

bioRxiv preprint doi: <https://doi.org/10.1101/2021.10.07.463499>; this version posted October 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



C



d

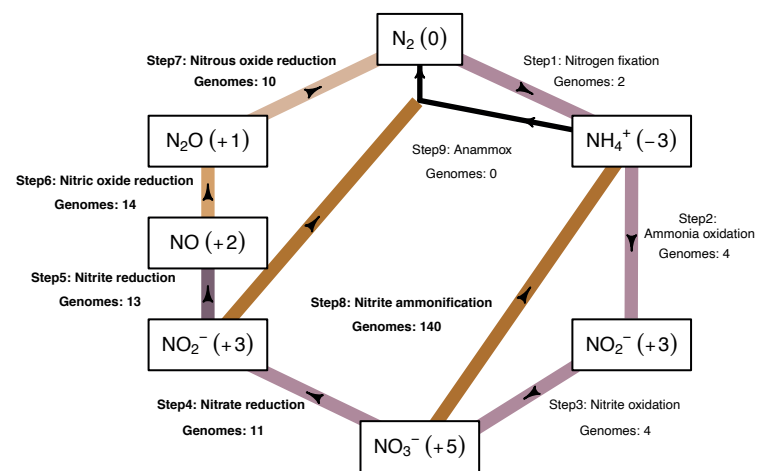
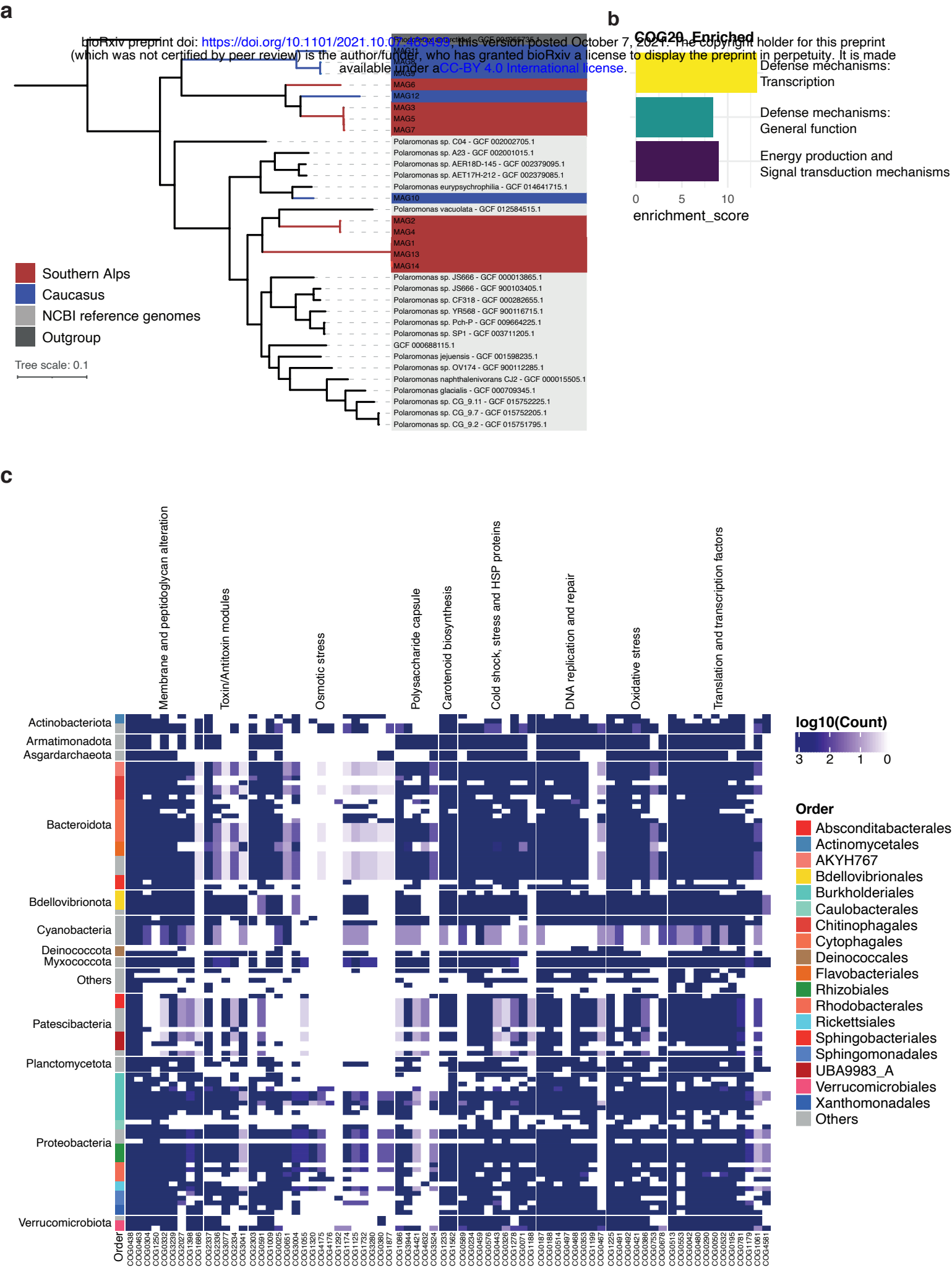


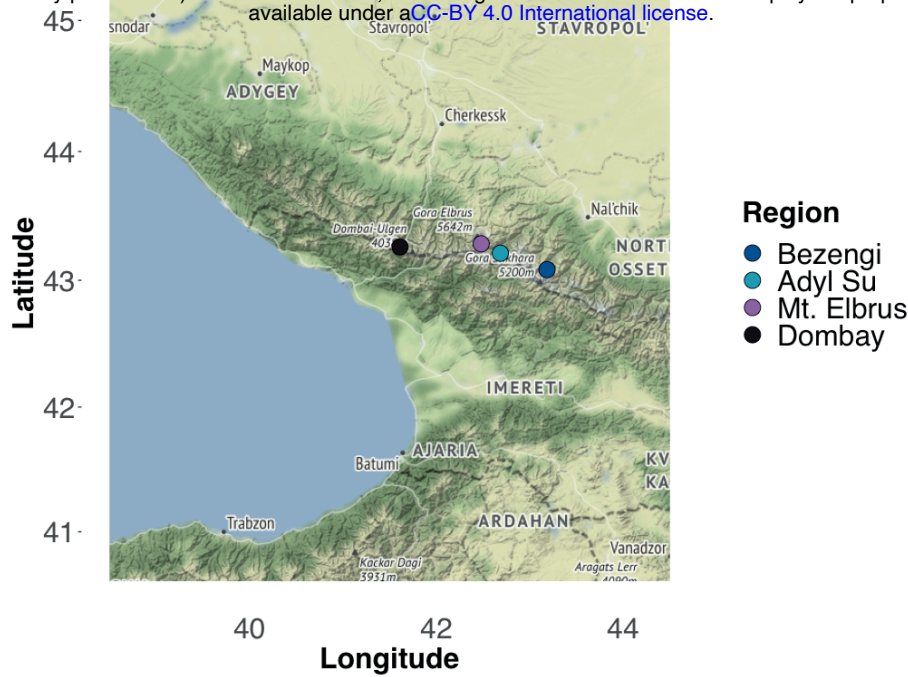
Fig.5. Genomic underpinnings of adaptation to the extreme GFS environment



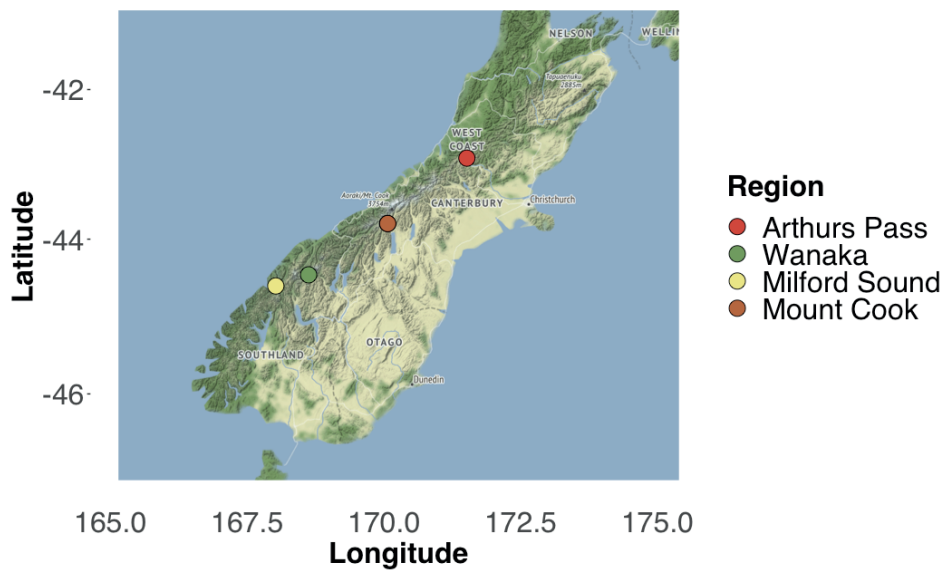
Supplementary figure 1. Sediment and epilithic biofilm sites

a

bioRxiv preprint doi: <https://doi.org/10.1101/2021.10.07.463499>; this version posted October 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

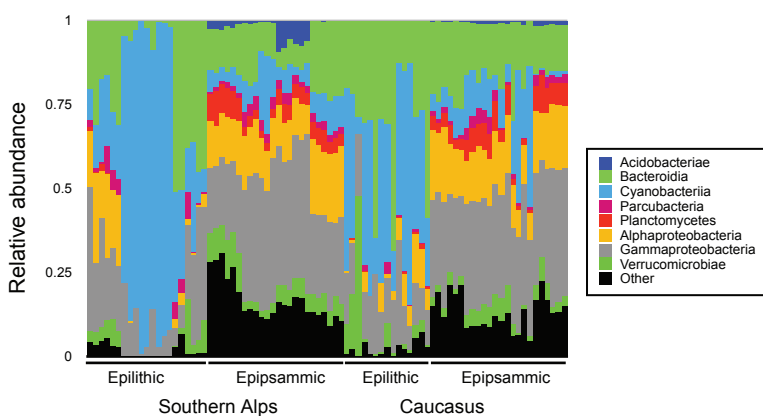


b



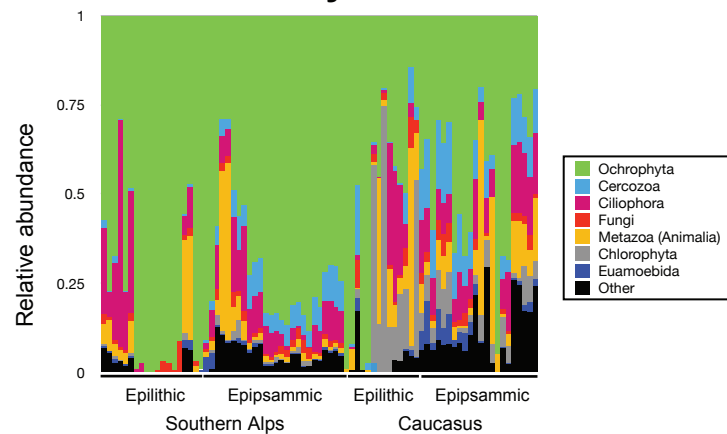
c

Prokaryotes



d

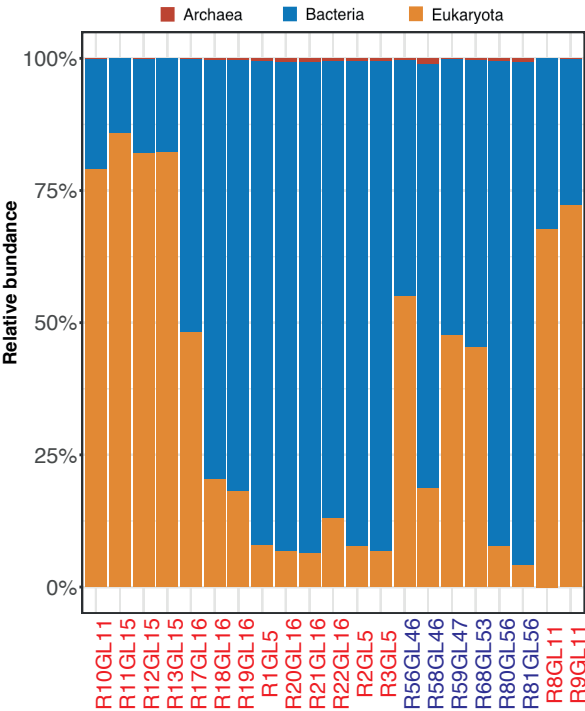
Eukaryotes



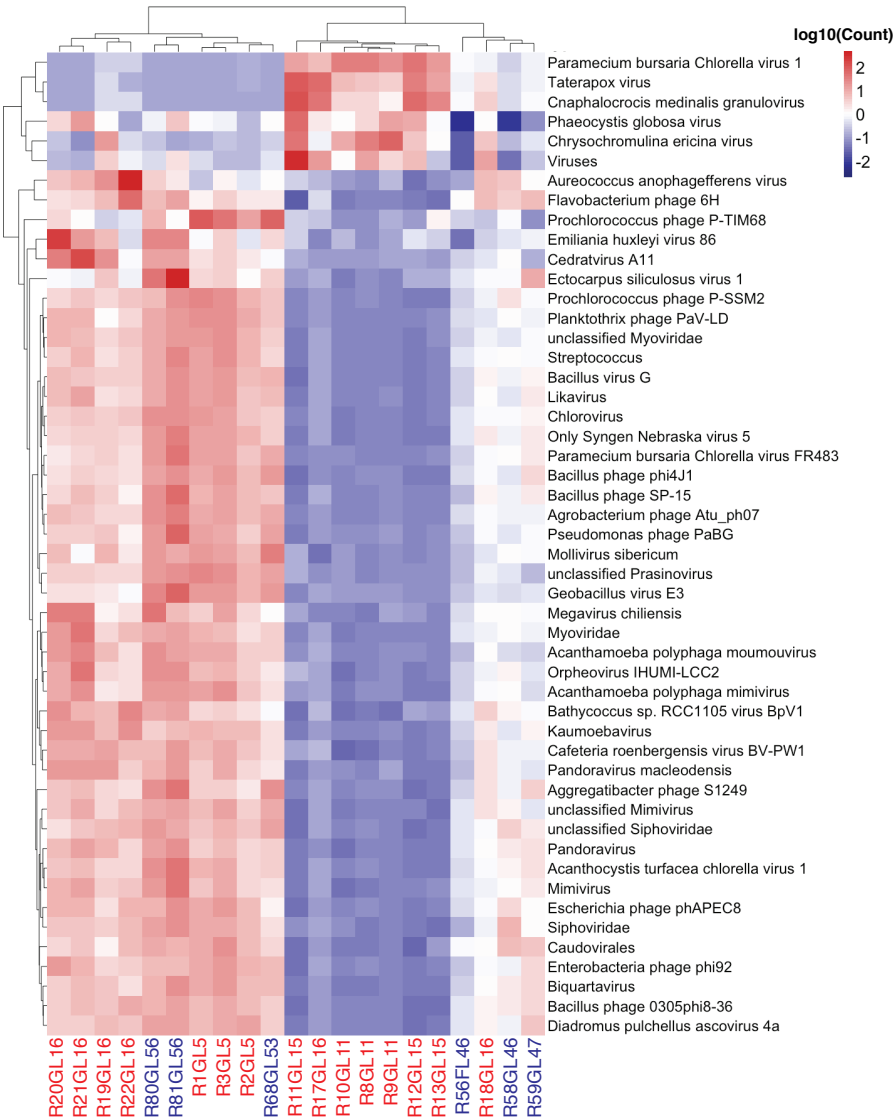
Supplementary figure 2. Epilithic biofilm metagenomic profiles

