Computational profiling of antimicrobial resistance genes and mobile genetic elements in the human microbiome

Butt, Vicky

Awarding institution:
King's College London

Computational profiling of antimicrobial resistance genes and mobile genetic elements in the human microbiome

Victoria Carr

Centre for Host-Microbiome Interactions

Faculty of Dental, Oral and Craniofacial Sciences

Thesis submitted to King's College London

for the degree of Doctor of Philosophy

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisors, Dr David Moyes and Dr David Gómez-Cabrero. Dr David Moyes has been an incredible all-round mentor for my academic research (especially getting me up to speed with microbiology), my life goals and my wellbeing. I am also incredibly grateful for all the feedback he has given me on this thesis from draft to final edit. Dr David Gómez-Cabrero equipped me with knowledge in computational methods and statistics. I particularly admire his meticulous, mathematical approaches to problems, which certainly kept me on my toes!

Doing a PhD in the last three tumultuous years has been frankly remarkable and bizarre, especially having experienced writing the majority of my thesis about antimicrobial resistance during the COVID-19 pandemic, at home on a couple of floorboards from a skip supported by my keyboard stand. Although the national lockdown provided a large chunk of time for me to immerse myself in thesis writing, it was the most mentally challenging thing I have ever done. I would like to give a special thank you to my partner, Adam Dyster, and my parents who have been unwavering in their unconditional love, support and patience during that time. I am also extremely grateful for Adam proofreading the entire thesis before the first submission. During especially intense periods of writing in isolation, I found much respite playing Dungeons & Dragons with my partner and friends via Skype. This would not have been possible without the Dungeon Master, Alex Tyndall, who led the campaigns and provided hours of entertainment and joyous escapism.

I would also like to extend thank yous to Nadine Mogford, coordinator of the London Interdisciplinary Biosciences Doctoral Training programme, Prof Agi Grigoriadis, my postgraduate coordinator at the Faculty of Dental, Oral and Craniofacial Sciences, and Sarah Easen, Administrator for the Centre for Host-Microbiome Interactions, for listening to my concerns and providing pastoral support. This PhD experience was made a lot more enjoyable by my friends, Ceri Proffitt and Neelu Begum, who I met at the Centre for Host-Microbiome Interactions, and Ismael Kherroubi García, who I met at the The Alan Turing Institute.

In addition to my primary supervisors, I would like to acknowledge the hard work and infallible advice from my University College London supervisor, Prof Peter Mullany, and collaborators: Prof Colin Hill, Dr Andrey Shkoporov, Dr Solon Pissis, Dr Fatima Vayani, Dr Sunjae Lee, Dr Saeed Shoaie, Dr Elizabeth Witherden, Prof Gordon Proctor and Prof William Wade.

Finally, this research was conducted at the Centre for Host-Microbiome Interactions at King’s College London and would not have been possible without my funders, the Biotechnology and Biological Sciences Research Council (BBSRC) (grant BB/M009513/1) and The Alan Turing Institute under the Engineering and Physical Sciences Research Council (EPSRC) (grant EP/N510129/1).

Abstract

Antimicrobial resistance is one of the greatest global health threats of this generation. Antimicrobial resistance in pathogens is leading to infections becoming untreatable with antimicrobial chemotherapy treatments. The number of different types of antimicrobial drugs are limited, meaning pathogens are rapidly developing resistance to commonly used antimicrobial drugs. This has been driven by the overuse of antimicrobial drugs in health care and agriculture, leading pathogens to evolve biological mechanisms to adapt to anthropogenic levels of antimicrobials. Microorganisms, including antimicrobial resistant pathogens, spread and colonise between animals, humans and the environment. Microorganisms have another insidious mechanism of spreading antimicrobial resistance, which is by transferring their genetic resistance determinants between their genomes. This process, known as horizontal gene transfer, has enabled pathogens to acquire antimicrobial resistance genes from other non-pathogenic and pathogenic microbes in close proximity within microbial communities. These antimicrobial resistance genes are usually carried by mobile genetic elements that can integrate into the genome of these pathogens.

Global surveillance using whole genome sequencing and molecular techniques have been adopted to monitor the spread, genetic evolution and resistance severity of antimicrobial resistant pathogens in human and animal populations. Whole genome sequencing has allowed scientists to determine antimicrobial resistance genes in microbial genomes that cause antimicrobial resistance, and in some cases, how these may have been acquired, e.g. carried by mobile genetic elements. Classical surveillance

techniques rely on sequencing a single genome from an isolated, cultured strain. However, this cannot be achieved for microbes that are unculturable. Further, it is incredibly labour-intensive to characterise genomes from all possible strains across microbial communities. Metagenomic sequencing is a more rapid approach that sequences as many genomes from a microbial community as possible, without relying on culturing. Metagenomics has revolutionised the ability to characterise genomes from a variety of species, including profiling antimicrobial resistance genes and mobile genetic elements. A caveat with metagenomics is that it is unable to directly show whether microbes in the community produce antimicrobial resistance traits, which can be achieved with culture-based techniques. However, advances in sequencing technologies and computational methods to interpret metagenomic data may help predict how antimicrobial resistance genes and mobile genetic elements lead to antimicrobial resistance in clinical settings.

In this study, I developed computational tools to profile antimicrobial resistance genes and three types of mobile genetic elements: bacteriophages, plasmids and insertion sequences/unit transposons, from whole, short-read metagenomic data. These tools were applied to publicly available metagenomic sequences of microbial communities in the human gastrointestinal tract across different countries worldwide. This study presents the first attempt at comparing the antimicrobial resistance gene profiles and their associations with mobile genetic elements from metagenomes between sites in the oral cavity and the gut with computational methods. Differences between these profiles are found particularly between gut and oral sites. The gut, surface of the tongue and dental plaque host the greatest diversity of antimicrobial resistance genes and mobile genetic

elements. Antimicrobial resistance genes are rarely found on bacteriophages, but are commonly associated with plasmids and insertion sequences. Insertion sequences are found to be associated with a greater diversity of antimicrobial resistance genes than plasmids, but plasmids encoding antimicrobial resistance genes are highly prevalent. These methodologies and results provide a framework for future development in surveillance and clinical predictions of antimicrobial resistance using metagenomic sequencing technologies.

Contents

Acknowledgements....2
Abstract....4
List of Figures....12
List of Tables....15
Abbreviations....16
Chapter 1: Introduction....19
1.1 The human microbiome....20
1.1.1 The microbiome of the human gastrointestinal tract....22
1.1.1.1 Microbial colonisation in early life....23
1.1.1.2 Human genetics and immune system in early life....24
1.1.1.3 Diet....25
1.1.1.4 Lifestyle....25
1.1.1.5 Antimicrobials and medication....26
1.2 Antimicrobial resistance....27
1.2.1 The use and misuse of antimicrobial drugs....28
1.2.1.1 Human health....28
1.2.1.2 Agriculture....30
1.2.2 Tackling antimicrobial resistance....30
1.2.3 Biochemical mechanisms of antimicrobial agents....31
1.2.4 Biochemical mechanisms of antimicrobial resistance....33
1.2.4.1 Antimicrobial resistance genes....33
1.2.4.2 Intrinsic resistance....35
1.2.4.3 Acquired resistance....36
1.2.4.4 The resistome....36
1.3 Horizontal gene transfer of antimicrobial resistance genes....37
1.3.1 Mechanisms of horizontal gene transfer....39
1.3.1.1 Conjugation....39
1.3.1.2 Transduction....39
1.3.1.3 Transformation....40
1.4 Mobile genetic elements....40
1.4.1 Bacteriophages....40
1.4.2 Plasmids....41
1.4.2.1 *Enterobacteriaceae* plasmids....42
1.4.2.2 *Staphylococci* plasmids....46
1.4.2.3 *Enterococci* plasmids....47

1.4.2.4 Other clinically important resistance plasmids....47
1.4.3 Transposable elements....47
1.4.3.1 Insertion sequences and composite transposons....48
1.4.3.1.1 IS*26* and related elements....50
1.4.3.1.2 IS*Ecp1* family and related elements....50
1.4.3.1.3 IS*Apl1* containing *mcr-1*....51
1.4.3.1.4 IS*91*-like and IS*CR* elements....51
1.4.3.1.5 Unit transposons....51
1.4.3.2 Miniature inverted-repeat transposable elements....55
1.4.4 Integrative conjugative elements....55
1.4.5 Gene cassettes/integrons....56
1.4.6 Integrative and mobilisable genetic elements....57
1.4.7 The mosaic of mobile genetic elements....57
1.4.8 The mobilome in the microbiome....58
1.4.9 The impact of horizontal gene transfer on antimicrobial resistance....59
1.5 Surveillance of antimicrobial resistance....61
1.5.1 Culture-based methods....62
1.5.1.1 Phenotypic testing....62
1.5.1.2 Whole genome sequencing....64
1.5.2 PCR-based methods....65
1.5.3 Metagenomics....66
1.5.3.1 Metagenomic DNA extraction....66
1.5.3.2 Targeted amplicon metagenomics....67
1.5.3.3 Whole metagenomic sequencing....69
1.5.3.4 Functional metagenomics....70
1.5.4 Sequencing technologies....70
1.5.4.1 Short-read sequencing....72
1.5.4.2 Long-read sequencing....73
1.5.4.3 Computational processing....75
1.5.5 Bioinformatic methods of profiling the resistome....76
1.5.5.1 Reference-based detection....76
1.5.5.2 *De novo* discovery....78
1.5.6 Future trends....79
1.6 Profiling the mobilome....80
1.6.1 Abstract....81
1.6.2 Introduction....81
1.6.3 Targeted metagenomic approaches and challenges in extracting MGEs....82
1.6.4 Whole metagenomics....85
1.6.4.1 Challenges in sequencing technologies....85
1.6.4.2 Bioinformatic methods in MGE sequence annotation....87
1.6.5 Technological challenges in host prediction of MGEs....92
1.6.5.1 Wet-lab technologies for microbial host prediction....92