bioRxiv preprint doi: <https://doi.org/10.1101/2021.10.07.463499>; this version posted October 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

a



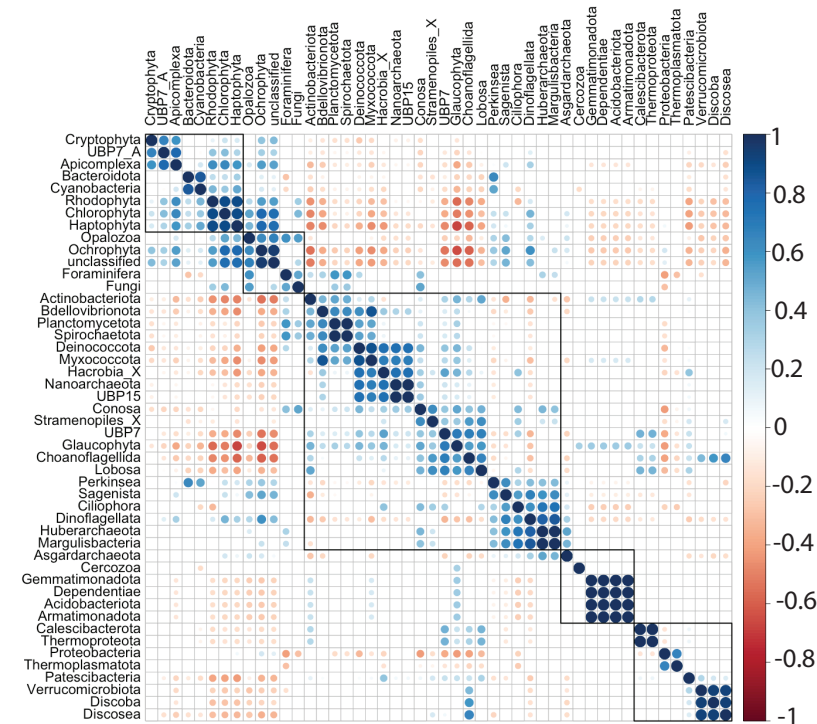
b



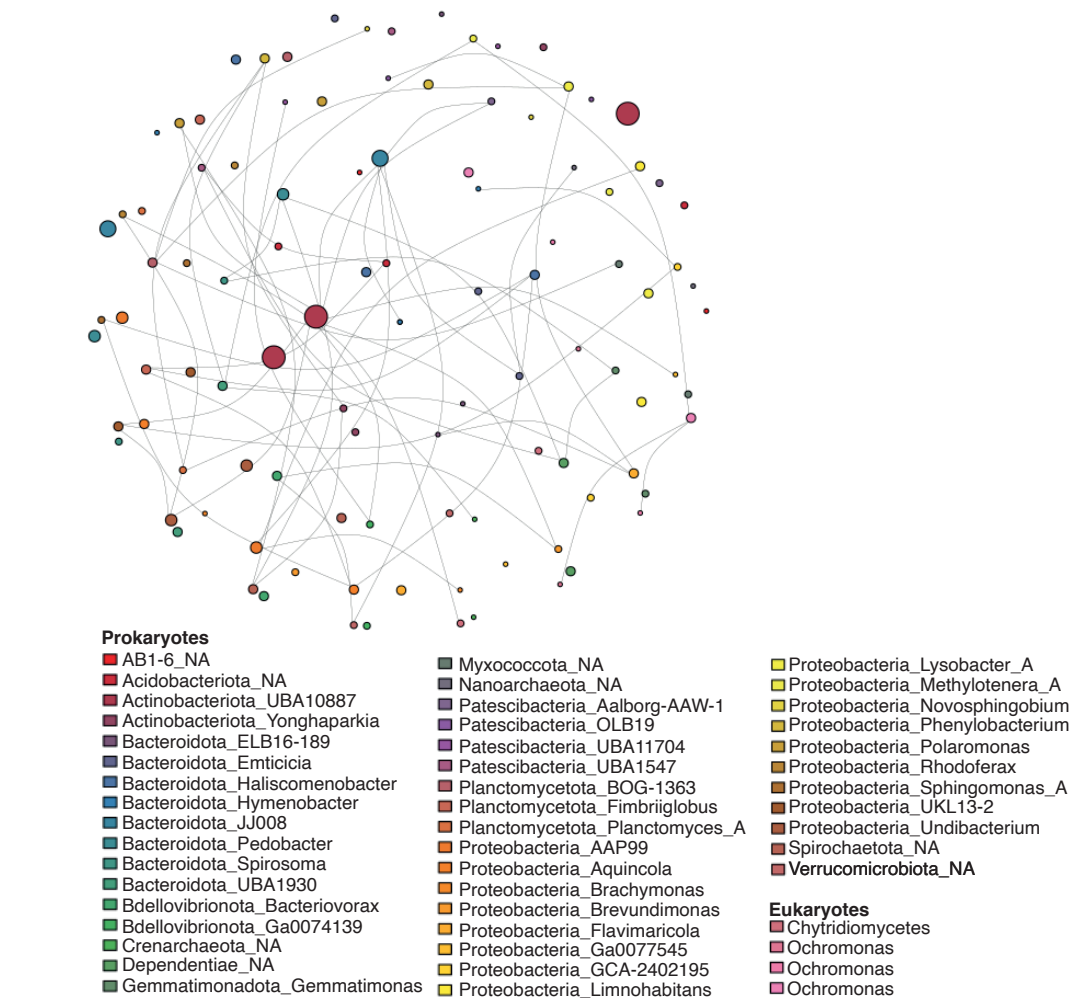
Supplementary figure 3. Cross-domain interactions and adaptations of epilithic biofilms

bioRxiv preprint doi: <https://doi.org/10.1101/2021.10.07.463499>; this version posted October 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

a

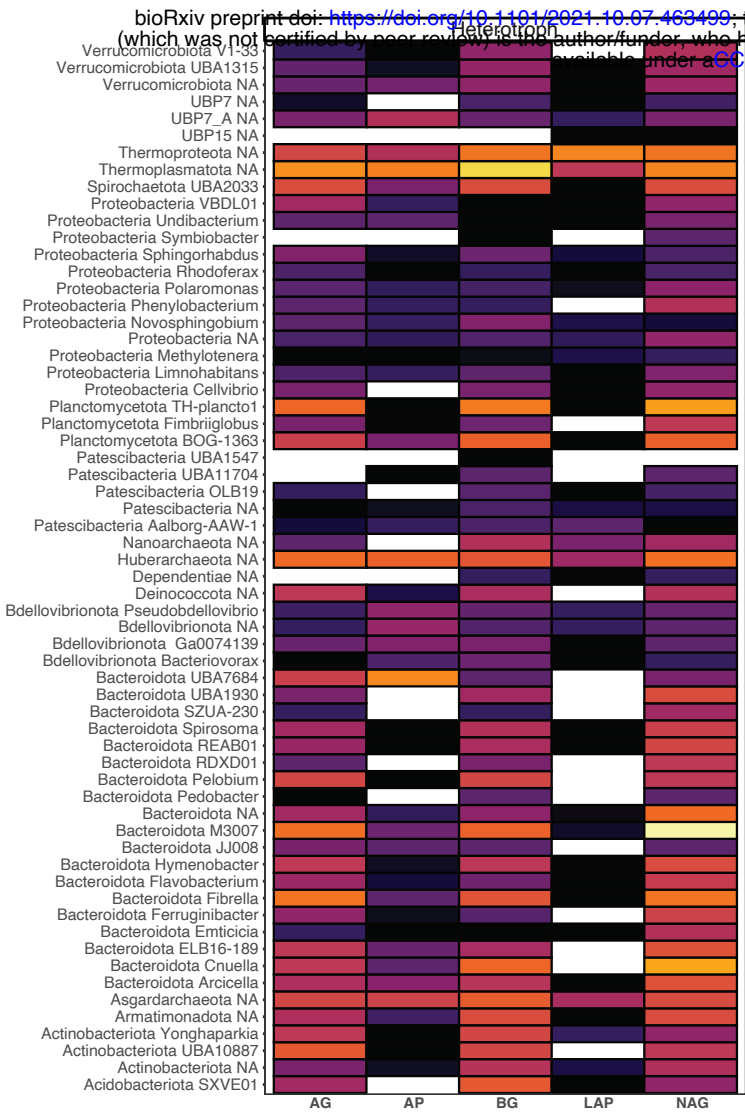


b

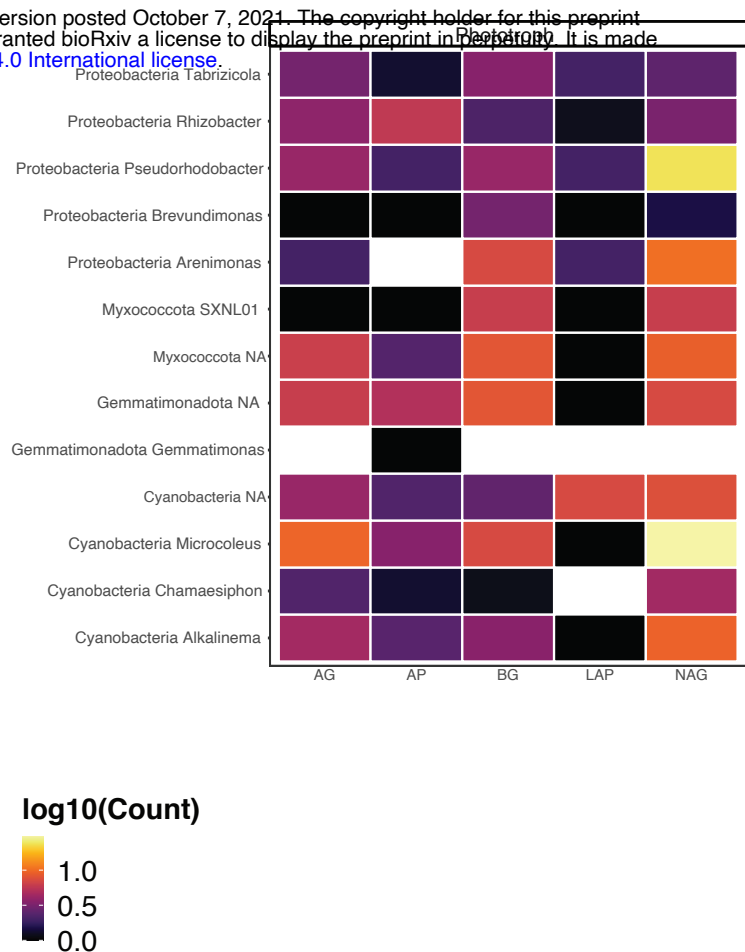


Supplementary figure 4. Extracellular enzyme genes based on lifestyle

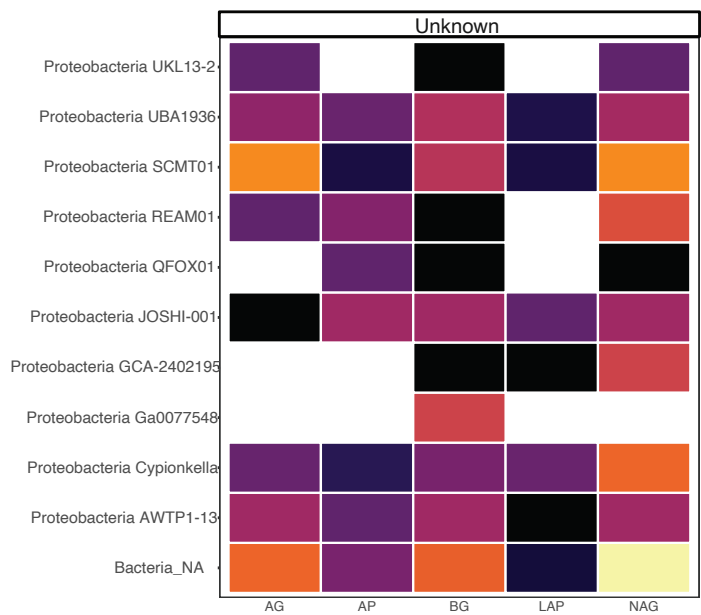
a



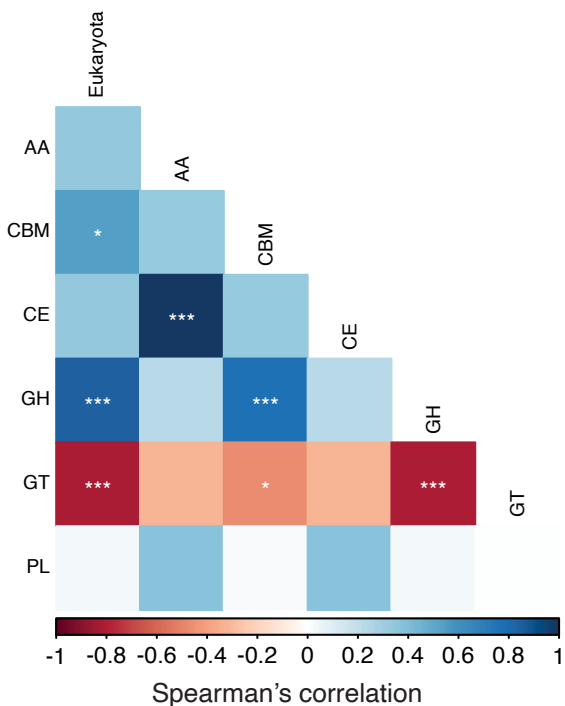
b



c

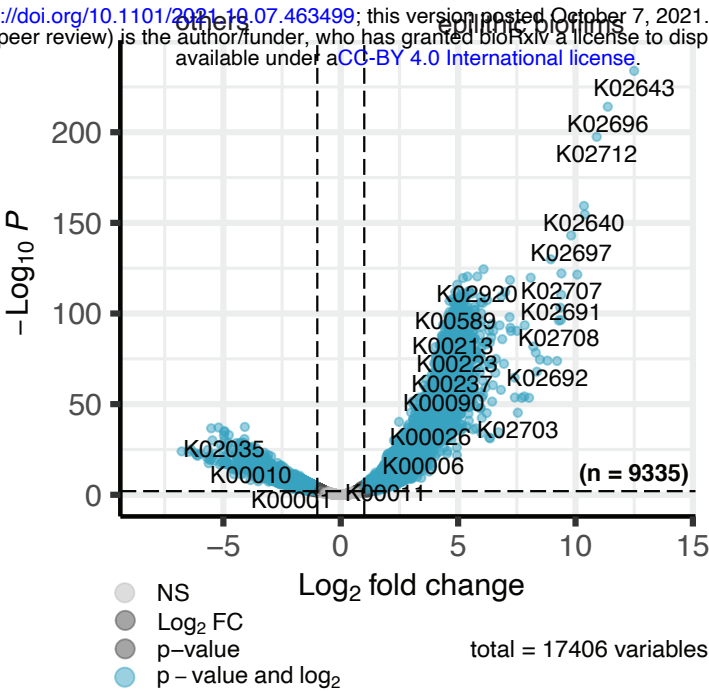


d



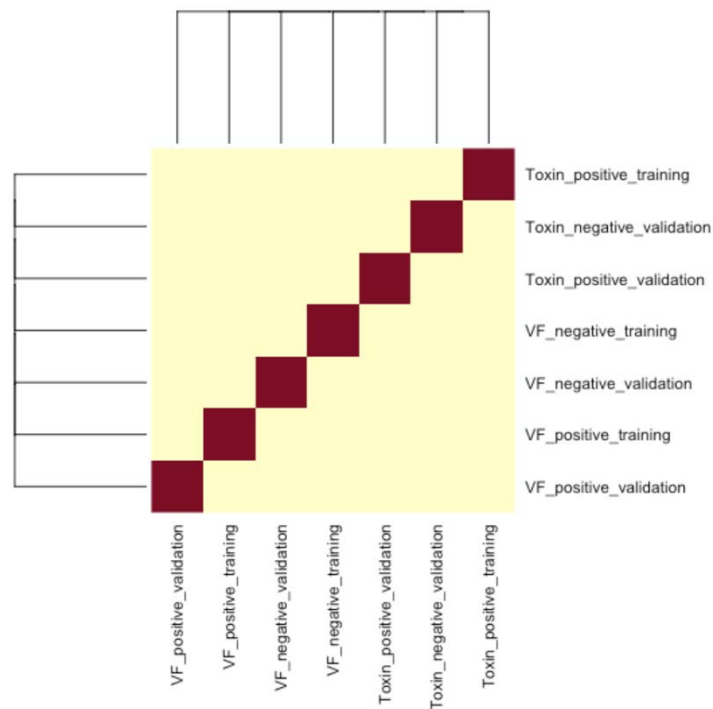
Supplementary figure 5. Comparison to public metagenomes reveals differential gene abundances

bioRxiv preprint doi: <https://doi.org/10.1101/2021.10.07.463499>; this version posted October 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

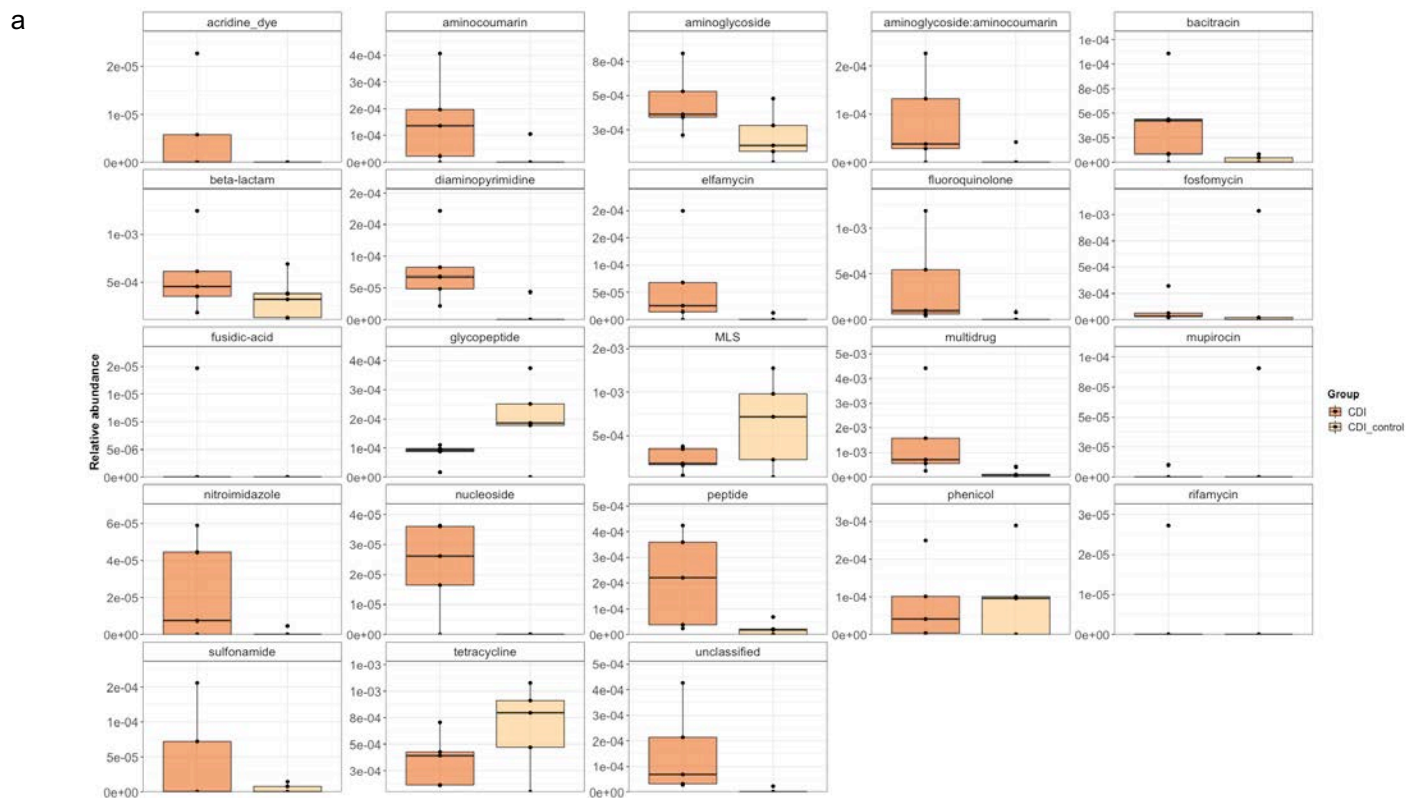


Appendix B. Supplementary Figures

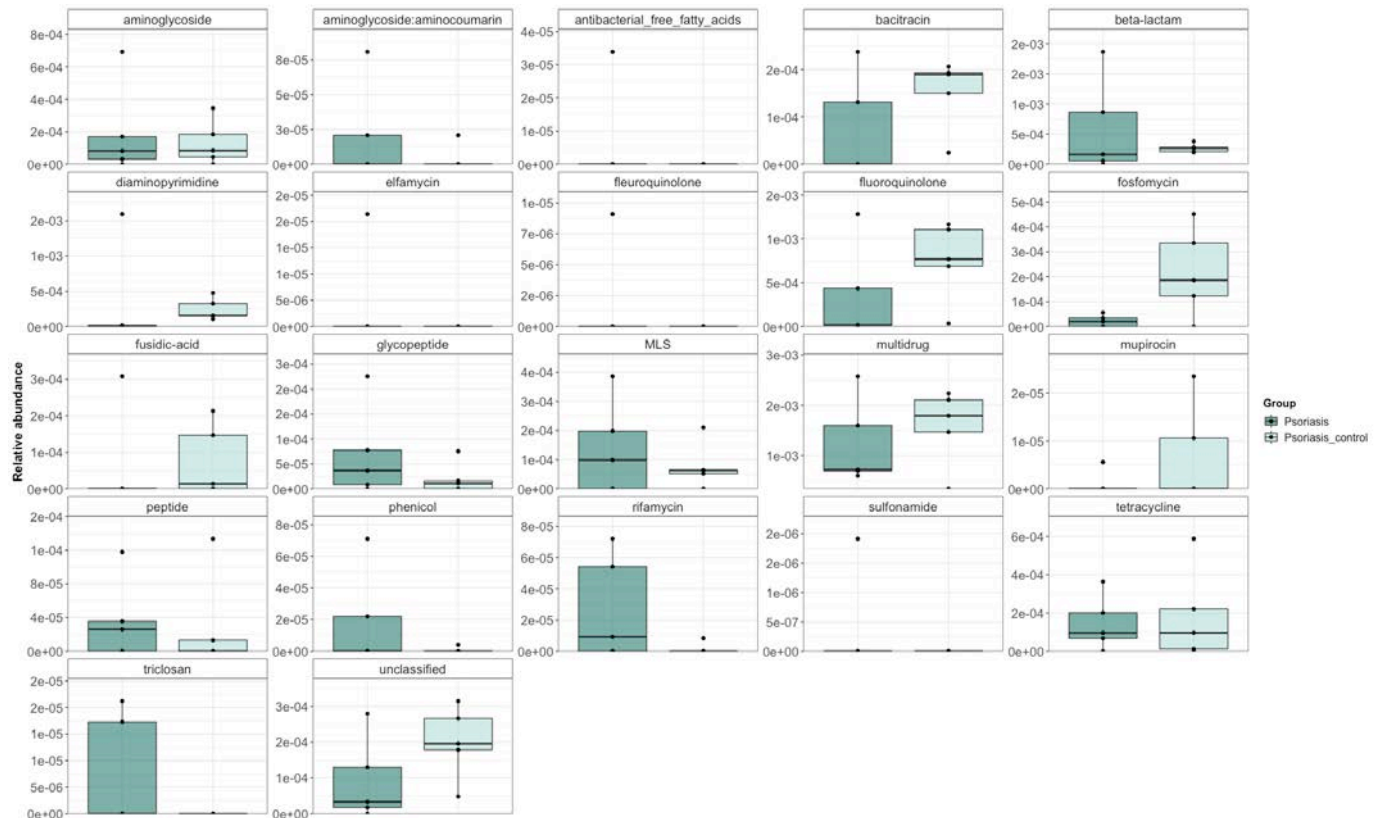
Appendix B.1



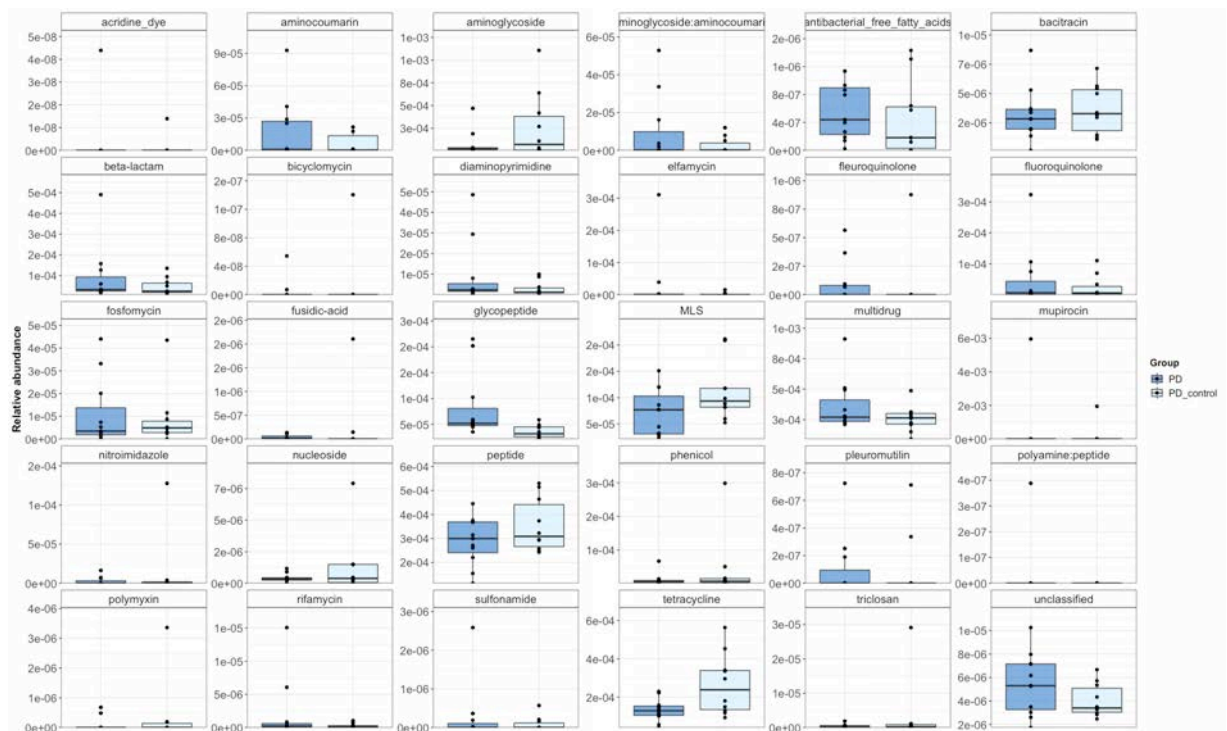
Supplementary figure 2.1: Sequence similarity comparison between validation and training datasets.



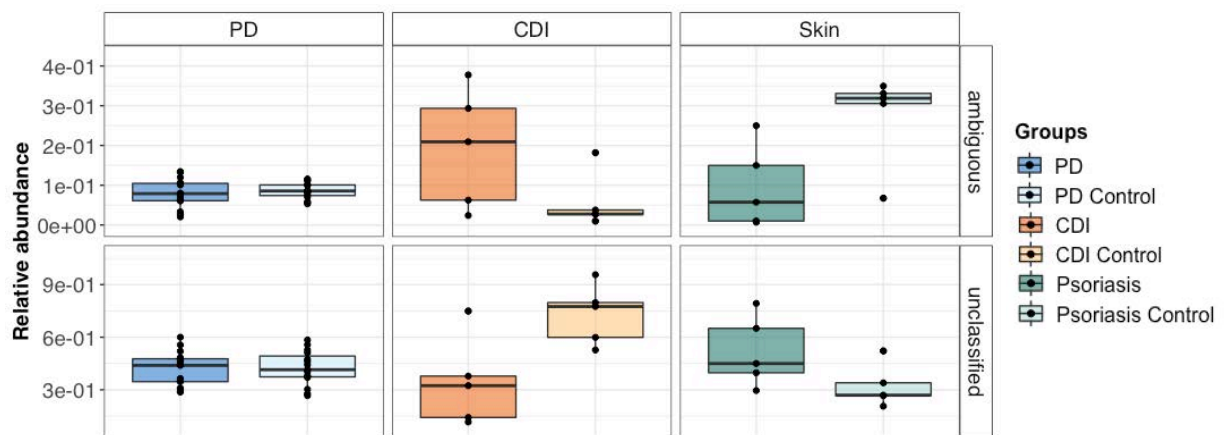
b



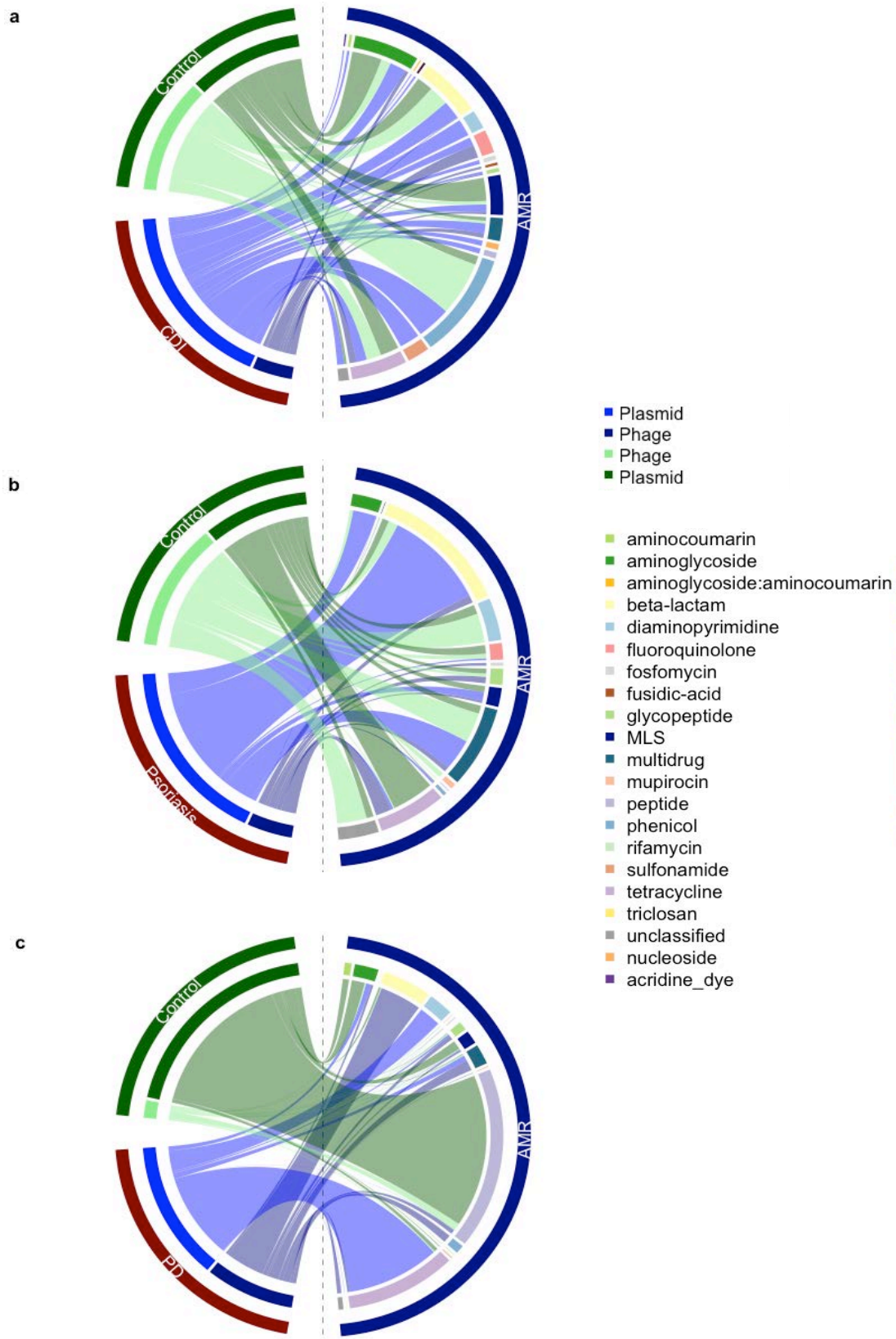
c



Supplementary figure 2.2: Relative abundance of antimicrobial resistance categories in three case-control metagenomic datasets. Relative abundance (%) of all identified resistance categories **a.** 23 antimicrobial resistance categories within *Clostridioides difficile* infection **b.** 22 antimicrobial resistance categories within the skin metagenome (psoriasis) and **c.** 30 antimicrobial resistance categories within the Parkinson's disease study.