1.6.5.2 Bioinformatic methods in microbial host prediction.....94
1.6.6 Conclusions and further perspectives.....95
1.7 The mobile resistome of the human gastrointestinal tract using whole metagenomics.....97
1.7.1 The resistome.....97
1.7.1.1 *The* GIT resistome in early life.....98
1.7.1.2 Adult GIT resistome.....99
1.7.1.3 The GIT resistome after antimicrobial intervention.....99
1.7.2 The mobile resistome.....101
1.7.3 Opportunities and challenges in using whole metagenomics.....104
1.8 Objectives.....106
Chapter 2: The Resistome.....108
2.1 Introduction to study.....109
2.2 Published paper.....110
2.2.1 Abstract.....111
2.2.2 Introduction.....111
2.2.3 Methods.....113
2.2.3.1 Metagenomic sequence data.....113
2.2.3.2 Processing metagenomic data.....116
2.2.3.3 Identifying ARGs.....116
2.2.3.4 Abundance of ARGs.....117
2.2.3.5 ARG class abundance and antibiotic prescription rate.....118
2.2.3.6 Percentage of samples with ARGs, ARG classes and mechanisms.....118
2.2.3.7 Principal Coordinates Analysis.....119
2.2.3.8 ARG diversity.....120
2.2.3.9 Correlation analysis.....122
2.2.3.10 Data availability.....122
2.2.3.11 Code availability.....123
2.2.4 Results.....123
2.2.4.1 Country and body site-specific differences in resistomes.....123
2.2.4.2 ARG composition differs between the oral cavity and gut.....129
2.2.4.3 Oral and gut ARG profiles associate with species.....134
2.2.5 Discussion.....136
2.2.6 Further discussion.....140
Chapter 3: Bacteriophages and their Association with the Resistome 143
3.1 Introduction.....144
3.2 Methods.....146
3.2.1 Metagenomic data for creating the phage catalogue.....146
3.2.2 Phage contig catalogue.....147
3.2.3 Sequence and functional annotation of jumbo phages.....148
3.2.4 Phage annotation in metagenomes.....149

3.2.5 Phage diversity....150
3.2.6 Microbial composition....152
3.2.7 Longitudinal analysis of phages....152
3.2.8 ARG annotation of phages....153
3.2.9 Code availability....153
3.3 Results....153
3.3.1 Phage composition and diversity differs between GIT sites....153
3.3.2 Phage hosts match varied microbial composition across GIT sites....158
3.3.3 Stability of phage clusters across longitudinal metagenomes....161
3.3.4 Very few phage genomes contain ARGs....164
3.3.5 Circular jumbo phages are commonly found in the oral cavity but not in the gut....166
3.4 Discussion....169
Chapter 4: Plasmids and Resistance Plasmids....174
4.1 Introduction....175
4.2 Methods....177
4.2.1 Creating a plasmid catalogue....177
4.2.2 Plasmid annotation in metagenomes....179
4.2.3 ARG annotation of plasmids in metagenomes....179
4.2.4 Plasmid analysis....179
4.2.5 Plasmid diversity....180
4.2.6 Microbial composition....181
4.2.7 Longitudinal analysis of plasmids....182
4.3 Results....182
4.3.1 Plasmid composition across GIT sites....182
4.3.2 Stability of plasmids across longitudinal metagenomes....189
4.3.3 Resistance plasmids in both oral and gut metagenomes....193
4.4 Discussion....199
Chapter 5: De novo Identification of Transposable Elements and their Association with the Resistome....202
5.1 Introduction....203
5.2 Developing the pal-MEM software....205
5.2.1 The E-MEM algorithm....205
5.2.2 The pal-MEM algorithm....208
5.2.2.1 Finding reverse complements only....208
5.2.2.2 Identifying MEMs from one metagenomic library....209
5.2.2.3 Faster search for inverted repeats....209
5.2.2.4 Ignoring technical reverse complements....209
5.2.2.5 Changing how the MEMs are recorded....212
5.2.2.6 Changing the output....212
5.3 Developing PaliDIS....213
5.3.1 Assembling metagenomic reads to contigs....215

5.3.2 Clustering metagenomic reads to representative reads....215
5.3.3 Identifying inverted repeats using pal-MEM....215
5.3.4 Identifying contigs associated with inverted repeats....217
5.3.5 Identifying ITRs from transposases....218
5.4 Methods to test PaliDIS....219
5.4.1 Creating a catalogue of non-redundant inverted repeats....219
5.4.2 Searching for ARGs....220
5.4.3 IS β-diversity....221
5.5 Results....222
5.5.1 Detecting ITRs and ISs from metagenomic data using PaliDIS....222
5.5.2 Profiles of ISs across GIT sites....222
5.5.3 ARGs are associated with ITRs....225
5.6 Discussion....229
Chapter 6: Discussion....232
6.1 Antibiotic use and prevalence of ARG-associated MGEs....234
6.2 Comparing prevalence of ARG-associated MGEs....237
6.3 The influence of MGE incidence on ARG abundance....239
6.4 Which ARGs that associate with MGEs may be important in AMR?....241
6.5 To what extent can short-read, whole metagenomic data be useful in predicting AMR?....245
6.5.1 Profiling acquired ARGs....246
6.5.2 Identifying pathogens carrying ARGs....247
6.5.3 ARG expression of a resistance phenotype....250
6.5.4 Future work with short-read whole metagenomes....251
6.6 Interventions against AMR....252
6.6.1 Surveillance....252
6.6.2 Diagnostics....253
6.6.3 Therapies to prevent resistant infections....253
6.7 Concluding remarks....255
References....256
Appendices....297
Appendix 2A: Choosing an ARG reference database....297
Appendix 2B: Selecting a mapping tool....300
Appendix 2C: Selecting a breadth of coverage threshold....302
Appendix 2D-M: Supplementary Materials from integrated paper....304
Appendix 3: Chapter 3 Supplementary Figures....328
Glossary....342

List of Figures

Figure 1.1. Molecular AMR mechanisms of antimicrobials in a cell.....35
Figure 1.2. Mechanisms of HGT between donor and recipient cells.....38
Figure 1.3. A typical structure of a composite transposon with two insertion sequences. 50
Figure 1.4. Measuring MICs using a) broth dilutions and b) disk diffusion.....63
Figure 1.5. Targeted and whole metagenomic technologies for extracting MGEs.....83
Figure 1.6. Wet-lab protocols for microbial host identification of MGEs (applicable to plasmids and prophages) using a) SMRT sequencing and b) Hi-C.....94
Figure 2.1. Percentage of individuals that contain ARG classes and ARG mechanisms. 126
Figure 2.2. Clustering of ARG incidence profiles into distinct groups, and comparing ARG abundance to antibiotic use.....128
Figure 2.3. Comparing ARG abundance between the oral cavity and gut.....132
Figure 2.4. Comparing ARG richness between paired body sites.....134
Figure 2.5. Spearman's correlation of ARG and species abundance from saliva samples. 136
Figure 3.1. Phage incidence and abundance profiles.....156
Figure 3.2. Relationship between phage profiles and microbial composition, and abundance of predicted phage hosts.....160
Figure 3.3. Phage cluster stability in longitudinal USA oral and gut samples.....163
Figure 3.4. ARGs in 77 phages.....165
Figure 3.5. Prevalence of jumbo phages.....167
Figure 3.6. Functions and associations of protein-coding gene in jumbo phages.....169
Figure 4.1. Plasmid incidence.....184
Figure 4.2. Plasmid composition across GIT sites.....186
Figure 4.3. Plasmid Cluster Richness between paired GIT sites.....188
Figure 4.4. Overlay of microbiome composition and plasmid composition.....189
Figure 4.5. Plasmid stability in longitudinal USA GIT sites.....192
Figure 4.6. ARG-carrying plasmids of GIT sites from China and the USA.....195
Figure 4.7. Relative abundance of species of the *Enterobacteriaceae* family.....197
Figure 4.8. Resistance plasmid stability in longitudinal USA GIT sites.....198
Figure 5.1. *k*-mer positions that are saved into the hash table at the beginning of a reference sequence.....206
Figure 5.2. A MEM between two technical reverse complement reads.....210
Figure 5.3. Buffer to capture *k*-mer matches at the end of the reads that indicate technical reverse complements.....211
Figure 5.4. Schematic of PaliDIS pipeline for paired-end, short-read whole metagenomes.....214

Figure 5.5. IS profiles in metagenomic samples.....224
Figure 5.6. IS profiles across longitudinal USA samples.....225
Figure 5.7. Percentage of contigs with or without ITRs carrying ARGs.....226
Figure 5.8. Number of samples with ARGs associated with ITRs.....227
Figure 5.9. Number of individuals with ARGs from the same ITR cluster-ARG pairs found in at least 1,2 or 3 timepoints.....228
Figure 6.1. Prevalence of ARG-carrying ISs, phages and plasmids against DDD Per 1,000 in 2015 for each antibiotic class.....236
Figure 6.2. Percentage of samples containing ARGs associated with MGE types.....238
Figure 6.3. MGE incidence versus ARG abundance for MGEs with statistical significance.....240
Figure 2A. Pairwise overlap of annotations from five ARG databases.....299
Figure 2B. Relative abundance of ARG classes using four read mapping tools.....301
Figure 2C. Benchmarking of breadth of coverage threshold.....303
Figure 2Da. ARGs that are found >= 70% of saliva samples.....304
Figure 2Db. ARGs that are found >= 70% of dental plaque samples.....305
Figure 2Dc. ARGs that are found >= 70% of stool samples.....306
Figure 2E. Longitudinal USA samples clustered by ARG abundance profiles.....307
Figure 2F. Comparing ARG abundance between oral cavity sites.....308
Figure 2G. Comparing ARG abundance of different body sites between individuals.. 308
Figure 2Ha. ARG abundance shown for each ARG from China.....309
Figure 2Hb. ARG abundance shown for each ARG from the USA.....310
Figure 2Hc. ARG abundance shown for each ARG from Fiji.....311
Figure 2Hd. ARG abundance shown for each ARG from Western Europe.....312
Figure 2Ia. Differential analysis of ARG abundance between stool and saliva samples from China.....313
Figure 2Ib. Differential analysis of ARG abundance between saliva and dental samples from China.....314
Figure 2Ic. Differential analysis of ARG abundance between stool and dental samples from China.....315
Figure 2Id. Differential analysis of ARG abundance between stool and buccal mucosa samples from the USA.....316
Figure 2Ie. Differential analysis of ARG abundance between stool and dorsum of tongue samples from the USA.....317
Figure 2If. Differential analysis of ARG abundance between stool and dental samples from the USA.....318
Figure 2Ig. Differential analysis of ARG abundance between dorsum of the tongue and buccal mucosa samples from the USA.....319
Figure 2Ih. Differential analysis of ARG abundance between dental and buccal mucosa samples from the USA.....320
Figure 2Ii. Differential analysis of ARG abundance between dorsum of the tongue and dental samples from the USA.....321

Figure 2Ij. Differential analysis of ARG abundance between stool and saliva samples from Western Europe.....322
Figure 2Ik. Differential analysis of ARG abundance between stool and saliva samples from Fiji.....323
Figure 2J. Log2 fold change of ARGs exclusively found in one geographical location between paired samples.....324
Figure 2K. Comparing ARG richness between paired body sites excluding ARGs that are part of or regulate an efflux pump complex.....325
Figure 2L. Spearman's correlation of ARG and species abundance from China stool samples.....326
Figure 3A. Linear regression of Phage Cluster Richness against number of phage contigs.....328
Figure 3B. Frequency of samples containing a number of unique a) phages and b) phage clusters.....328
Figure 3C. Log10 of the proportion of reads mapped to differentially abundant phage clusters for each sample.....329
Figure 3D. Phage incidence and abundance profiles.....330
Figure 3E. Relative abundance of phage taxonomy across GIT sites.....330
Figure 3F. Percentage of phage contigs of phage families with predicted bacteria hosts. 331
Figure 3Ga. Heatmap of log10 relative abundance of phage clusters for the crAss-like phage family.....332
Figure 3Gb. Heatmap of log10 relative abundance of phage clusters for the *Inoviridae* phage family.....333
Figure 3Gc. Heatmap of log10 relative abundance of phage clusters for the *Microviridae* phage family.....334
Figure 3Gd. Heatmap of log10 relative abundance of phage clusters for the *Myoviridae* phage family.....335
Figure 3Ge. Heatmap of log10 relative abundance of phage clusters for the *Podoviridae* phage family.....336
Figure 3Gf. Heatmap of log10 relative abundance of phage clusters for the *Siphoviridae* phage family.....337
Figure 3H. Proportion of phage clusters with predicted phage hosts for each phage family and GIT site.....338
Figure 3I. Heatmap of log10 relative abundance of bacterial genera.....339
Figure 3J. Heatmap of log10 relative abundance of *Eubacterium*, *Haemophilus, Prevotella*, *Streptococcus* and *Veillonella* species.....340
Figure 3K. Percentage of contigs with contig size for each GIT site.....341

List of Tables

Table 1.2. Examples of insertion sequences and unit transposons carrying ARGs and classes they confer resistance to.....53
Table 1.3. Published tools for *de novo* MGE discovery intended for whole metagenomes.....90
Table 6.1. ARGs that have the highest prevalence of being associated with an MGE..242
Table 2M. Coefficients and p-values from linear regression between ARG class abundance and antibiotic prescriptions in 2015 for each country and body site.....327

Abbreviations

A	adenine
AMR	antimicrobial resistance
ARG	antimicrobial resistance gene
ARP	antimicrobial resistance proteins
bp	base pair
C	cytosine
CARD	the Comprehensive Antibiotic Resistance Database
CI	confidence interval
CRISPR	Clustered Regularly Interspaced Palindromic Repeats
DDD	defined daily dose
ddNTP	dideoxynucleotide
DNA	deoxyribonucleic acid
ECOFF	epidemiological cut-off value
EPS	extracellular polymeric substances
ESBL	extended-spectrum β-lactamase
G	guanine
GIT	gastrointestinal tract
HGT	horizontal gene transfer
HMM	Hidden Markov Model
ICE	integrative conjugative element
IME	integrative and mobilisable element
indel	short insertion or deletion

IS	insertion sequence or unit transposon
ITR	inverted terminal repeat
ITS	internal transcribed spacer
kbp	kilobase pair
MEM	maximal exact match
MGE	mobile genetic element
MIC	minimum inhibitory concentration
MITE	miniature inverted-repeat transposable element
MLS	macrolide, lincosamide and streptogramin
NDMS	non-metric multidimensional scaling
NGS	next-generation sequencing
nt	nucleotide
PacBio	Pacific Biosciences
PBP2A	penicillin-binding protein 2A
PCR	polymerase chain reaction
PERMONOVA	permutational multivariate analysis of variance
PROTEST	procrustean randomisation test
pVOG	Prokaryotic Virus Orthologous Group
RNA	ribonucleic acid
rRNA	ribosomal RNA
RPKM	reads per kilobase of read per million
SMRT	single-molecule real-time
T	thymine
TRACA	transposon-aided capture

TSD	target site duplication
VLP	virus-like particles
WGS	whole genome sequencing
ZMW	zero-mode waveguide

Chapter 1: Introduction

1 Introduction

1.1 The human microbiome

Microorganisms are widespread on Earth thriving in living hosts (including humans), indoor environments (such as hospitals), in outdoor environments (like soil), and even in extreme environments (like hydrothermal vents). Microorganisms, or microbes, co-exit and interact with each other and the surrounding environment as microbial communities, also known as the microbiota. The microbiota consists of a complex blend of bacteria, fungi, viruses, archaea and protists in various amounts depending on the environment. Microbiota composition can be broken down by taxonomic rank. Bacteria, archaea and eukaryotes make up three domains, which represent the highest taxonomic rank. Fungi are part of eukaryotes and represent the next taxonomic rank of kingdom. Protists are also eukaryotes but do not belong to a clade (a natural group that includes descendants of a common ancestor). The term protist informally categorises any eukaryotic organism that are not within a kingdom. After kingdom, the hierarchy of taxonomic rank moves down from phylum, class, order, family, genus, to species. Viruses exist within their own group, separate from other domains. Unlike other living organisms, viruses can only reproduce inside other host cells, and they do not belong to any domain. Viruses have their own similar taxonomic ranking system, which starts from realm, moving on to kingdom, phylum, class, order, to genus or sub-family. Within these communities, microbes of these communities can either function as commensals, mutualists, amensalists, pathogens or opportunistic pathogens (pathobionts). A commensal microbe gains benefits from living in the community without benefitting nor

damaging another microorganism, whereas mutualism between two microbes is where they both benefit. Amensalism describes the interaction between two microbes where one is inhibited and the other is unaffected. Pathogens are microorganisms that can cause infections and diseases. Pathobionts can act as commensals that do no damage to the host or microbiota, but can be pathogenic when the host or microbiota are compromised. Until recently, the 10:1 ratio was used to summarise the ratio of microbial to human cells. However, this estimate has been disputed as ill-evidenced dogma[1] and reevaluated to be closer to a 1:1 ratio[2]. Other studies scale human and microbial matter by comparing number of genes. For instance, the human gut microbiota has an estimated three million bacterial genes compared to the human body with approximately 20,000 human genes[3]. The genomes of all microbes within microbial communities are collectively known as the microbiome. Predictions from a collection of studies estimate ~93% bacterial, ~5.8% viral, ~0.8% archaeal, ~0.2% protista and ~0.1% fungal DNA make up the total gut microbiome[4]. Although bacterial genomes are the dominant component of the microbiome, the importance of lower abundant non-bacterial microorganisms can be exemplified by their unique metabolic and mechanistic contributions. Viruses that infect bacteria, known as bacteriophages, shape bacterial communities by lysing bacterial cells or modifying bacterial genomes[4]. Larger archaeal, fungal and protist cells, act like metabolic powerhouses[5–7]. These microbial communities can be stable or dynamic in composition over time depending on the environmental context[8]. A community's dynamics are shaped by metabolic and structural interactions between constituent microorganisms, determined by their individual fitness traits and adaptations to changing conditions. Microbial communities, residing across multiple body sites in humans and animals (including the gut, oral

cavity, oesophagus, skin and vagina)[9], also interact with and have an important influence on the immune system and metabolism of their host[8,10]. Their compositions differ widely between body sites and specific biogeography due to differences in environmental factors such as temperature, pH, oxygen and nutrient availability. When exposed to environmental stressors, the growth rate and survival of particular microbes can fluctuate. For example, the gut microbiota is sensitive to changes in diet[11]. The microbiota's collective metabolism interacts with the human host (i.e. the immune, endocrine and nervous system), which can influence a range of diseases, from inflammatory bowel disease[12], cancer[13] and major depressive disorder[14]. Particularly in the last decade, there has been a surge of research into how the microbiota of the gastrointestinal tract impacts human health and disease, and what interventions can be made.

1.1.1 The microbiome of the human gastrointestinal tract

The gastrointestinal tract (GIT) includes the oral cavity and the gut (stomach and intestine). Although the oral cavity and the gut are connected, their microbial compositions differ due to variations in their environments, including differences in pH dynamics, mechanical force, nutritional availability and oxygen levels[15]. Despite facing changes in their environments, the gut and oral cavity microbiota remain relatively stable. The stool microbiota is dominated by the *Bacteroides* genus. Whilst in the oral cavity, the buccal mucosa, keratinised gingiva and throat are dominated by the *Streptococcus* genus, whereas the saliva, dorsum of the tongue, tonsils, throat and dental plaque tends to have a more even distribution of abundant genera, including *Streptococcus*, *Veillonella*, *Prevotella*, *Neisseria*, *Fusobacterium*, *Actinomyces* and

Leptotrichia[15]. More invasive biopsy studies in patients sampling specific areas of the human intestine have revealed heterogeneous patterns specific to the mucosal layer and lumen, and distal and proximal sites of the intestine[16]. Some individuals will also have significantly different GIT microbiota that may even vary between the most conserved taxa. Hunter-gatherer communities have a higher level of microbial richness and biodiversity in their gut microbiota, with enrichment in *Prevotella*, *Treponema* and *Bacteroidetes* species, than in humans living in urbanised areas[17,18]. These variations are governed by multiple factors, including human genetics and immune system in early life, colonisation of microbes in early life, diet, lifestyle, and use of antimicrobials and other medication, which will be introduced hereafter.

1.1.1.1 Microbial colonisation in early life

The general consensus is that a human's first exposure to microorganisms is during birth[19]. During and soon after a vaginal birth, newborns are inoculated with maternal vaginal and faecal microbes[20,21]. These pioneering microbes seed the expansion and colonisation of other microbes within months. Within the first two and half years, phylogenetic[i] diversity increases gradually, but the abundance of emerging major taxonomic groups can vary with changes in diet and health status, conforming more to the characteristics of an adult microbiome[22]. Dietary changes from weaning further shifts the core gut microbiota, which remains relatively stable throughout a healthy adult's life[23]. Neonates born via caesarian section have a higher level of colonisation by opportunistic pathogens from the hospital environment, including *Enterococcus*,

i Phylogeny is the evolutionary relationship of genetic or physical characteristics between species. (Included in Glossary)

Enterobacter and *Klebsiella* species, and less exposure from maternal *Bacteroides* strains[ii] than infants born vaginally[24].