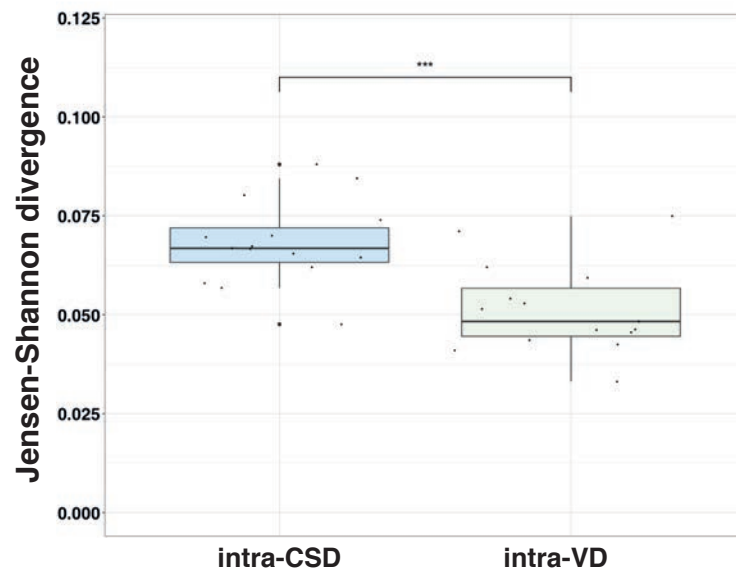


Supplementary figure 2.3: Distribution of virulence factors, including bacterial toxins, and AMR over unclassified and ambiguous (predicted to be both plasmid and phage or phage and chromosome)

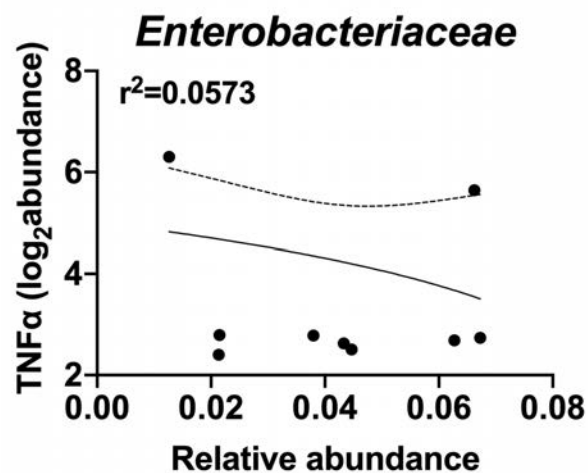


Supplementary figure 2.4: The prevalence of different resistance categories within the MGEs. a. prevalence of antimicrobial resistance genes within MGEs in *Clostridioides difficile* infection and control. b. psoriasis and control c. Parkinson's disease and control.

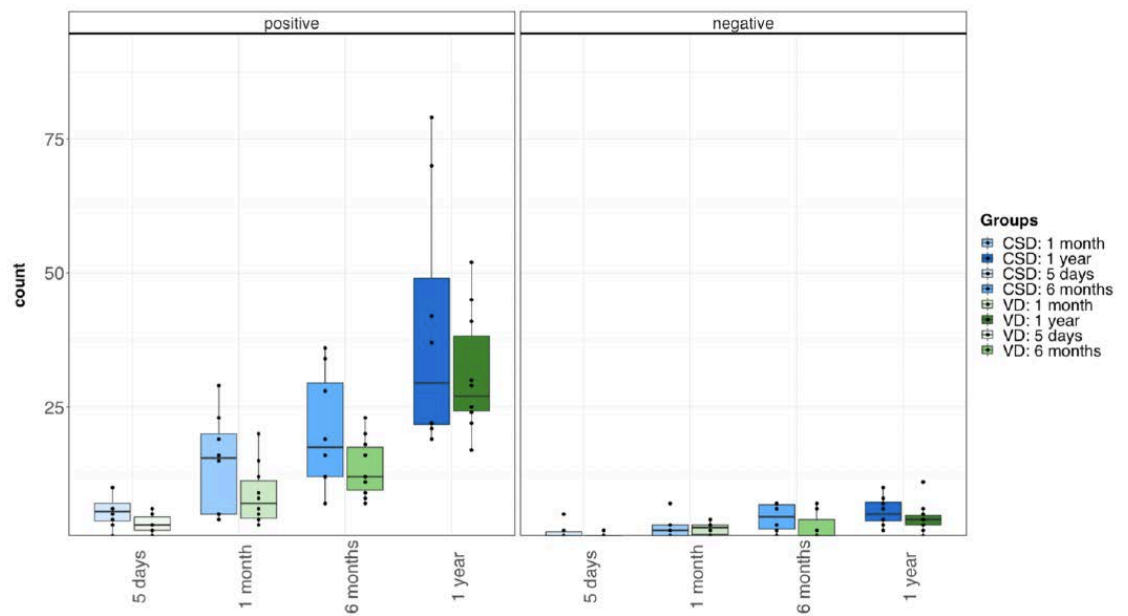
Appendix B.2



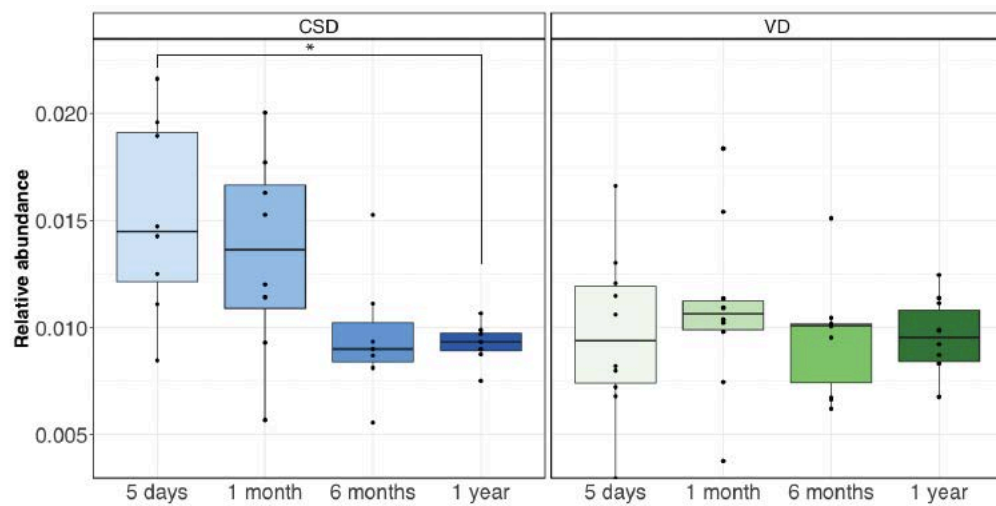
Supplementary figure 3.1: Intra- and inter-birth mode variability. Intra- and inter-birth mode Jensen-Shannon divergences of the mOTU profiles were calculated and tested to assess intra- and inter-group variability. Statistically significant differences were determined using the Wilcoxon-Mann-Whitney test; *** $p < 0.001$.



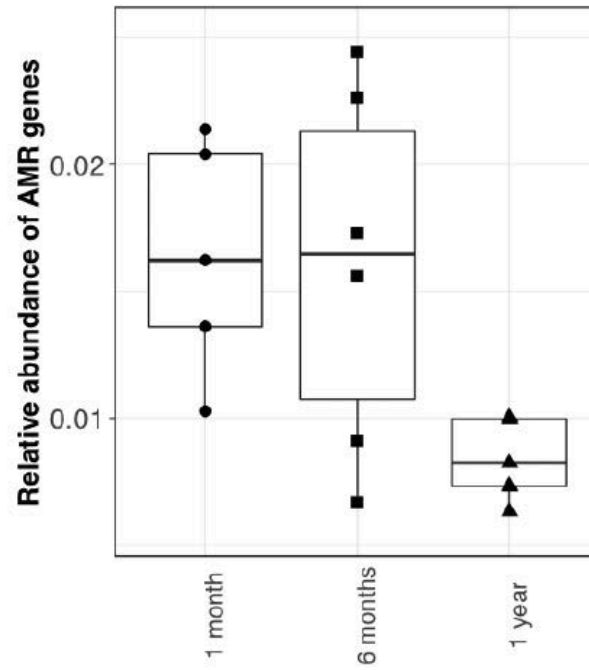
Supplementary figure 3.2: TNF-alpha correlation with Enterobacteriaceae. Log2 abundance of TNF-alpha levels were tested for correlation with the percent relative abundance of Enterobacteriaceae based on read counts mapping to the family. r^2 indicates the Spearman correlation coefficient.



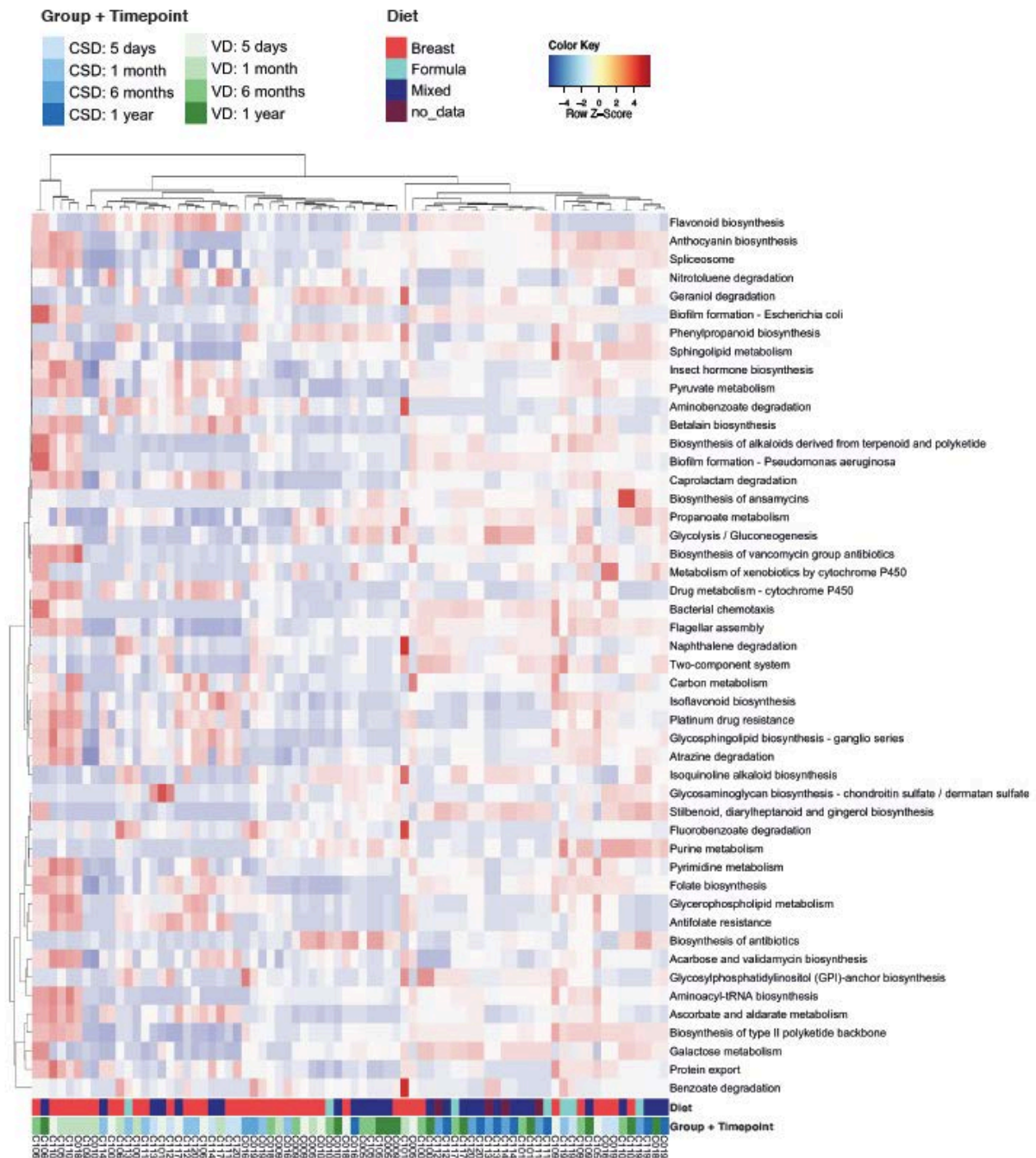
Supplementary figure 3.3: The number of Gram -ve and +ve organisms in CSD and VD.



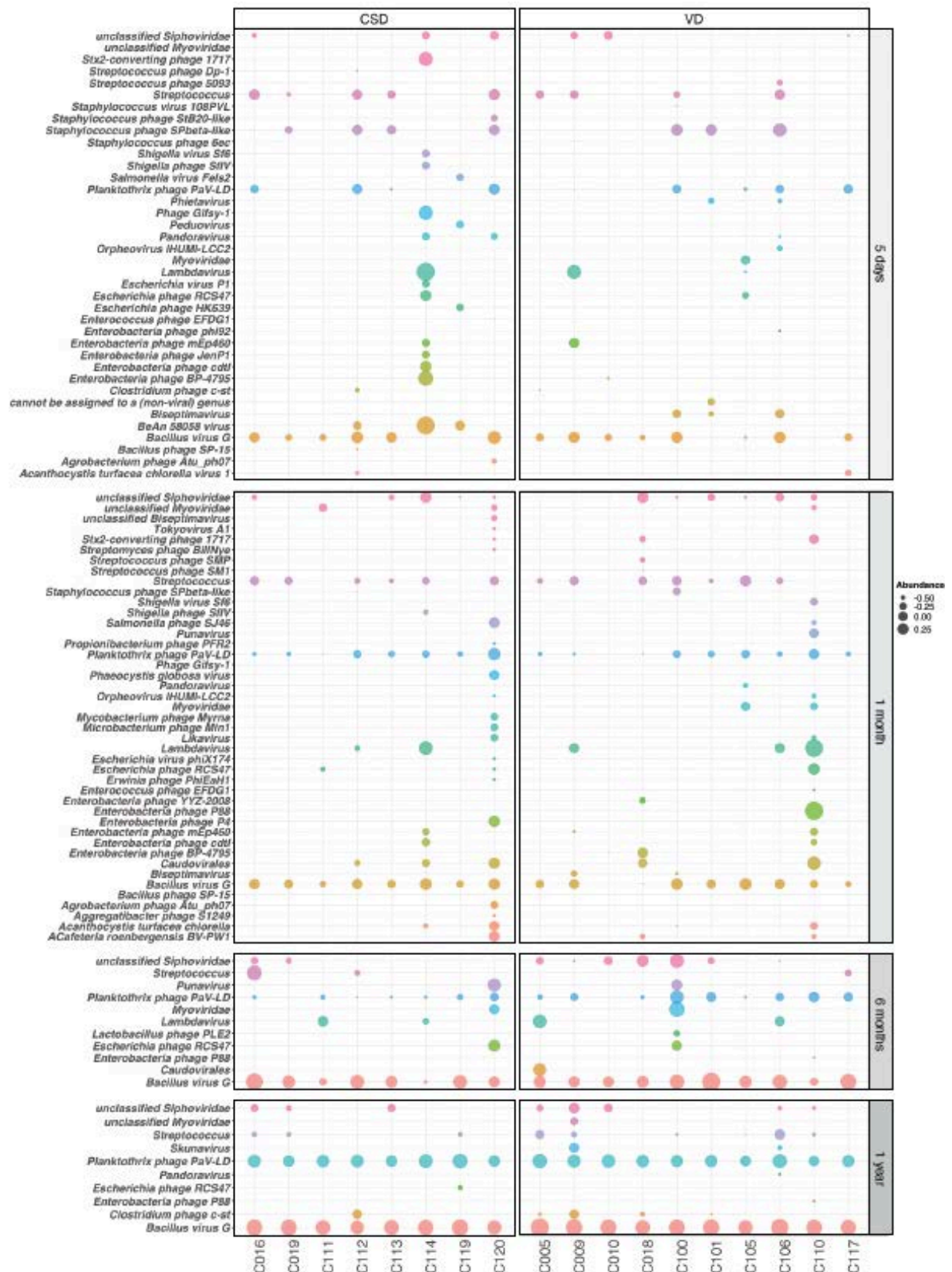
Supplementary figure 3.4: Overall AMR abundance at early and at the one year timepoint. Barplots depicting the differential AMR abundance between CSD and VD groups comparing the metagenomic from day five after birth with one year of age.