1.1.1.2 Human genetics and immune system in early life

The microbial composition is unique for an individual. Identical (monozygotic) twins have almost as varied microbial communities as non-identical (dizygotic) twins, detected by a small but significant host genetic effect in a large twins cohort study[25]. The impact of genetic differences can be seen in the immune response. Variability in the human immune response is partly driven by human genetics, and it is predicted up to 10% of this variability is associated with the microbiome[26]. The immune-microbiota interactions that are established in early life are critical[27]. Immunoregulatory T helper 17 (Th17) cells in the lamina propria of the small intestine (a thin layer of connective tissue that forms part of the mucosal layer) are induced upon microbial colonisation of specific commensals that direct the maturation of the developing immune system[28]. The absence of commensal microbes in germ-free mice lead to profound defects in the lymphoid system (specifically lacking formation of isolated lymphoid follicles) and immune functions of the intestine[29,30]. Although the immune-microbiota mechanisms are poorly defined in humans, early-life microbial colonisation and the interactions that take place with the immune system could be important determinants of susceptibility to infections, allergies and inflammatory diseases in later life[27].

ii A strain is a genetic variant or subtype of a species. (Included in Glossary)

1.1.1.3 Diet

As well as in early life, the GIT microbial composition is influenced by dietary changes throughout adult life[11,31]. The dynamics of the gut microbiome in response to dietary changes have been studied extensively. Long-term (and in some cases short-term) dietary changes can alter the gut microbiome significantly. For instance, in the Hadza hunter-gather community in Tanzania, a significantly lower abundance of the *Bacteriodetes* phylum was detected during berry foraging and honey consumption in the wet season compared to the dry season when hunting becomes more dominant[18]. As there are many variations in human diet, it is challenging to pinpoint dietary nutrients that determine the existence of particular taxa from human studies alone. Alternatively, controlled experiments using mice specifically colonised with a small number of microbial strains have been conducted to measure more precise interactions between diet, and microbial and mouse metabolism[32]. However, this is unlikely to fully reflect the complexity of metabolic interactions in naturally occurring animal and human microbiota.

1.1.1.4 Lifestyle

There is an incomprehensible variation of lifestyles that could impact the human GIT. Out of these, associations have been found with pet ownership[33], exercise[34], stress[35], occupation[36], smoking[37] and even sleep deprivation[38]. However, it is difficult to account for other confounding factors in a non-controlled environment and to dissect how a combination of lifestyle factors influence the microbiome.

1.1.1.5 Antimicrobials and medication

Antimicrobials can have a profound effect on the human microbiome. Studies observing effects of antibiotic exposure on the adult GIT show that the resident microbiota can be permanently modified in response to these compounds[39]. For instance, broad-spectrum antibiotics (that act on a wide range of bacteria, both Gram-positive and Gram-negative[iii]) reduce the bacterial diversity and select for bacteria that are resistant to these antibiotics[40]. The microbiome can be modified in various ways depending on the type of antimicrobial drug, the dose or the repetition. Two five-day courses of antibiotic ciprofloxacin treatment (separated by six months) have a greater impact on the gut microbiome than a single course[41]. A single antimicrobial drug as well as a cocktail of several antimicrobials can wipe out many species that are core to what is currently known as a healthy microbiota[42,43]. Other drugs as well as antimicrobials, such as metformin[44], proton pump inhibitors (PPIs)[45], nonsteroidal anti-inflammatory drugs (NSAIDs)[46] and atypical antipsychotics (AAPs)[47], can also lead to long-term changes in the gut microbiota, including eradicating key species related to health status[48]. Any recovery from drug-induced changes in microbial composition may depend on the state of the microbiota after treatment: what species are present and how they interact with each other, alongside diet and lifestyle factors. Exposure to antimicrobials during pregnancy affects gut microbiota maturation of infants in the first two years of life[49]. As the immune response at the GIT is critically dependent on microbial colonisation, antimicrobial perturbation of microbiota maturation can disrupt the homeostasis of the

iii Bacteria can be classified into two categories: either Gram-positive or Gram-negative. Gram-negative bacteria have an outer membrane outside their cell wall, whereas Gram-positive bacteria do not. This means Gram-positive bacteria are more susceptible to cell wall targeting by antibiotics that Gram-negative bacteria. (Included in the Glossary)

immune system leading to inflammatory diseases like inflammatory bowel disease[50] and asthma[51].

1.2 Antimicrobial resistance

In order to survive, microbes need to find an optimum way of living with other microbes in a limiting environment. Microbes compete with each other for limited nutrients and space in their environment. Thus, they have developed tactics to regulate their own requirements by interacting with other microbes. One effective way of doing so is by producing antimicrobial chemicals that can impair or kill another microbe. However, some microbes have evolved ways of living under the exposure of naturally occurring antimicrobials, making them a stable member of a microbial community. Microorganisms can have either intrinsic or acquired resistance to antimicrobials. Intrinsic resistance is where a microorganism may have naturally occurring resistance to an antimicrobial. Intrinsic resistance mechanisms have been present in microorganisms for millennia, driving their co-evolution and integration with microbial communities. In contrast, acquired resistance is the process of a microbe gaining a new resistance mechanism to an antimicrobial drug (described in more detail in Sections 1.2.4.1. and 1.2.4.2.). The discovery and use of antimicrobial drugs to treat and eradicate microbial infections is undoubtedly the greatest achievement in modern medicine. Penicillin, the first mass-produced antibiotic to be used on a large scale from WW2, saved millions of lives, and has pioneered the discovery and synthesis of hundreds of different antimicrobial drugs to target specific pathogens. All known and used antimicrobial drugs have originated from naturally occurring microbial sources, which some microbes

had already developed intrinsic resistance to. Increased use of antimicrobial drugs coupled with pre-existing resistance have led to an upsurge of microbes developing acquired resistance to these antimicrobial drugs. Antimicrobial resistance (AMR) occurs when bacterial, viral or fungal microbes become less susceptible to antimicrobials. This leads to infections potentially becoming difficult to eradicate.

Today, AMR is one of the greatest threats to global health, food security and economic development[52]. The World Health Organisation has estimated 700,000 people die of AMR infections per year, and without sufficient interventions this figure is likely to rise[53]. An increase in the rate of AMR cases in the last 20 years is due to the overuse and misuse of antimicrobial treatments in a rapidly growing global economy and population[54]. Antimicrobials that are regarded as a panacea to eradicating infections have driven the evolution of AMR in pathogens[55].

1.2.1 The use and misuse of antimicrobial drugs

1.2.1.1 Human health

Although antimicrobials have been used for thousands of years, it was not known that infections were caused by microbes until the late 19th century when Robert Koch determined the cause of infectious diseases. Paul Ehrlich discovered arsphenamine was effective against syphilis, which became the first modern antibiotic. Sulfonamides, widely used in the dye-making industry, were also used as antibacterials to treat a range of infections until the early years of WW2, though they had toxic effects. In 1928, Alexander Fleming discovered that *Penicillium notatum* could prevent the growth of

Staphylococcus at concentrations less toxic to humans. Penicillin was then mass produced and widely used to treat soldiers for infections during WW2. The development of mass production and the discovery of penicillin paved the way for development of other relatively non-toxic, naturally occurring antimicrobials between 1945 and 1960. Following this period until 1980, the pace of antimicrobial discovery slowed[56]. The only new antimicrobials created were modifications and elaborations on the biochemical structure of existing ones. Despite the shadow of AMR, antibiotics that were discovered in the mid-twentieth century are still commonly used today. In 2015, the most commonly consumed antibiotics were broad-spectrum penicillins, followed by cephalosporins, quinolones and macrolides[54]. Amoxicillin, one of the most commonly used penicillin drugs, is used to treat a variety of infections, including respiratory, dental and urinary tract infections, and is often used in combination with clathromycin (a macrolide) to treat stomach ulcers. Not only are the same antibiotics still in use, global antibiotic consumption is rising. Between 2000 and 2015, the defined daily dose (DDD, defined by the World Health Organisation as the assumed average maintenance dose per day for a drug used for its main indication in adults) per 1,000 inhabitants by 39%[54]. Particularly in rapidly developing countries and rural regions where antimicrobial use is less regulated, citizens can buy antimicrobial treatments and remedies across the counter without a prescription, with limited understanding of appropriate use[57–59]. Populations who are exposed to more antimicrobials have a higher incidence of developing AMR meaning antimicrobial treatments are rendered less effective, promoting a greater risk of untreatable infections[60].

1.2.1.2 Agriculture

Antibiotics are not only used in clinical health settings; they are also used in agriculture. It is still common practice in animal husbandry worldwide that sub-therapeutic doses of antibiotics are added to animal feed to prevent onset of infections, as well as to produce the desirable side-effect of growing larger animals[61]. Fungicides, such as azoles, are widely used in arable farming.

1.2.2 Tackling antimicrobial resistance

The rate of AMR infections is rising as a result of the increasing selective pressures by the exposure of antimicrobial drugs[54]. Antimicrobial stewardship initiatives have been set up and regulations have been legislated around the world to prevent misuse and overuse of antimicrobials. For instance, the European Union member nations banned the use of all antimicrobials for use as animal growth promoters in 2006[62]. Although measures are being taken to regulate the use of antimicrobials, AMR is present and continues to persist in our ecosystem. Microbial strains have now been found to be resistant to last-resort antimicrobial drugs (glycylcyclines, oxazolidinones, carbapenems and polymixins)[63]. Although the development of AMR can be slowed by cutting back on antimicrobial use, existing antimicrobial resistant pathogens, also known as "superbugs", can easily spread between humans and animals. Outbreaks are commonly seeded in communities with high levels of contact between individuals, especially in hospitals, community social care settings and burgeoning urbanised environments, which are also where usage of antibiotics is higher[64]. The spread of superbug infections (superinfections) are further exacerbated in regions of people living in unsanitary

conditions or with poor access to healthcare[65,66]. Already resistant pathogens, such as *Neisseria gonorrhoeae*[67], are spreading between countries worldwide. Superbugs that persist under ongoing antimicrobial treatment can cause complications in individuals and disrupt healthcare practices, such as surgery, in the long term[68]. Resistant infections, fatal or otherwise, are emerging as silent epidemics.

Currently, there is no panacea to preventing AMR. Instead, multidisciplinary approaches are being combined to tackle AMR from various angles, including drug discovery of alternative antimicrobials[69], better point-of-care diagnostics[70], governmental stewardship to control use of antimicrobials and to incentivise development of alternative antimicrobials[71], and surveillance[72]. However, out of these approaches, global surveillance has been recognised by the World Health Organisation[52] as key to informing governments and institutions (such as Wellcome[73]) of appropriate actions in emergencies or policies to prepare for future outbreaks.

1.2.3 Biochemical mechanisms of antimicrobial agents

There are three main types of antimicrobial agents. These are: 1) disinfectants, such as bleach, that are non-selective and have the purpose of killing a range of microbes on material surfaces; 2) antiseptics which can be applied topically to skin or tissue; and 3) antibiotics which are more selective for particular bacteria and can be administered orally, topically and intravenously. Antimicrobial agent is also a general term for collectively describing any agent that can kill or inhibit the growth of a microorganism, including antibacterials and antifungals. Antibiotics and some antiseptics can be classed

as antibacterial agents that can be either bactericidal, where bacteria are killed, or bacteriostatic, which disrupts bacterial growth.

Antibiotics can be grouped into classes based on their molecular mechanisms: inhibition of cell wall synthesis (β-lactam antibiotics, fosfomycins, isoniazid, glycopeptides), plasma membrane disruption (lipopeptides, polymixins), inhibition of protein synthesis (aminoglycosides, tetracyclines, phenicols, lincosamides, macrolides, oxazolidinones and streptogramins) or nucleic acid synthesis (fluoroquinolones and rifamycins) and disruption of metabolic pathways (sulfonamides, triclosans and diaminopyrimidines)[74]. There are also antibiotics that do not group into one specific class. Nitrofurans, for example, are an unusual class that targets bacteria in a variety of ways. Nitrofurantoin, is a broad-spectrum antibiotic used to treat bladder infections. It disrupts bacterial ribosomal proteins leading to inhibition of protein and nucleotide synthesis, metabolic and cell wall synthesis. Fusidic acid is another example that inhibits protein synthesis but is bacteriostatic like tetracycline, and often applied topically on infected skin or as eyedrops.

These agents can be grouped further into subclasses by their molecular structures. For example, carbapenems, carbapenams, cephalosporins, monobactams and penicillins are all subclasses of β-lactams that differ slightly in their structure. However, they all retain the β-lactam four-atom ring. Antibiotics can also be classified based on their target range; some antibiotics can be classified as narrow-spectrum or broad-spectrum antimicrobials. Broad-spectrum antibiotics act on a wide range of bacteria or on Gram-positive and Gram-negative bacteria, whereas narrow-spectrum antibiotics are only

effective against a limited group of bacteria. Humans may also be exposed to other antimicrobial agents that are not used for clinical or agricultural purposes but may be implicated in AMR. Many acridines, such as proflavine, whilst having antiseptic properties have also been used as dyes for fabrics.

There are fewer types of antifungals than antibacterials for treatment. The four main classes of antifungal treatments are: 1) antimetabolites that inhibits the use of a metabolite (such as antimetabolite flucytosine); 2) azoles that inhibit fungal membrane ergosterol synthesis; 3) echinocandins that inhibit the synthesis of a 1,3-β-glucan that is necessary for maintaining the structure of the fungal cell walls; and 4) polyenes that interfere with permeability and with transport functions in fungal cell membranes.

1.2.4 Biochemical mechanisms of antimicrobial resistance

1.2.4.1 Antimicrobial resistance genes

In some cases, the mechanisms of AMR are caused by the presence of an antimicrobial resistance gene (ARG) within chromosomal or plasmid DNA. DNA encodes this genetic information in a double-stranded sequence of four types of nucleotides (or bases): adenine (A), cytosine (C), guanine (G) and thymine (T). ARGs arise from genes gaining mutations or genes acquired from another microorganism resulting in AMR pathways in the new host. Sometimes, the presence of an ARG is not sufficient to cause resistance, either because it is not expressed because there is no promoter or cannot be translated, or expression levels are low as parts of the protein synthesis pathway are not available.

To increase resistance, ARGs might be duplicated to increase the copy number of transcripts or a promoter can increase ARG expression levels. In other cases, the deletion of a gene, such as those encoding transporter proteins that take up antimicrobials into the cell, can make a microbe more resistant. However, these genes are not considered ARGs, as their absence rather than their presence can cause AMR.

ARGs encode antimicrobial resistance proteins (ARPs) that act against antimicrobial drugs in the five ways **(Fig. 1.1)**. ARPs can function in several different ways. Firstly, the ARPs can directly inactivate the antimicrobial. β-lactamases are a family of ARP enzymes that break the four carbon ring structure in a β-lactam antibiotic, such as penicillin[75]. Narrow-spectrum β-lactamases are enzymes that hydrolyse penicillins but not extended-spectrum cephalosporins, whereas extended-spectrum β-lactamases (ESBLs) hydrolyse most β-lactams including extended-spectrum cephalosporins and monobactams[76,77]. Carbapenemases are a group of β-lactamases that hydrolyse carbapenems that are commonly the last resort treatment for ESBL-producing bacteria[78]. Secondly, the antimicrobial target can be altered. Ligases VanA and VanB are examples of ARPs that alter the target of vancomycin antibiotic. ARGs, *vanA* and *vanB*, code for VanA and VanB that synthesise D-Ala-D-Lac instead of D-Ala-D-Ala, reducing the vancomycin binding affinity[79]. Thirdly, the target can be protected. Quinolone resistance proteins are a family of ARPs that mimic the DNA structure as a pentapeptide repeat and protect DNA gyrases from damage by fluoroquinolone antibiotics[80]. Fourthly, the target can also be replaced. ARG *mecA* encodes penicillin-binding protein 2A (PBP2A), a transpeptidase that replaces the wild type DD-transpeptidase to form the bacterial cell

wall. PBP2A has a lower affinity for β-lactams than DD-transpeptidase meaning it does not bind to the carbon ring of these antibiotics, which prevents them from inhibiting cell wall synthesis[81]. Finally, the cell can reduce the uptake of antimicrobials or increase the activation of efflux mechanisms to extrude antimicrobials. For instance, the MarA activator protein, when overexpressed, downregulates the OmpF porin that makes cells more permeable to multiple antimicrobials and can also induce the efflux pump AcrAB[82].

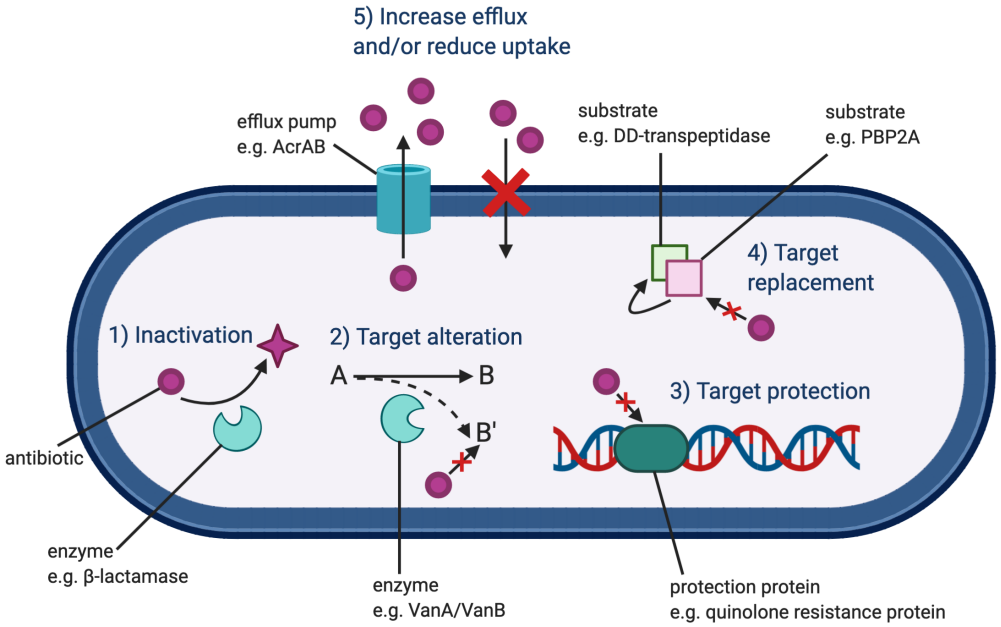

Figure 1.1. Molecular AMR mechanisms of antimicrobials in a cell.
The predominant mechanisms of AMR are: 1) inactivation; 2) altering the target; 3) protecting the target; 4) replacing the target; and 5) increasing efflux and/or reducing permeability of the antimicrobial.

1.2.4.2 Intrinsic resistance

Intrinsic resistance is the innate ability of a microbe to resist or tolerate the activity of an antimicrobial through its inherent structural or functional characteristics. The most common intrinsic resistance mechanisms are the impermeability of the outer membrane

to large or hydrophilic molecules (like vancomycin) entering the cell and the presence of multidrug efflux pumps that can transport antimicrobials out of the cell. These intrinsic resistance mechanisms can be applied to counteract antimicrobial drugs. For example, *Pseudomonas aeruginosa* has a low number of porins in its outer member meaning many classes of antibiotics cannot enter the interior of the cell[83]. Another example is the AcrAB efflux pump that is thought to have evolved in *Escherichia coli* to pump out bile acid but is also able to expel a variety of antimicrobial drugs[84].

1.2.4.3 Acquired resistance

Microorganisms can also acquire ARGs either through mutation of existing genes or by gaining an ARG from another microbe via horizontal gene transfer (Section 1.3). Unlike ARGs encoding intrinsic resistance, acquired ARGs can be part of plasmids as well as integrated in chromosomal DNA. The overuse and misuse of antimicrobial drugs that have caused an increase in selective pressures have led to the rapid emergence of drug-resistant microorganisms with acquired ARGs. It is not uncommon that microbes with short generation times (some *Escherichia coli* strains can double every 20 minutes) can adapt to acquire resistance during exposure to an antimicrobial drug, which then multiply and replace susceptible strains.

1.2.4.4 The resistome

Different microbial communities have different profiles of ARGs with varying abundances. The profile of ARGs in a microbiome are collectively termed the resistome. The size (the number of total ARGs) and diversity (the number of unique ARGs) of the

resistome is influenced by the microbial composition that can harbour different types of ARGs and on the type of exposure the microbiota has had to anthropogenic antimicrobials that drive acquired resistance. Resistome profiles are often used to describe a microbiota's AMR load, reservoir or potential. An AMR load is the number of ARGs that a microbial community can carry at a single point in time[85]. An AMR reservoir is often used to describe the presence ARGs that may be able to spread into other ecosystems by ARG-carrying microbes[86]. AMR potential describes the presence of ARGs in the resistome that have potential to be expressed under antimicrobial drug exposure[87].

1.3 Horizontal gene transfer of antimicrobial resistance genes

Horizontal gene transfer (HGT) is the movement of genetic material between genomes of different organisms. Mobile genetic elements (MGEs) are a type of genetic material that can move within or between genomes via HGT. MGEs can transport genes, including virulence factors and ARGs, most commonly between prokaryotic (bacterial and archaeal) genomes. For example, it is suspected that the methicillin-resistant *Staphylococcus aureus* (MRSA) acquired a gene cassette (a type of MGE with a recombination site) containing the ARG *mecA* via HGT from a non-pathogenic *Staphylococcus* species[88]. These MGEs can then transfer ARGs between different genomes in a microbial community by three major HGT mechanisms: conjugation, transduction and transformation **(Fig. 1.2)**. MGEs can be broadly classified by their

sequence structure into: 1) bacteriophages; 2) plasmids; 3) transposable elements; 4) integrative conjugative elements; 5) gene cassettes/integrons; and 6) integrative and mobilisable genetic elements, which are described in Section 1.4.