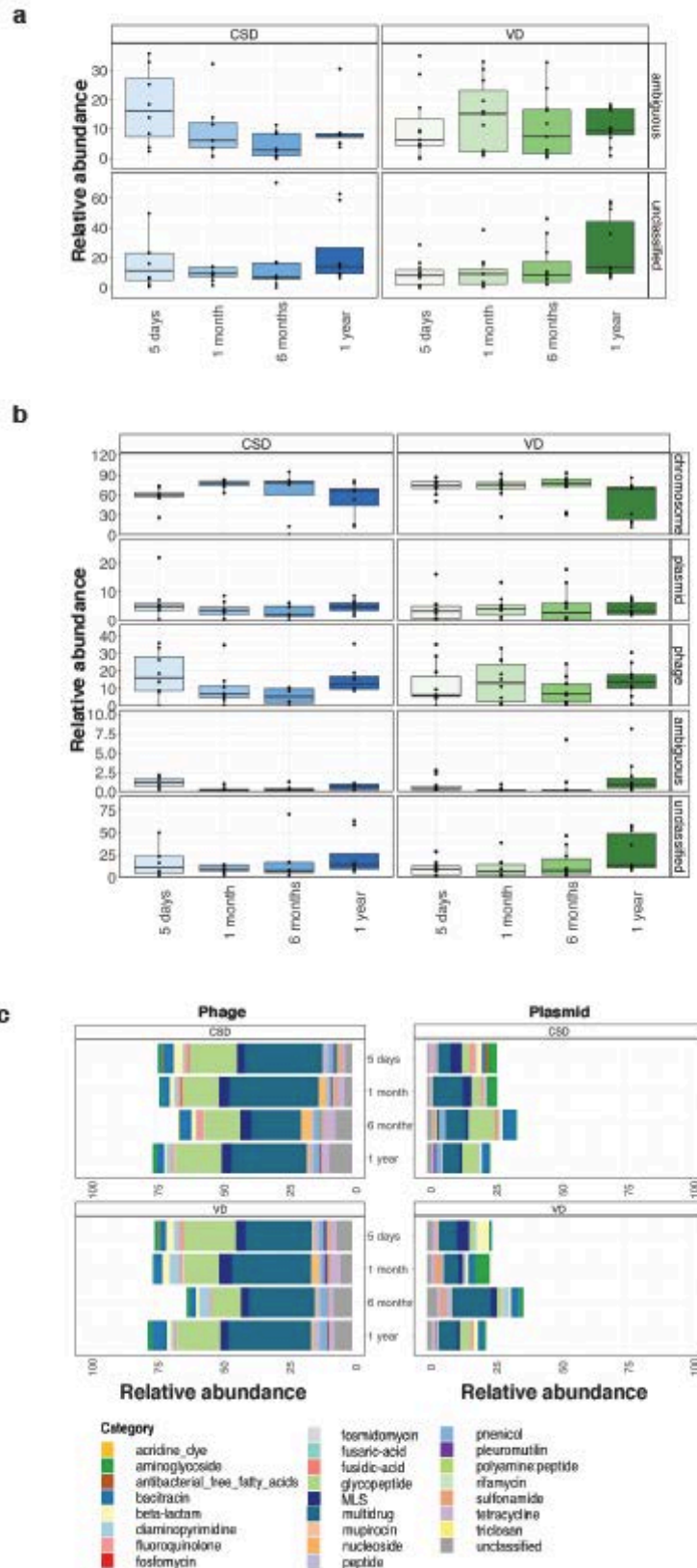
Supplementary figure 3.5: Overall AMR relative abundances observed in samples from the Gasparrini *et al.* study ranging from one month through to one year of age.



Supplementary figure 3.6: Functional pathways from day 5 after birth through to one year of age. Heatmap showing the relative abundances of the functional (KEGG) pathways through the first year of life. The plot includes annotations for different categories including group, timepoint and diet, i.e. breast-milk, formula, mixed or no data available.

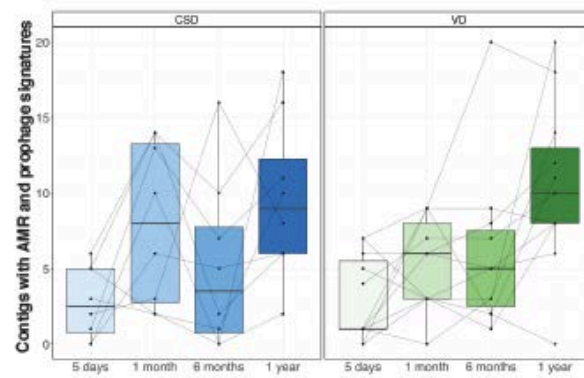


Supplementary figure 3.7: Longitudinal virome profiles. Bubble plots showing the log relative abundance of viruses identified within the CSD and VD samples through the first year of life. Only those with relative abundance greater than 0.01% are depicted. Significance was tested using a Two-way ANOVA, for a FDR-adjusted $p < 0.05$.

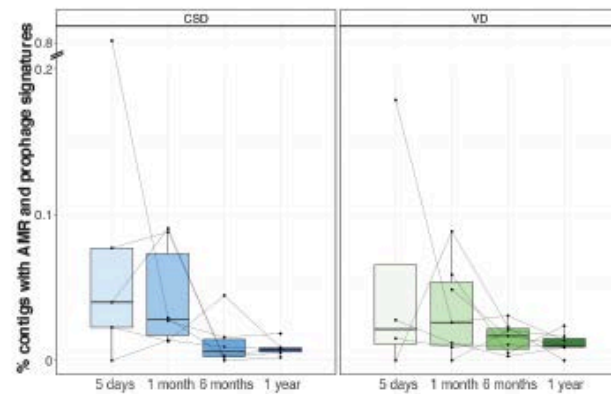


Supplementary figure 3.8: Abundances of bacterial chromosome and MGEs across time. **a.** Barplots depicting the relative abundance of ambiguous and unclassified sequences with respect to chromosomal or MGE classification. **b.** Distribution of chromosomal and MGE classification of AMR sequences at different time points for CSD and VD. Sequences assigned to both chromosome and bacteriophages were classified as belonging to bacteriophage under the assumption that the former represent likely prophages. **c.** The association of bacteriophage and plasmids with resistance categories at all timepoints comparing CSD and VD groups.

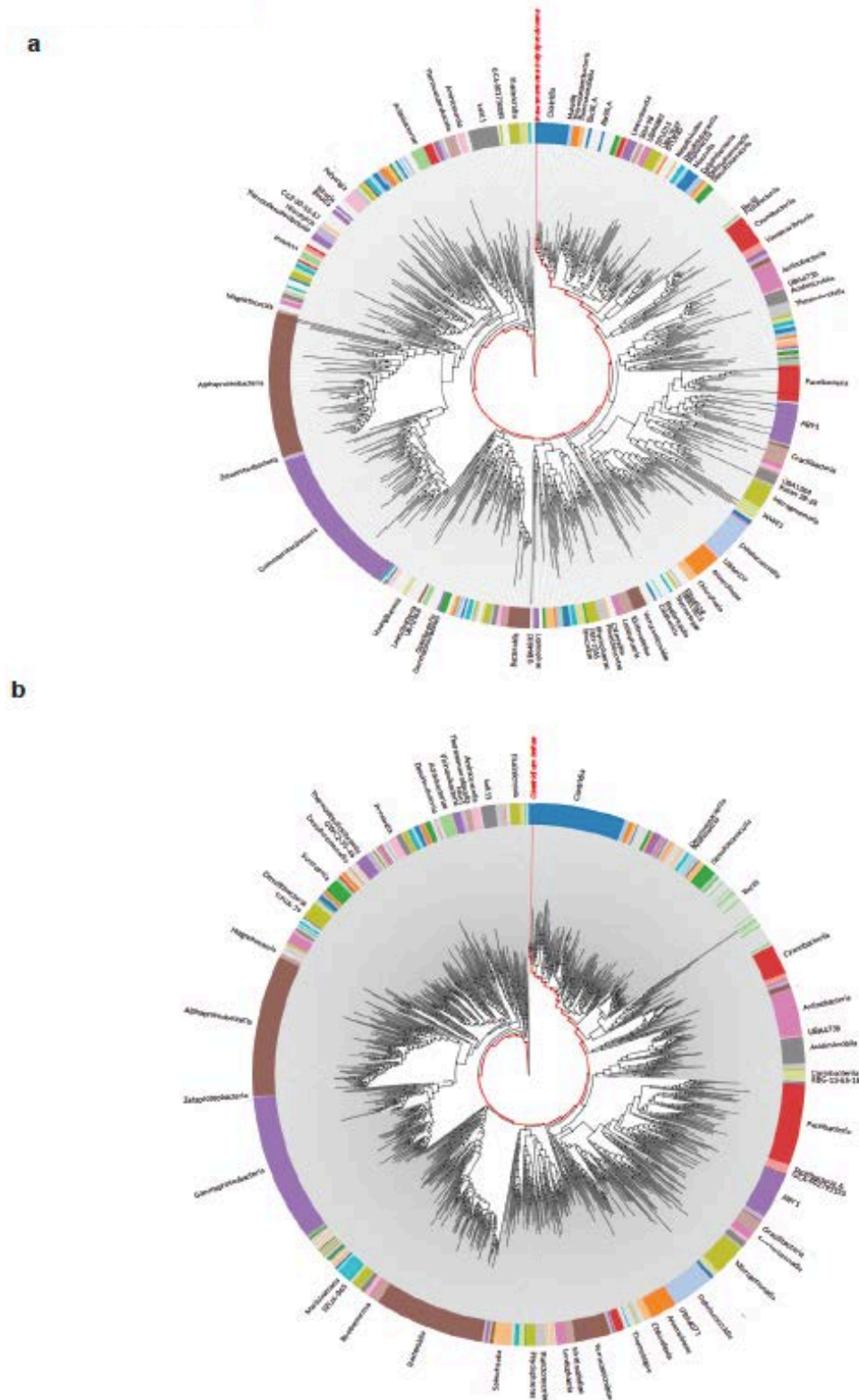
a



b

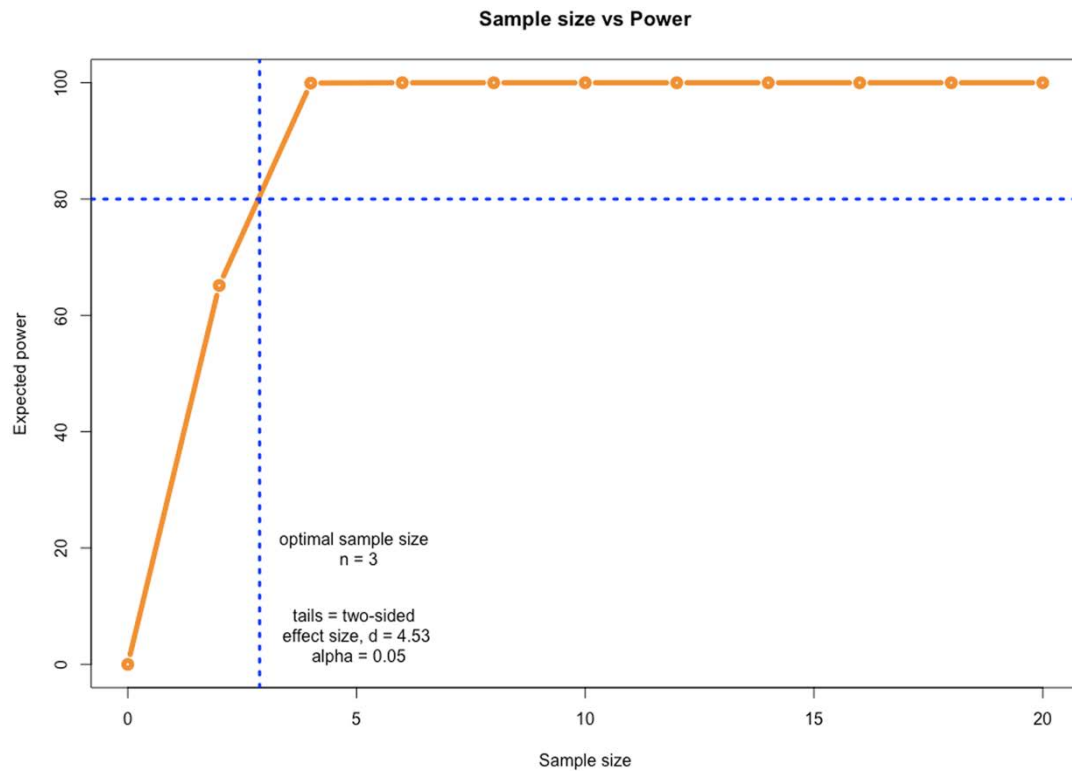


Supplementary figure 3.9: Abundance of AMR and prophage signature genes. **a.** Barplots depicting the abundance of contigs with AMR and prophage signature genes along the y-axis. The x-axis shows the longitudinal abundance of CSD and VD samples. **b.** Relative abundance of AMR and prophage signature genes in CSD and VD samples from day five after birth through to one year of age.

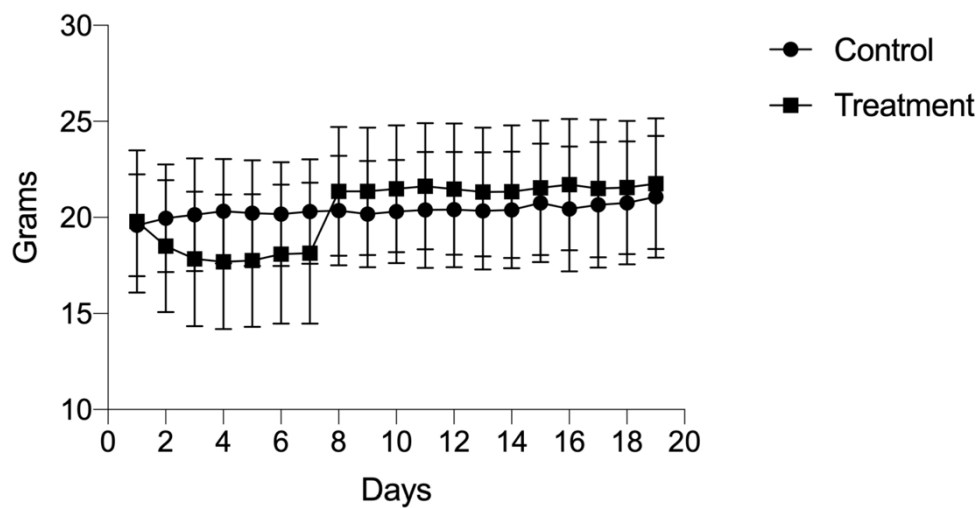


Supplementary figure 3.10: Strain resolution of recovered genomes. **a.** Neighbour-joining tree of the *Intestinimonas* genome involved in HGT recovered from the one month faecal sample of CSD sample C119. The tree was generated using ribosomal proteins in comparison to 1925 complete genomes obtained from the RefSeq database and visualised using the AnnoTree webserver¹⁰⁴. **b.** Taxa identified as belonging to the genus *Clostridium* were mapped against the RefSeq database and identified using the above approach.

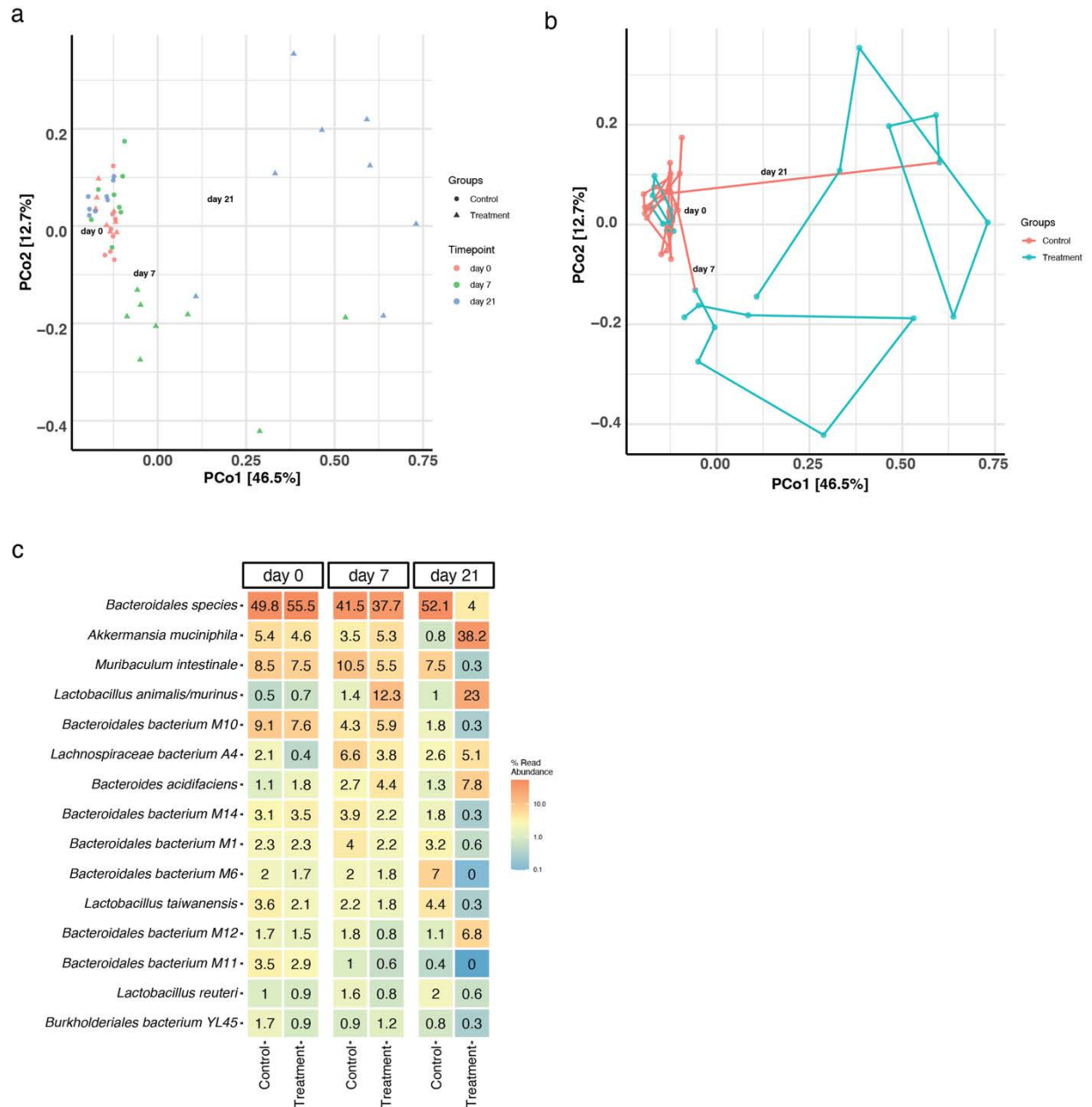
Appendix B.3



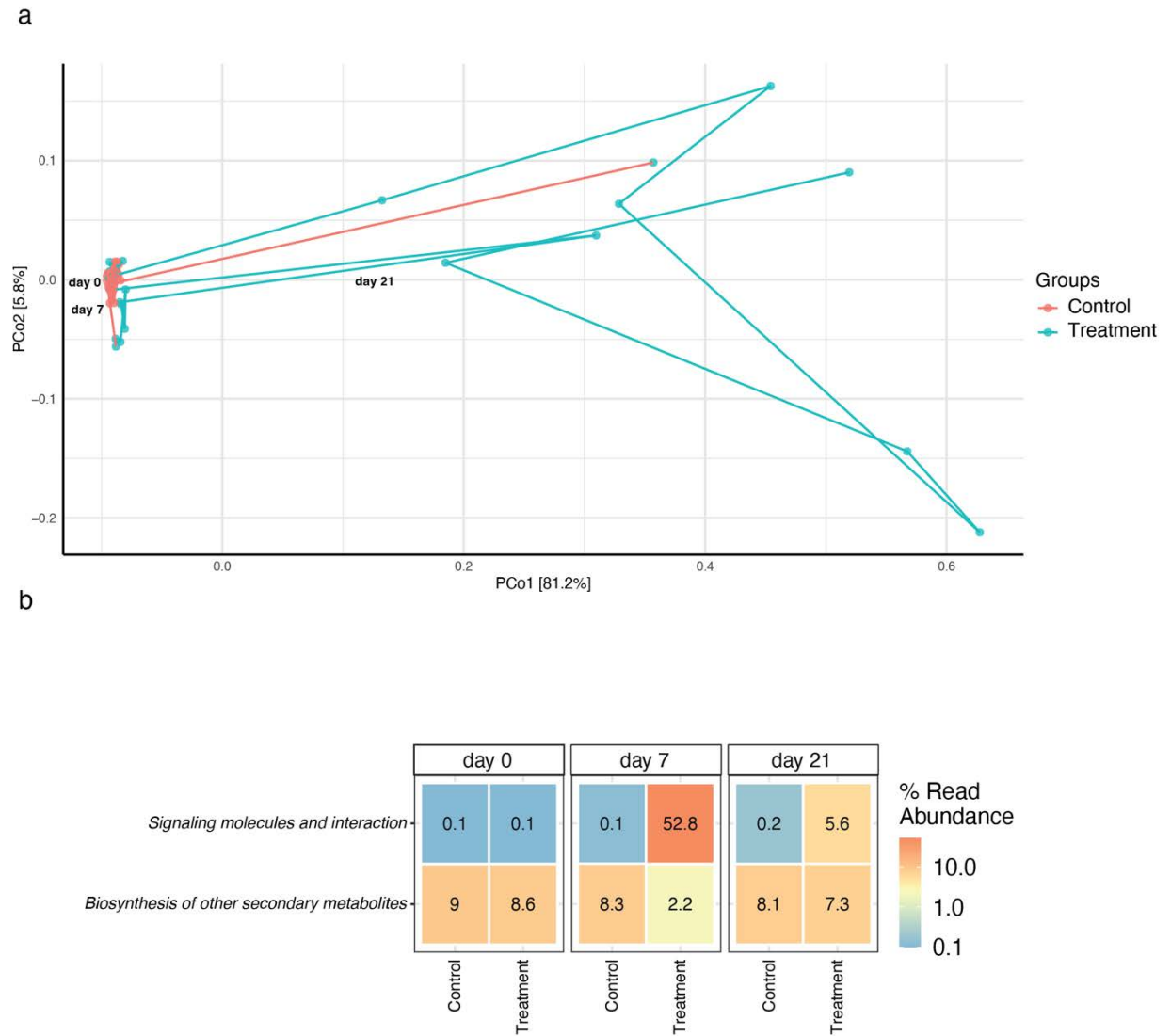
Supplementary figure 4.1: Multifactorial power analysis (based on Raymond *et al.*) to determine number of animal required per treatment and control group.



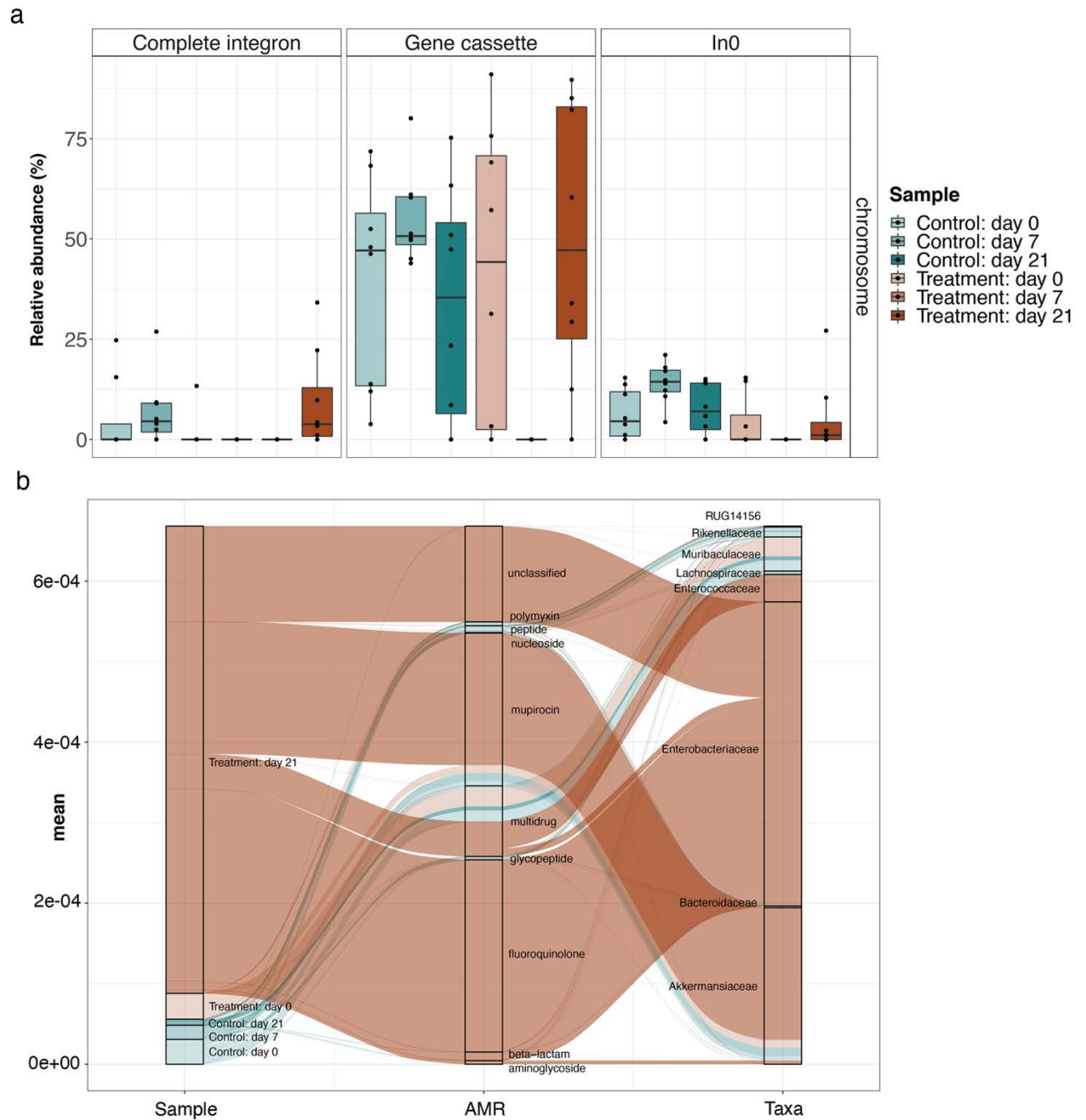
Supplementary figure 4.2: Physiological characterization of mice across timepoints. Daily weight measurements throughout the experimental duration.



Supplementary figure 4.3: Depletion of taxa after antibiotic treatment. **a.** Principal component analyses generated from metagenomic operational taxonomic unit (mOTUS) profiles for the control and treatment groups at days 0, 7 and 21. **b.** Time-tracked ordination plot representing the overall changes in community profile between day 0, through day 7 to day 21. **c.** Relative abundance of the significantly different mOTUs ($adj.p < 0.05$, Two-way ANOVA) in the control and treatment groups at different timepoints

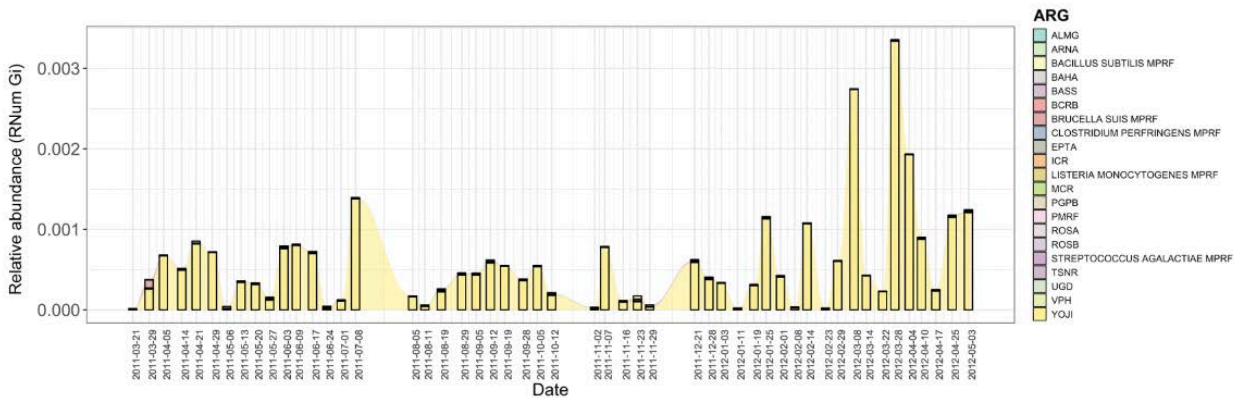


Supplementary figure 4.4: Differential KEGG pathway analysis. **a.** Time-tracked ordination plot of the metagenomic functional profile indicating the changes of the KEGG ortholog functions across time between the control and antibiotic-treated mice. **b.** KEGG pathways that are significantly different ($adj.p < 0.05$, Two-way ANOVA) between the treatment and control groups, are shown as a heatmap displaying longitudinal changes.