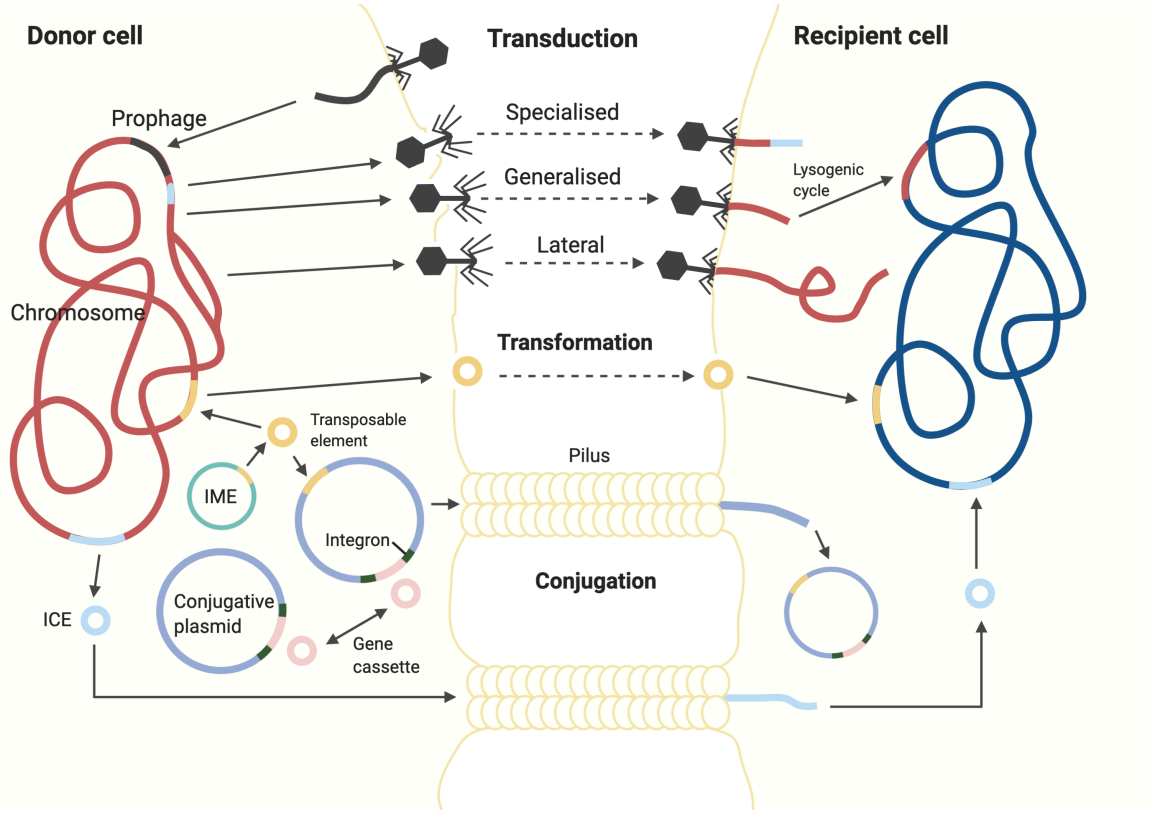

Figure 1.2. Mechanisms of HGT between donor and recipient cells.
There are three major mechanisms of HGT. Transduction is the process of transferring genetic material via a viral vector. Transformation is the natural competence of the recipient cell to incorporate exogenous genetic material. Conjugation is transfer of genetic material between two contacting cells.

1.3.1 Mechanisms of horizontal gene transfer

1.3.1.1 Conjugation

Conjugation is the HGT of MGEs between two directly contacting bacterial cells via sex pili. This structure is the mode of transfer for integrative conjugative elements and

conjugative plasmids, and integrative and mobilisable genetic elements that exploit the conjugative functions of these MGEs. The donor cell produces a pilus that attaches to the recipient cell. A single strand of DNA from the conjugate plasmid or integrative conjugative element is transferred to the recipient. A complementary strand is synthesised in both cells to produce a double stranded circular plasmid, or an integrative conjugative element that inserts into the chromosome.

1.3.1.2 Transduction

Transduction is a mechanism of transferring genetic material into a cell by a viral vector. This is the typical mechanism for HGT of bacteriophages that infect bacterial hosts. There are three types of transduction: 1) generalised; 2) specialised; and 3) lateral. Generalised transduction occurs when bacteriophages can package any bacterial DNA and transfer it to another bacterium, whereas specialised transduction is limited to transferring a particular set of genes[89]. Because specialised transduction is more complex and limited to a certain set of genes, it is thought that most transduction events are generalised[90]. More recently, lateral transduction was discovered in *Staphylococcus aureus,* where very long fragments of bacterial DNA are transferred to another bacterium[90]. In contrast to generalised and specialised transduction, bacteriophage DNA that is integrated into the host genome as a prophage initiates replication before excision, leading to replication of the bacterial host DNA.

1.3.1.3 Transformation

Transformation is a process where exogenous DNA is integrated into a recipient bacterial genome by homologous recombination (the exchange of DNA sequences between identical or similar regions)[91]. Unlike conjugation and transduction, transformation relies on the recipient genome encoding and expressing the required proteins during natural competence (the ability of bacterial to uptake exogenous DNA naturally). A number of MGEs can be transferred via transformation, including plasmids, insertion sequences and integrons/gene cassettes.

1.4 Mobile genetic elements

1.4.1 Bacteriophages

Bacteriophages (phages) are viruses ranging in size from a few to hundreds of kilobases that replicate within bacteria and archaea[92] and can transfer genetic material by transduction. They replicate rapidly, have huge genetic diversity and have genomes that can be comprised of single- or double-stranded DNA or RNA. Phages can replicate through either the lytic or a lysogenic cycle. In the lytic cycle, the phage infects the host cell, replicates and lyses their host at the completion of their replication cycle. In contrast, during the lysogenic cycle, phages inject their genetic material into the cell, which then integrates into the host genome becoming a prophage as part of their replication cycle. Phages that can replicate using both lytic and lysogenic cycles are called temperate phages, while phages that only replicate by the lytic cycle are known as virulent phages. Bacteriophages play an influential role in shaping microbial

composition[93], which may impact the resistome[94]. In fact, bacteriophage cocktails can be used to selectively kill bacterial species that may harbour ARGs in a technique known as phage therapy[95].

Bacteriophages themselves rarely encode ARGs[96]. However, it is still debatable whether phages significantly contribute to the spread of ARGs in microbial communities via HGT. Studies have argued that while a small minority of phages contain ARGs, they still act as a vehicle of ARG transmission and may support an environmental reservoir with potential to cross ecosystems[97]. The reasons behind why phages rarely contain ARGs remain speculative. It has been shown that ARGs are ten-fold less abundant in phage particulate genomes than in integrated prophages[98]. An ARG within another MGE that integrates into an already existing prophage may render it inactive and unable to excise from the host genome during prophage induction. Prophage induction is a process whereby the prophage DNA is excised, transcribed and translated to create viral structural proteins for the lytic cycle. There is some evidence to suggest prophages that carry ARGs do not have any lytic potential[96]. Whether prophages can sustain an active reservoir of transferable ARGs within replicable viral genomes is still under scrutiny. However, phages could encode other determinants of resistance, such as alternative metabolic pathways to bypass antimicrobials that disrupt host metabolic processes[99].

1.4.2 Plasmids

Plasmids are extra-chromosomal replicons present in bacteria, archaea and fungi and mostly transfer by the process of conjugation in bacteria (described in Section 1.3.1.1). They range in size from less than a kilobase to the megabase size range[100]. Conjugative

plasmids consist of a complex system of functional DNA, including its own origin of replication (*ori*) (a specific sequence where replication is initiated), at least one replication initiation protein (Rep), and an origin of transfer (*oriT*) (a short sequence required for the transfer of DNA during conjugation)[101]. The transfer regions of the plasmids encode proteins for mating pair formation (MPF) that functions as a secretion system pore and DNA transfer replication (DTR) that processes the plasmid DNA, such as the relaxase protein that specifically nicks the *oriT* of the DNA strand that is exported to the recipient cell[102]. A conjugative plasmid can cause the transfer of other DNA, when it recombines with another plasmid or chromosomal DNA. Plasmids have an array of functions. Some plasmids are cryptic (have no known function), but many carry genes encoding important functions in the survival and fitness of their host. These include virulence traits and, notably, resistance to antimicrobials. There is currently no systematic approach to classifying resistance plasmids, i.e those that carry ARGs. However, it is widely considered that many resistance plasmids implicated in human health mostly originate from the Gram-negative family, *Enterobacterioceae*[103], and the Gram-positive genera, *Enterococcus* and *Staphylococcus*[104] **(Table 1.1)**. Although they have particular origins, plasmids are highly mobile between different species, but have mostly been found to transfer within their taxonomic families[105].

1.4.2.1 Enterobacteriaceae plasmids

The *Enterobacter* species, *Escherichia coli*, *Enterococcus faecium* and *Klebsiella pneumoniae* are problematic pathogens of the *Enterobacterioceae* family and carriers of resistance plasmids, which are often reported in nosocomial (hospital-acquired) infections. Resistance plasmids in the *Enterobacterioceae* family can be up to 200 kb

(kilobases) long, and include conjugative plasmids and small, mobilisable plasmids[106] **(Table 1.1)**. *Enterobacterioceae* resistance plasmids can be categorised by their mechanisms of replication but are often described by their resistance in the clinic. ARGs encoding for narrow-spectrum β-lactamases, ESBLs and, less commonly, carbapenemases are found in *Enterobacterioceae* plasmids of resistant clinical isolates[107]. Non-β-lactamase ARGs, such as those encoding for aminoglycoside and quinolone resistance, are frequently located in the same plasmid encoding ESBLs and carbapenemases, making *Enterobacterioceae* carriers more difficult to eradicate with antibiotics[108].

Table 1.1. ARGs, resistance plasmids and classes they confer resistance to in *Enterobacteriaceae spp.*, *Enterococci spp.*, *Staphylococci spp.*, *Pseudomonas aeruginosa* and *Acinetobacter baumannii*.

Collated from Partridge et al., 2018[106]. MLS is an abbreviation of macrolide, lincosamide and streptogramin.