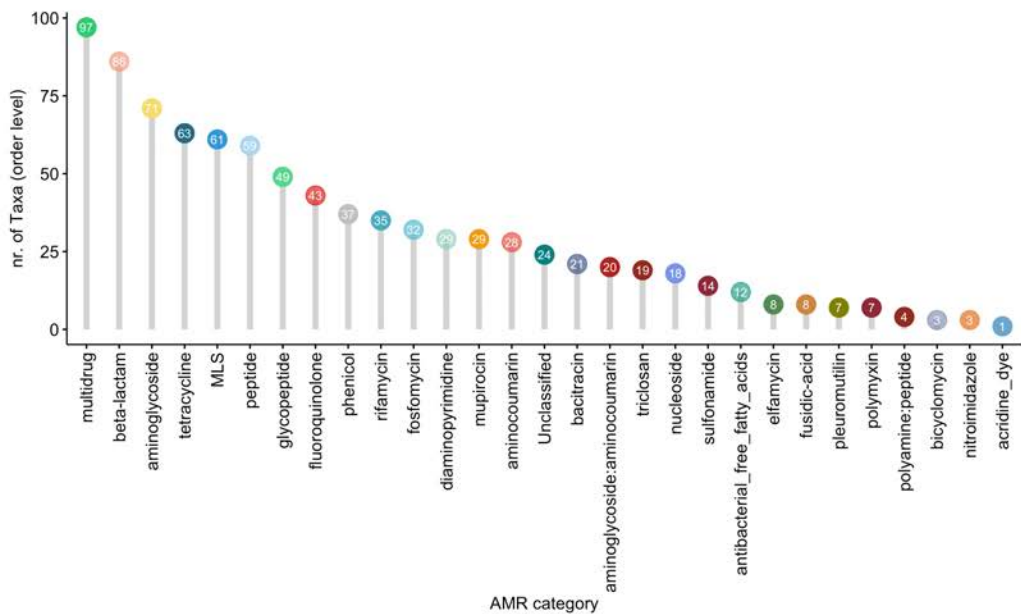


Supplementary figure 4.5: Integron-derived AMR. **a.** Barplots depicting the relative abundance of gene cassettes, complete and incomplete integrons (Ln0) linked to AMR found on the bacterial chromosome across the control and treatment groups at all timepoints. **b.** Alluvial plot indicating integron-mediated AMR categories at all timepoints and the corresponding taxa they are associated with.

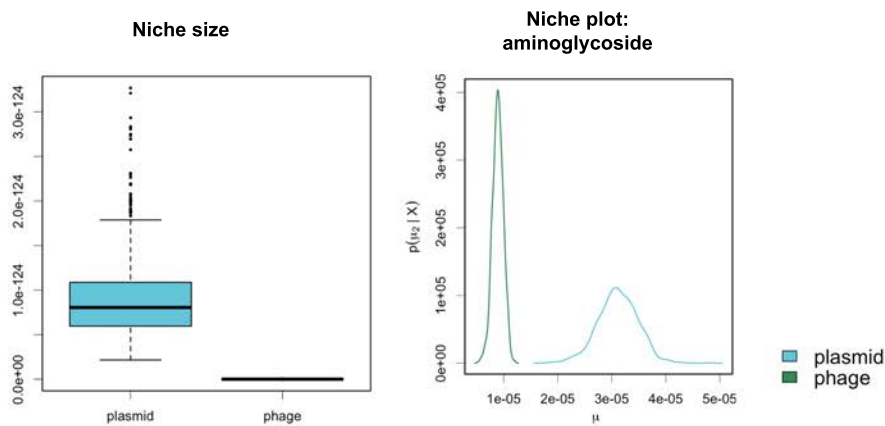
Appendix B.4



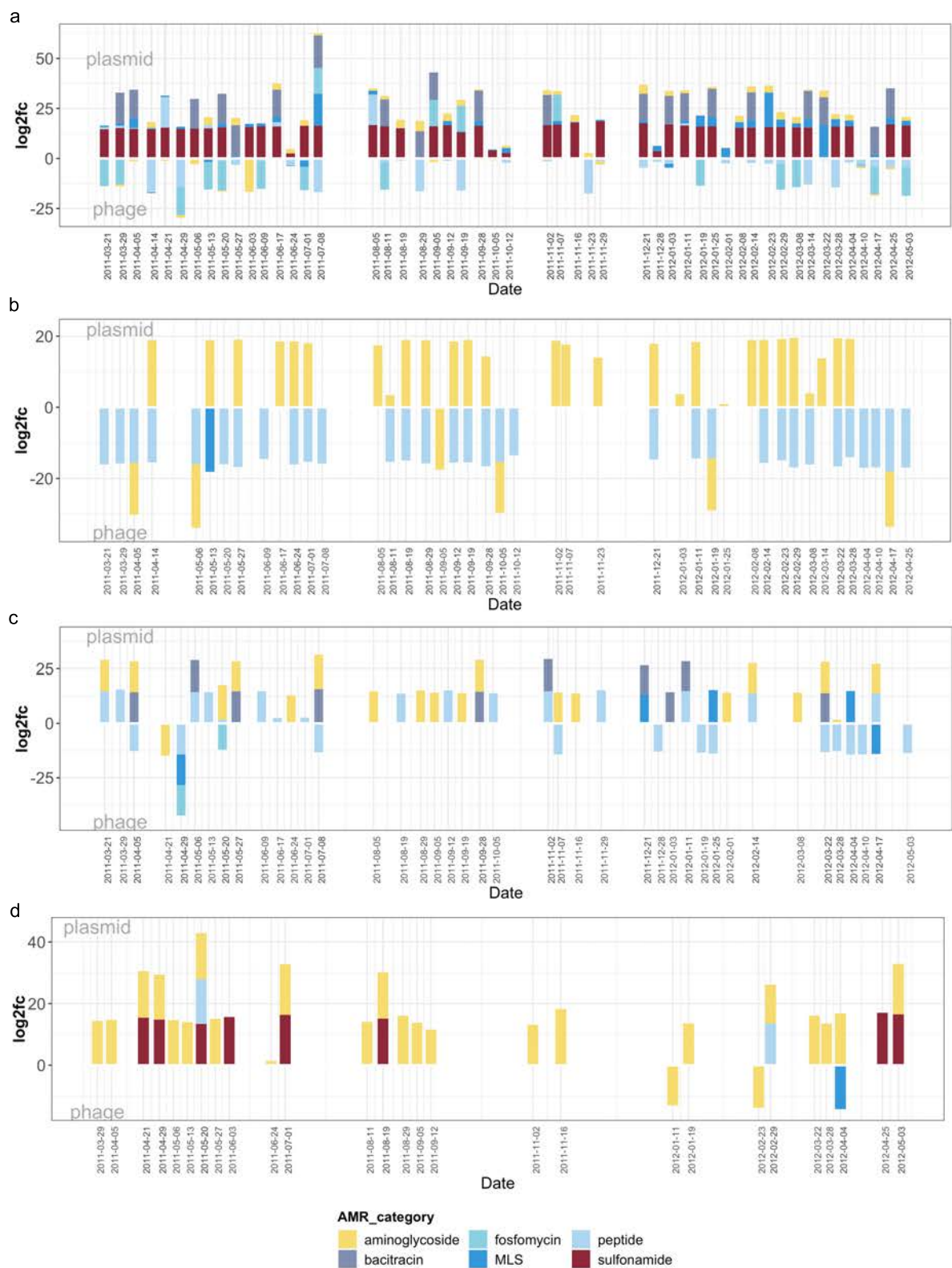
Supplementary figure 5.1: Expression levels of individual ARGs overtime within the BWWTP.



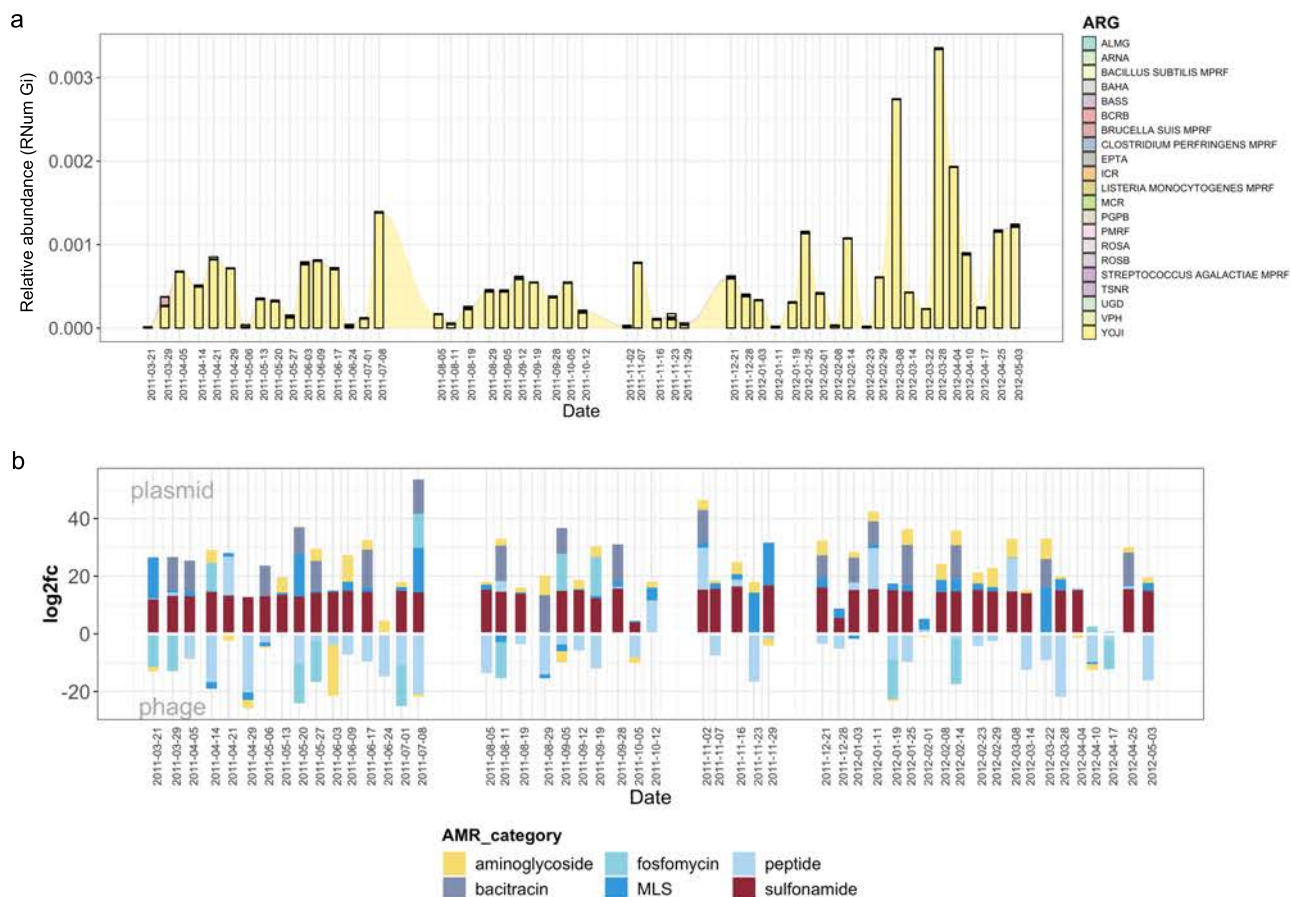
Supplementary figure 5.2: Taxonomic diversity of AMR. The plot indicated the number of taxa (order level) in which the corresponding AMR categories are identified.



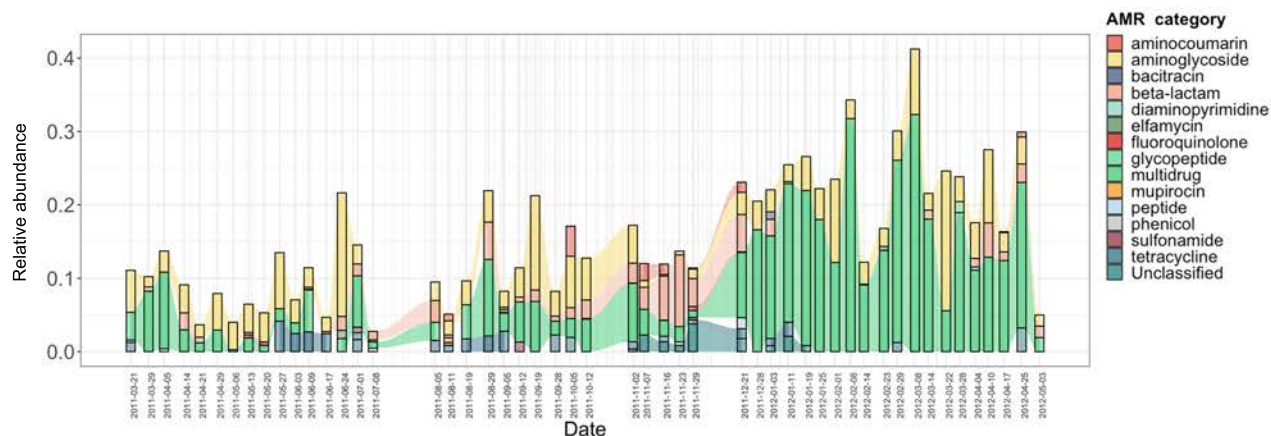
Supplementary figure 5.3: Partitioning of MGs through AMR. The boxplots indicate the niche size (left) for the MGEs (plasmids and phages) based on metagenomic assessment. Niche plots (right) reveal that plasmids tend to differentiate from phages based on their capacity to encode for aminoglycoside resistance.



Supplementary figure 5.4: Differential AMR abundance in MGEs. The barplots report the log₂foldchange of AMR categories over time in MGEs (plasmid versus phage) in: **a.** the general microbial population, **b.** *M. parvicella*, **c.** *Pseudomonas* spp. and **d.** *Comamonas* spp.



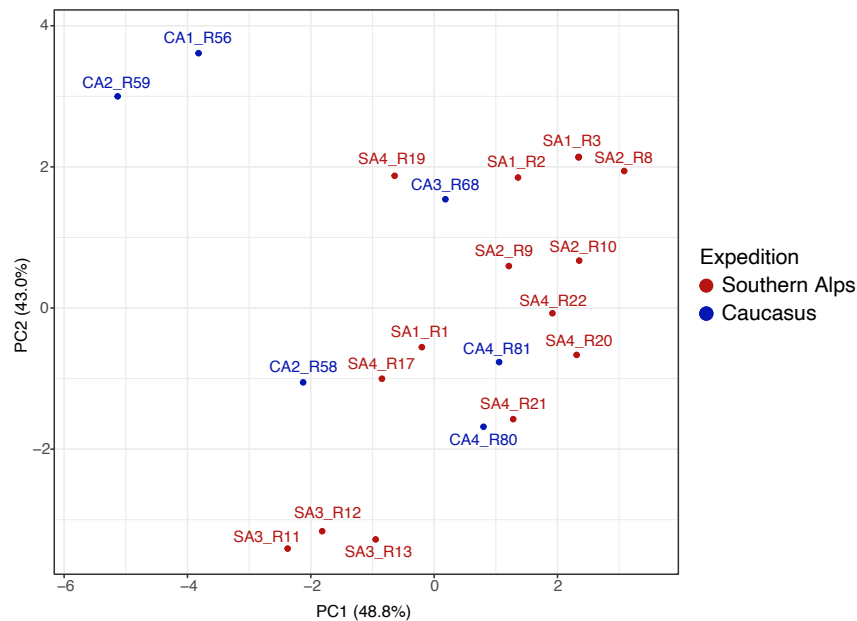
Supplementary figure 5.5: Expression of AMR categories in MGEs. The barplots report the expression levels of AMR categories over time in MGEs (plasmid versus phage) in **a**. Acidimicrobiales, and **b**. Burkholderiales



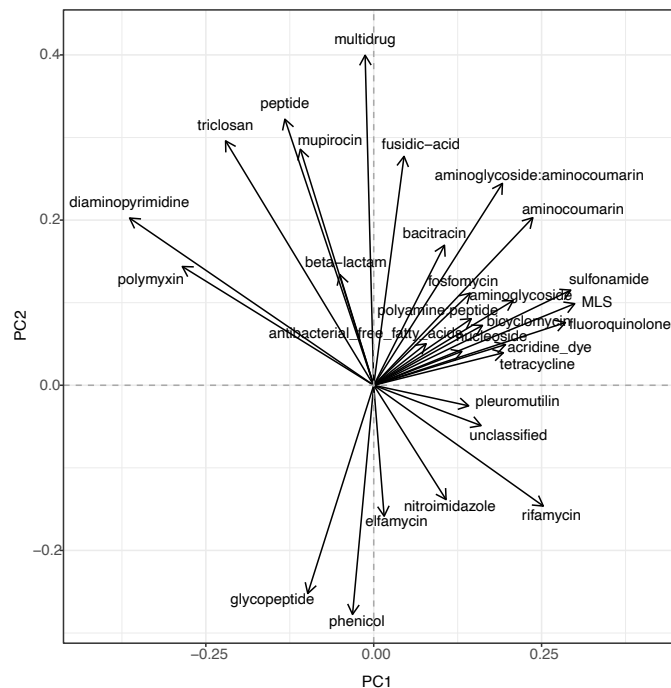
Supplementary figure 5.6: AMR protein abundances. Barplot depicting protein abundances of various AMR categories over time.

Appendix B.5

a



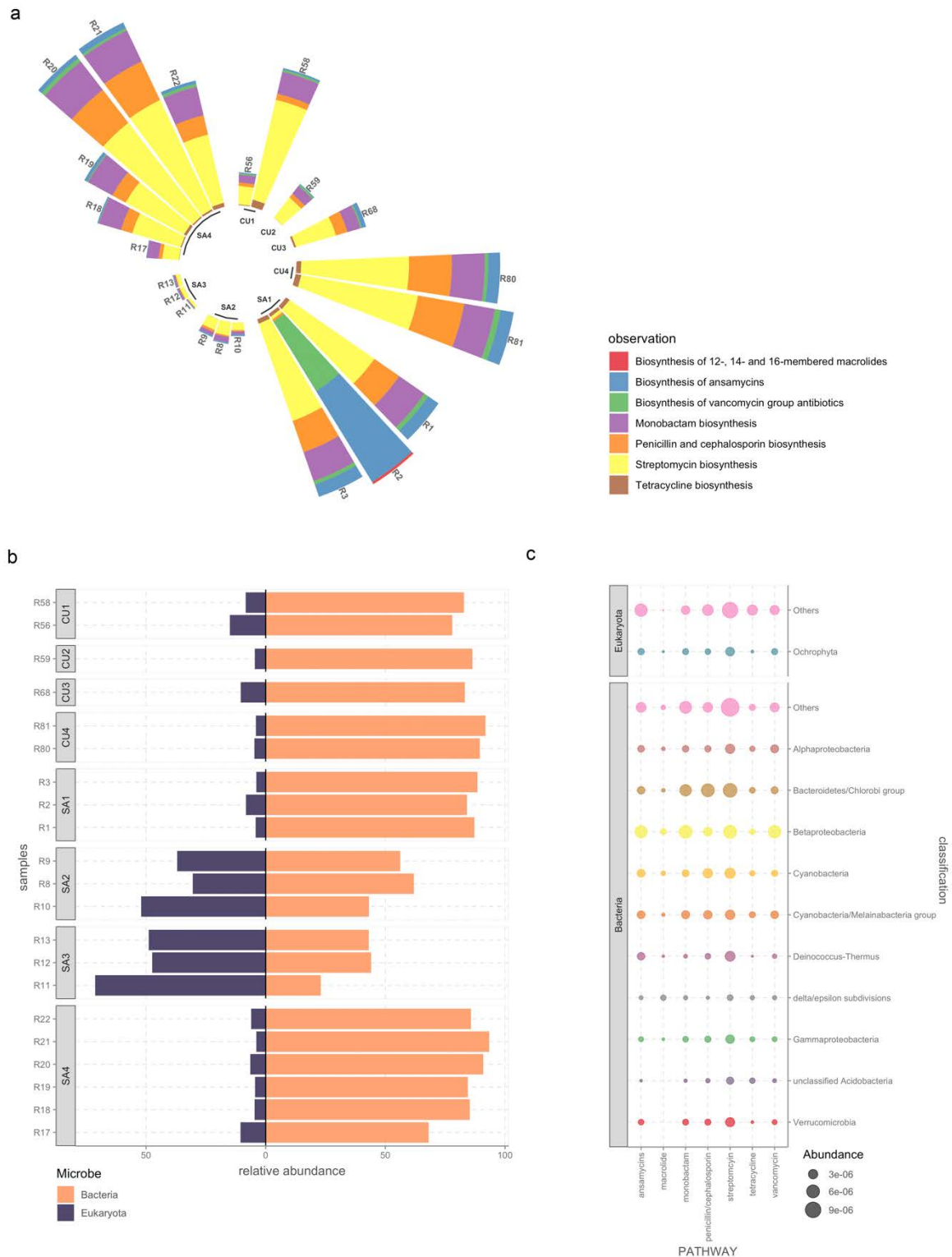
b



Supplementary figure 6.1: Ordination analyses reveal the (dis)similarity of the GFS resistomes. **a.** Principal component analyses depicting the overall similarity of the individual GFS resistomes. Each dot represents the resistome predicted from a single metagenome. SA: Southern Alps. CU: Caucasus. **b.** Biplot demonstrating the underlying factors, i.e. ARG abundances across 29 AMR categories, driving the similarity within the FS epilithic resistomes.



Supplementary figure 6.2: Bacteria and eukaryotic phyla encode AMR. **a.** Relative abundance of the bacteria associated with AMR. The stacked bar plots are faceted by the individual GFs where the epilithic biofilms were collected. The colors represent the individual phyla. **b.** Stacked bar plots indicating the relative abundance of the AMR encoded by eukaryotes.



Supplementary figure 6.3: Antibiotic synthesis pathway assessment via KEGG orthology. **a.** Relative abundance of KEGG pathways associated with antibiotic synthesis across the 21 epilithic biofilms. **b.** Bar plots indicating the relative abundance of the antibiotic associated KEGG pathways mediated by bacteria and eukaryotes. **c.** Normalized relative abundance of pathways associated with antibiotic production in the KEGG database, juxtaposed with the various phyla encoding these genes.

Appendix C. Supplementary Tables

Appendix C.1

Supplementary table 2.1: List of samples analyzed using the PathoFact pipeline grouped by originating study.

Study	Cohort	Group	Accession number	Sample
Bedarf. et al	Parkinson's disease	PD	ERP019674	ERS1647277
Bedarf. et al	Parkinson's disease	PD	ERP019674	ERS1647278
Bedarf. et al	Parkinson's disease	PD	ERP019674	ERS1647279
Bedarf. et al	Parkinson's disease	PD	ERP019674	ERS1647280
Bedarf. et al	Parkinson's disease	PD	ERP019674	ERS1647281
Bedarf. et al	Parkinson's disease	PD	ERP019674	ERS1647282
Bedarf. et al	Parkinson's disease	PD	ERP019674	ERS1647283
Bedarf. et al	Parkinson's disease	PD	ERP019674	ERS1647284
Bedarf. et al	Parkinson's disease	PD	ERP019674	ERS1647285
Bedarf. et al	Parkinson's disease	PD	ERP019674	ERS1647286
Bedarf. et al	Parkinson's disease	Control	ERP019674	ERS1647303
Bedarf. et al	Parkinson's disease	Control	ERP019674	ERS1647304
Bedarf. et al	Parkinson's disease	Control	ERP019674	ERS1647305
Bedarf. et al	Parkinson's disease	Control	ERP019674	ERS1647306
Bedarf. et al	Parkinson's disease	Control	ERP019674	ERS1647307
Bedarf. et al	Parkinson's disease	Control	ERP019674	ERS1647308
Bedarf. et al	Parkinson's disease	Control	ERP019674	ERS1647309
Bedarf. et al	Parkinson's disease	Control	ERP019674	ERS1647310
Bedarf. et al	Parkinson's disease	Control	ERP019674	ERS1647311
Bedarf. et al	Parkinson's disease	Control	ERP019674	ERS1647312
Milani. et al	Clostridioides difficile infection	CDI	PRJNA297269	SRR2582243
Milani. et al	Clostridioides difficile infection	CDI	PRJNA297269	SRR2582246
Milani. et al	Clostridioides difficile infection	CDI	PRJNA297269	SRR2582247
Milani. et al	Clostridioides difficile infection	CDI	PRJNA297269	SRR2582248
Milani. et al	Clostridioides difficile infection	CDI	PRJNA297269	SRR2582251
Milani. et al	Clostridioides difficile infection	Control	PRJNA297269	SRR2582233
Milani. et al	Clostridioides difficile infection	Control	PRJNA297269	SRR2582234
Milani. et al	Clostridioides difficile infection	Control	PRJNA297269	SRR2582237
Milani. et al	Clostridioides difficile infection	Control	PRJNA297269	SRR2582238
Milani. et al	Clostridioides difficile infection	Control	PRJNA297269	SRR2582241
Tett. et al	Psoriasis (Skin metagenome)	Psoriasis	PRJNA281366	SRR2005538
Tett. et al	Psoriasis (Skin metagenome)	Psoriasis	PRJNA281366	SRR2005673

Tett. et al	Psoriasis (Skin metagenome)	Psoriasis	PRJNA281366	SRR2005659
Tett. et al	Psoriasis (Skin metagenome)	Psoriasis	PRJNA281366	SRR2005707
Tett. et al	Psoriasis (Skin metagenome)	Psoriasis	PRJNA281366	SRR2005712
Tett. et al	Psoriasis (Skin metagenome)	Control	PRJNA281366	SRR2005670
Tett. et al	Psoriasis (Skin metagenome)	Control	PRJNA281366	SRR2005727
Tett. et al	Psoriasis (Skin metagenome)	Control	PRJNA281366	SRR2005657
Tett. et al	Psoriasis (Skin metagenome)	Control	PRJNA281366	SRR2005698
Tett. et al	Psoriasis (Skin metagenome)	Control	PRJNA281366	SRR2005710

Supplementary table 2.2: Comparison of virulence factor prediction with MP3. Evaluated performance of the virulence prediction model versus the MP3 prediction tool regarding sensitivity, specificity and accuracy.

	MP3	PathoFact
Sensitivity	0.125	0.886
Specificity	0.992	0.957
Accuracy	0.558	0.921

Supplementary table 2.3: Toxin domains differentially abundant in diseased versus control in *Clostridioides difficile* infection.

HMM Domain	Log2Fold Change	Name	Definition	Group
K11057	-9,61	cpb2	Beta2-toxin	Control
K12788	-6,20	espH	LEE-encoded effector EspH	Control
K01387	-5,94	colA	Microbial collagenase	Control
K11023	-5,01	ptxA, artA	Pertussis toxin subunit 1	Control
PF13945	-4,59	NST1	Salt tolerance down-regulator	Control
PF08998	-4,27	Epsilon antitox	Bacterial epsilon antitoxin	Control
PF15534	-3,83	Ntox35	Bacterial toxin 35	Control
K11062	-3,73	entD	Probable enterotoxin D	Control
K11045	-3,36	cfa	cAMP factor	Control
PF15643	-3,16	Tox-PL-2	Papain fold toxin 2	Control
TIGR03396	-3,10	PC_PLC	Phospholipase C	Control
PF05015	-2,13	HigB-like toxin	RelE-like toxin of type II toxin-antitoxin system HigB	Control
K12340	3,02	tolC	Outer membrane protein	Psoriasis
PF13935	4,70	Ead/Ea22	Ead/Ea22-like protein	Psoriasis
PF14449	4,78	PT-TG	Pre-toxin TG	Psoriasis
K11052	5,20	cylE	CylE protein	Psoriasis

Supplementary table 2.4: Toxin domains differentially abundant in diseased versus control in psoriasis

HMM Domain	Log2Fold Change	Name	Definition	Group
------------	-----------------	------	------------	-------

PF13954	-5,84	PapC_N	PapC N-terminal domain	CDI
PF06609	-3,36	TRI12	Fungal trichothecene efflux pump	CDI
PF13953	-2,90	PapC_C	PapC C-terminal domain	CDI

Supplementary table. 2.5: Toxin domains differentially abundant in diseased versus control in Parkinson's disease.

HMM Domain	Log2Fold Change	Name	Definition	Cohort
K10948	-2.03	hlyA	hemolysin	Control
PF15524	-2.31	Ntox17	Novel toxin 17	Control
PF09599	2.18	IpaC_SipC	Salmonella-Shigella invasion protein c	PD

Supplementary table 2.6: Antimicrobial resistance genes identified within the *Klebsiella pneumoniae* subsp. *Pneumoniae* HS11286 reference genome.

ARG	ARG_SNP	Database	Hits
acrA	n/a	DeepARG	1
acrB	n/a	DeepARG	1
acrB	n/a	DeepARG/RGI	1
acrD	n/a	RGI	1
acrF	n/a	DeepARG	1
AcrF	n/a	DeepARG	1
adeB	n/a	DeepARG	1
bacA	n/a	DeepARG	1
bacterial_regulatory_protein_LuxR	n/a	DeepARG	1
baeR	n/a	DeepARG/RGI	1
baeS	n/a	DeepARG	1
bicyclomycin-multidrug_efflux_protein_bcr	n/a	DeepARG	1
CBP-1	n/a	RGI	1
cob(I)alamin_adenyltransferase	n/a	DeepARG	1
cpxA	n/a	DeepARG	1
CRP	n/a	DeepARG/RGI	1
DNA-binding_protein_H-NS	n/a	DeepARG	2
emrD	n/a	DeepARG	1
emrR	n/a	DeepARG/RGI	1
eptA	n/a	DeepARG	1
Escherichia coli ampH beta-lactamase	n/a	RGI	1
Escherichia coli EF-Tu mutants conferring resistance to Pulvomycin	R234F	RGI	2
Escherichia coli gyrA conferring resistance to fluoroquinolones	S83I	RGI	1
Escherichia coli marR mutant conferring antibiotic resistance	n/a	RGI	1
Escherichia coli mdxA	n/a	DeepARG/RGI	1
Escherichia coli parC conferring resistance to fluoroquinolone	S80I	RGI	1

Escherichia coli UhpT with mutation conferring resistance to fosfomycin	E350Q	RGI	1
Escherichia coli LamB	n/a	DeepARG	1
Escherichia coli mipA	n/a	DeepARG	1
FosA6	n/a	DeepARG/RGI	1
Haemophilus influenzae PBP3 conferring resistance to beta-lactam antibiotics	D350N, S357N	RGI	1
kasugamycin resistance protein ksgA	n/a	DeepARG	1
kdpE	n/a	DeepARG	1
Klebsiella pneumoniae acrA	n/a	DeepARG/RGI	1
Klebsiella pneumoniae KpnE	n/a	RGI	1
Klebsiella pneumoniae KpnF	n/a	RGI	1
Klebsiella pneumoniae KpnG	n/a	DeepARG/RGI	1
Klebsiella pneumoniae KpnH	n/a	DeepARG/RGI	1
macA	n/a	DeepARG	1
marA	n/a	DeepARG/RGI	1
mdtB	n/a	RGI	2
mdtC	n/a	RGI	2
mdtD	n/a	DeepARG	1
mdtG	n/a	DeepARG	2
mdtH	n/a	DeepARG	1
mdtK	n/a	DeepARG	4
MdtK	n/a	DeepARG	4
mdtL	n/a	DeepARG	1
mdtM	n/a	DeepARG	2
mdtN	n/a	DeepARG	1
mexX	n/a	DeepARG	1
msbA	n/a	RGI	1
ompF	n/a	DeepARG	2
ompR	n/a	DeepARG	2
patA	n/a	DeepARG	1
PBP-1A	n/a	DeepARG	1
PBP-1B	n/a	DeepARG	1
penA	n/a	DeepARG	2
PmrF	n/a	DeepARG/RGI	1
ramA	n/a	DeepARG	1