Species	ARG	Class that ARG conferring resistance	Plasmid type
Enterobacteriaceae spp.	*armA*	aminoglycoside	L/M plasmids
	blaCMY-2	cephalosporin, cephamycin	A/C, I-complex plasmids
	blaCTX-M family	cephalosporin	F, HI, I2, X plasmids
	blaCTX-M-1	cephalosporin	I-complex plasmids
	blaCTX-M-14	cephalosporin	I-complex plasmids
	blaCTX-M-15	cephalosporin	F, I-complex, Y plasmids
	blaCTX-M-15	cephalosporin	R plasmids
	blaCTX-M-2	cephalosporin	T plasmids
	blaCTX-M-27	cephalosporin	F plasmids
	blaCTX-M-3	cephalosporin	L/M plasmids
	blaCTX-M-62	cephalosporin	N plasmids
	blaFOX family	cephalosporin, cephamycin	L/M plasmids
	blaGES-1	penam, carbapenem, cephalosporin	Q plasmids
	blaIMP family	carbapenem, cephalosporin, cephamycin, penam, penem	HI, N plasmids
	blaIMP-4	carbapenem, cephalosporin, cephamycin, penam, penem	L/M plasmids
	blaKPC family	monobacterium, carbapenem, cephalosporin, penam	A/C, F, I2, L/M, R, U and G/P-6, W, X, ColE1/ColE1-related plasmids
	blaNDM family	carbapenem, cephalosporin, cephamycin, penam	A/C, L/M, N, R plasmids
	blaNDM-1	carbapenem, cephalosporin, cephamycin, penam	HI, T plasmids
	blaNDM-4-like	carbapenem, cephalosporin, cephamycin, penam	X plasmids
	blaNDM-5	carbapenem, cephalosporin, cephamycin, penam	X plasmids
	blaOXA-181	penam, cephalosporin	T, X plasmids
	blaOXA-48-like	penam, cephalosporin	L/M plasmids
	blaSHV family	carbapenem, cephalosporin, penam	L/M plasmids
	blaSHV-12	carbapenem, cephalosporin, penam	X plasmids
	blaTEM family	penam, monobactam, cephalosporin, penem	ColE1/ColE1-related plasmids
	blaVIM family	carbapenem, cephalosporin, cephamycin, penam, penem	R plasmids
	blaVIM-1/4	carbapenem, cephalosporin, cephamycin, penam, penem	W plasmids
	catA1	phenicol	F plasmids
	floR	phenicol	A/C plasmids

Continues next page

	mcr-1	peptide	HI, I-complex, I2, P/P-1, X, Y plasmids
	mcr-2	peptide	X plasmids
	mcr-3	peptide	HI plasmids
	oqxAB	tetracycline, fluoroquinolone, glycylcycline, nitrofuran, diaminopyrimidine (efflux pump)	X plasmids
	qnr family	fluoroquinolone	X plasmids
	qnrB19	fluoroquinolone	ColE1/ColE1-related plasmids
	rmtC	aminoglycoside	A/C plasmids
	strAB	aminoglycoside	A/C plasmids
	sul2	sulfonamide	A/C plasmids
	tet	tetracycline	A/C plasmids
Enterococci spp.	*aadE*	aminoglycoside	Inc18, RepA_N plasmids
	aphA-3	aminoglycoside	Inc18, RepA_N plasmids
	cat	phenicol	Inc18 plasmids
	cfr	oxazolidinone, streptogramin, lincosamide, phenicol, pleuromutilin	Inc18 plasmids
	erm(B)	MLS	Inc18, RepA_N plasmids
	fexB	phenicol	Inc18 plasmids
	sat4	nucleoside	Inc18, RepA_N plasmids
	tet(L)	tetracycline (efflux pump)	Rep_3
	vanA	glycopeptide	Inc18, RepA_N plasmids
Staphylococci spp.	*aacA-aphD*	aminoglycoside	Multi-resistance, Conjugative multi-resistance plasmids
	aadD	aminoglycoside	RC-replicating plasmids
	aphA-3	aminoglycoside	Multi-resistance plasmids
	bcrA	peptide (efflux pump)	Multi-resistance plasmids
	bcrB	peptide (efflux pump)	Multi-resistance plasmids
	blaZ family	penam	Multi-resistance, Conjugative multi-resistance plasmids
	ble	glycopeptide	RC-replicating plasmids
	cat	phenicol	RC-replicating plasmids
	dfrA family	diaminopyrimidine	Multi-resistance, Conjugative multi-resistance plasmids
	erm(A)	MLS	Conjugative multi-resistance
	erm(B)	MLS	Multi-resistance plasmids
	erm(C)	MLS	RC-replicating plasmids
	fosB	fosfomycin	RC-replicating plasmids
	lnu(A)	lincosamide	RC-replicating plasmids
	mphC	macrolide	Multi-resistance plasmids
	msrA	oxazolidinone, tetracycline, streptogramin, macrolide, pleuromutilin, phenicol, lincosamide	Multi-resistance plasmids
	qacA	antiseptic/disinfectant	Multi-resistance plasmids
	qacC	antiseptic/disinfectant	Conjugative multi-resistance plasmids
	sat4	nucleoside	Multi-resistance plasmids

Continues next page

	spc	aminoglycoside	Conjugative multi-resistance
	str	aminoglycoside	RC-replicating plasmids
	tet(K)	tetracycline (efflux pump)	RC-replicating plasmids
	vanA	glycopeptide	Conjugative multi-resistance
Pseudomonas aeruginosa	*blaIMP-45*	carbapenem, cephalosporin, cephamycin, penam, penem	IncP-2 plasmids
	blaKPC family	monobacterium, carbapenem, cephalosporin, penam	plasmids carrying carbapenemase genes
	blaKPC-2	monobacterium, carbapenem, cephalosporin, penam	plasmids carrying carbapenemase genes
	blaSIM-2	penam, carbapenem, cephalosporin	IncP-2 plasmids and plasmids carrying carbapenemase genes
	blaVIM-1	carbapenem, cephalosporin, cephamycin, penam, penem	plasmids carrying carbapenemase genes
	blaVIM-2	carbapenem, cephalosporin, cephamycin, penam, penem	IncP-2 plasmids and plasmids carrying carbapenemase genes
	blaVIM-7	carbapenem, cephalosporin, cephamycin, penam, penem	plasmids carrying carbapenemase genes
Acinetobacter baumannii	*aadB*	aminoglycoside	
	aphA6	aminoglycoside	
	blaOXA-23	penam, cephalosporin	
	blaNDM-1	carbapenem, cephalosporin, cephamycin, penam	

1.4.2.2 Staphylococci plasmids

Staphylococci frequently contain one or more resistance plasmids and can be broadly grouped into three categories based on size: 1) small rolling circle-replicating[iv] (RC-replicating) plasmids that usually encode a single ARG (<1 to 10 kb); 2) multi-resistance plasmids that consist of multiple ARGs and ARG-associated transposable elements (> 15 kb); and 3) larger, conjugative multi-resistance plasmids that transfer at low frequencies and mediate conjugation of other smaller plasmids or even integrate into chromosomes (> 30 kb)[109,110] **(Table 1.1)**.

iv In rolling circle replication, the double-stranded DNA is nicked. The 3' end of the unnicked DNA is elongated and the 5' end strand is displaced. Once replication is complete, the displaced DNA circularises and the second strand is synthesised. The 3' end and 5' end of both strands represent the configuration of bonds between carbon atoms of the DNA pentose backbone. (Included in Glossary)

1.4.2.3 Enterococci plasmids

Enterococci resistance plasmids are generally conjugative, range from 3.3 to 375 kb in size[111,112], and can be classified into their replication initiators Rep_3, Inc18 and RepA_N families[104]. Sometimes they can encode multiple replication initiators which confound classification. Inc18 and RepA_N frequently harbour ARGs, but the Rep_3 family plasmids rarely encode ARGs, apart from *tet(L)*[113] **(Table 1.1)**.

1.4.2.4 Other clinically important resistance plasmids

Other plasmids that originate in important clinical pathogens *Acinetobacter baumannii* and *Pseudomonas aeruginosa* are not as well studied, but a few isolated plasmids have been found in contain ARGs[106] **(Table 1.1)**.

1.4.3 Transposable elements

Transposable elements are MGEs that can integrate into prokaryotic and eukaryotic genomes by transposition. Transposition is the transfer of a section of DNA from one genome to another or to another site on the same genome. Microbial transposable elements include insertion sequences, composite transposons and miniature inverted-repeat transposable elements. Transposition can either be conservative, where an insertion sequence is excised from the donor genome and inserted into the recipient genome, or replicative, when the insertion sequence is duplicated so the donor and recipient each receive of a copy of the insertion sequence[114]. There are two mechanisms of replicative transposition: copy-and-paste and copy-out-paste-in[114,115]. In the copy-and-paste mechanism, the donor and recipient genomes join, the insertion sequence is

replicated, and the genomes separate to leave the original copy in the donor genome and the duplicate in the recipient genome. The copy-out-paste-in mechanism describes an insertion sequence in the donor genome that is replicated out into a circular double-stranded intermediate, and is then integrated into the recipient genome.

1.4.3.1 Insertion sequences and composite transposons

Insertion sequences are short transposable elements containing genes that code for the proteins involved in their own transposition in both chromosomes and plasmids. Most insertion sequences contain one or sometimes two genes encoding transposases, the most ubiquitous gene in prokaryotic and eukaryotic sequences[116]. Insertion sequences and transposons can be broadly classified by their amino acids in their transposase, commonly DDE (aspartic acid, aspartic acid and glutamic acid), DEDD or HUH (two histidine residues separated by any large hydrophobic amino acid), and their mechanism of transposition (either conservative or replicative)[106]. Common DDE types of insertion sequences contain two terminal inverted repeats at each end that are reverse complement sequences[v] of each other. Some insertion sequences are flanked by unique shorter direct repeat sequences, also known as target site duplications (TSDs), which are formed by the duplication of the target site of the insertion sequence when it is inserted[117].

Composite transposons are mobile elements bounded by two copies of two different insertion sequences that can move together in a single unit. A composite transposon can

v The reverse complement of a DNA sequence is the reverse strand that has complementary base pairs. The complementary base pair rule states DNA base pairs are always paired A – T and C – G. This means the reverse complement of GCATGGA is TCCATGC, for example. (Included in the Glossary)

contain one or more passenger genes, such as ARGs, flanked by these two insertion sequences and with two TSDs at both ends **(Fig. 1.3)**. It is also possible for a single insertion sequence to mobilise an adjacent region containing one or more ARGs by being exercised from the genome as a circular translocatable unit[118]. Insertion sequences that associate with ARGs can be broadly categories into five major categories: 1) IS*26* and related elements; 2*)* IS*Ecp1* family and related elements; 3) IS*Apl1* containing *mcr-1*; 4) IS*91*-like and IS*CR* elements; and 5) unit transposons **(Table 1.2)**. Although these represent just some of the types of transposable elements out of so many that are potentially capable of carrying ARGs, these categories cover the mechanisms of less commonly known ones.

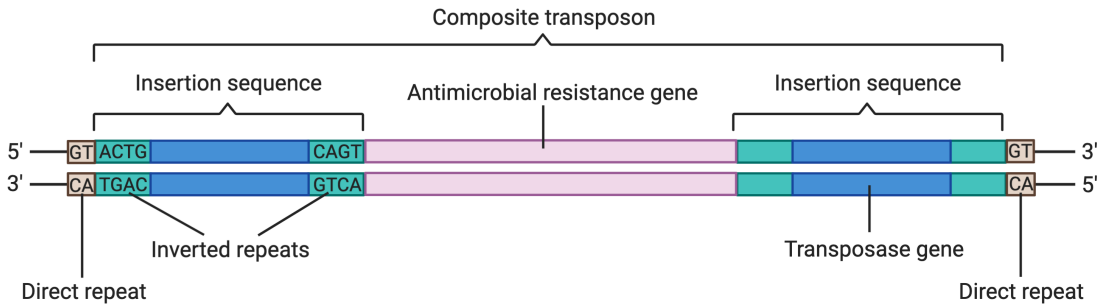

Figure 1.3. A typical structure of a composite transposon with two insertion sequences.
An insertion sequence consists of at least one transposase gene, flanked by terminal inverted repeat sequences. A composite transposon contains passenger genes, such as an ARG, surrounded by two insertion sequences, which are flanked by two TSDs (direct repeats).

1.4.3.1.1 IS*26* and related elements

IS*26*, IS*257* and IS*1216* in Gram-negative and Gram-positive bacteria are important in the dissemination of ARGs **(Table 1.2)**. IS*26* elements are commonly found within arrays of ARGs in resistance plasmids[118]. One copy of IS*26* and an adjacent region containing ARGs can form a circular translocatable unit, which tends to insert next to another copy of IS*26*.

1.4.3.1.2 IS*Ecp1* family and related elements

The insertion sequence family IS*Ecp1* and related elements IS*1247*, IS*Kpn23* and IS*Enca1* have a unique method of transporting ARGs. The IS*Ecp1* family is able to move an adjacent region beyond one of its inverted repeats to create a transposition unit[119]. Their ability to capture many different ARGs is thought to be mediated in this way **(Table 1.2)**. In addition, IS*Ecp1* also contains at least one promoter that can increase the expression of ARGs, such as *blaCTX-M*, an ARG of the CTX-M β-lactamase family conferring resistance to cephalosporin[120].

1.4.3.1.3 IS*Apl1* containing *mcr-1*

IS*Apl1* is another insertion sequence involved in the capture and mobilisation of the *mcr-1* (mobile colistin resistance) gene, conferring resistance to the last-line antimicrobial drug, colistin[63]. It is likely IS*Apl1* uses a copy-out-paste-in replicative transposition mechanism[121]. This case exemplifies how important transposable elements are to the transfer of emerging resistance to antimicrobials of last resort.

1.4.3.1.4 IS*91*-like and IS*CR* elements

IS*91*-like and IS*CR* elements are responsible for carrying a number of ARGs. IS*91*-like elements lack conventional terminal inverted repeats and are suspected to capture adjacent sequences by rolling circle replication[122]. IS*CR* elements are common regions associated with many types of ARGs in class 1 integrons[123] **(Table 1.2)**. Although this has not be demonstrated experimentally, IS*CR*s are thought to capture adjacent sequences by rolling circle replication as well as IS*91*-like elements[124].

1.4.3.1.5 Unit transposons

Unit transposons are a large family of mobile elements that are thought to contain a pair of inverted repeats at both ends rather than insertion sequences, and includes a transposase gene as well as a possible passenger gene, such as an ARG. The distinction between unit transposons and insertion sequences, however, is not clearly defined. There are examples of common insertion sequences that have unit transposon names and relatives of unit transposons with insertion sequence names[125]. Broadly, unit

transposons that are most important in disseminating ARGs either belong to the Tn*3* family that are much larger than typical insertion sequences and undergo replicative copy-and-paste transposition[126] or the Tn*7*-like family that have different transposition mechanisms. These are reviewed extensively in Partridge et al., 2018[106].

Table 1.2. Examples of insertion sequences and unit transposons carrying ARGs and classes they confer resistance to.

Collated from Partridge et al., 2018[106]

Insertion sequence/ Unit transposon	ARG	Class that ARG conferring resistance	Insertion sequence/ Unit transposon type
IS*26* and related elements (in Gram-negative bacteria)	*aph(3')-IIa-ble-aph(6)-Ic*	aminoglycoside, glycopeptide	IS*50*
	aphA1	aminoglycoside	IS*26*, IS*903*
	aphA6	aminoglycoside	IS*Aba14*
	blaFOX-5	cephamycin, cephalosporin	IS*As2*
	blaNDM	carbapenem, cephalosporin, cephamycin, penam	IS*Aba125*
	blaOXA-23	penam, cephalosporin	IS*Aba1*
	blaOXA-237	penam, cephalosporin	IS*Aba1*
	blaOXA-48-like	penam, cephalosporin	IS*1999*
	blaSHV family	carbapenem, cephalosporin, penam	IS*26*
	catA1	phenicol	IS*1*
	catA2	phenicol	IS*26*
	cfr	oxazolidinone, streptogramin, lincosamide, phenicol, pleuromutilin	IS*26*, IS*256*
	mcr-1	peptide	IS*Apl1*
	mcr-2	peptide	IS*Ec69*
	tet(B)	tetracycline (efflux pump)	IS*10*
	tet(C)	tetracycline (efflux pump)	IS*26*
	tet(D)	tetracycline (efflux pump)	IS*26*
IS*26* and related elements (in Gram-positive bacteria)	*aacA-aphD*	aminoglycoside	IS*256*, IS*257*, IS*1216*
	aadD	aminoglycoside	IS*257*, IS*21-558*, IS*Sau10*
	aadE	aminoglycoside	IS*1182*, IS*1216*
	aphA-3	aminoglycoside	IS*257*, IS*1182*
	bcrAB	peptide (efflux pump)	IS*257*
	blaZ family	penam	IS*1216*
	ble	glycopeptide	IS*257*
	cfr	oxazolidinone, streptogramin, lincosamide, phenicol, pleuromutilin	IS*256*, IS*1216*, IS*21-558*, IS*Enfa4*
	dfrA	diaminopyrimidine	IS*257*
	dfrK	diaminopyrimidine	IS*257*, IS*Sau10*
	erm(C)	MLS	IS*257*, IS*Sau10*
	erm(B)	MLS	IS*256*, IS*1216*
	erm(T)	MLS	IS*Sau10*
	fabI	isoniazid, triclosan	IS*1272*
	fosB5	fosfomycin	IS*257*
	fusB	fusidic acid	IS*257*
	ileS2	mupirocin	IS*257*

Continues next page

	lsa(B)	phenicol, macrolide, pleuromutilin, lincosamide, streptogramin, tetracycline, oxazolidinone	IS*21-558*
	sat4	nucleoside	IS*257*, IS*1182*
	spc	aminoglycoside	IS*257*
	str	aminoglycoside	IS*1216*
	tet(K)	tetracycline (efflux pump)	IS*257*
	tet(L)	tetracycline (efflux pump)	IS*257*, IS*Sau10*
	tet(M)	tetracycline (efflux pump)	IS*1216*
	vanA	glycopeptide	IS*1216*
	vanB1	glycopeptide	IS*16*, IS*256*
	vat(A)	streptogramin	IS*257*
	vga(A)	streptogramin, oxazolidinone, tetracycline, lincosamide, phenicol, pleuromutilin, macrolide	IS*257*
	vgb(A)	streptogramin	IS*257*
IS*Ecp1* family and related elements	*aac(3)-Iib*	aminoglycoside	IS*Kpn23*
	aac(3)-IIf-arr	aminoglycoside	IS*1247*
	aph(2")-Ie	aminoglycoside	IS*Enca1*
	blaACC family	monobactam, cephalosporin, penam	IS*Ecp1*
	blaBKC family	carbapenem	IS*Kpn23*
	blaCMY-2-like	cephalosporin, cephamycin	IS*Ecp1*
	blaCTX-M-1	cephalosporin	IS*Ecp1*
	blaCTX-M-2	cephalosporin	IS*Ecp1*
	blaCTX-M-25	cephalosporin	IS*Ecp1*
	blaCTX-M-9	cephalosporin	IS*Ecp1*
	blaOXA-181-like	penam, cephalosporin	IS*Ecp1*
	qnrB	fluoroquinolone	IS*Ecp1*
	qnrE1	fluoroquinolone	IS*Ecp1*
	rmtC	aminoglycoside	IS*Ecp1*
IS*Apl1*	*mcr-1*	peptide	IS*Apl1*
IS*91*-like and IS*CR* elements	*ant(4')-Iib*	aminoglycoside	IS*CR6*
	armA	aminoglycoside	IS*CR1*
	blaAIM-1	penam, cephamycin, cephalosporin	IS*CR15*
	blaCMY/MOX-like	cephalosporin, cephamycin, penam	IS*CR1*
	blaDHA family	cephalosporin, cephamycin	IS*CR1*
	blaNDM family	carbapenem, cephalosporin, cephamycin, penam	IS*CR27*
	blaOXA-45	cephalosporin, penam	IS*CR5*
	blaSPM-1	carbapenem	IS*CR4*
	catA2	phenicol	IS*CR1*
	dfrA10	diaminopyrimidine	IS*CR1*
	floR	phenicol	IS*CR3*
	qnrB	fluoroquinolone	IS*CR1*
	rmtB	aminoglycoside	IS*CR14*
	rmtD	aminoglycoside	IS*CR14*

Continues next page

	sul2	sulfonamide	IS*CR2*
	tet(31)	tetracycline (efflux pump)	IS*CR2*
Unit transposons (in Gram-positive bacteria, *Staphylococci* and *Enterococci)*	*aadE*	aminoglycoside	Tn*5404*
	aphA-3	aminoglycoside	Tn*5404*
	blaZ family	penam	Tn*552*
	dfrK	diaminopyrimidine	Tn*559*
	erm(A)	MLS	Tn*554*, Tn*6133*
	erm(B)	MLS	Tn*551*, Tn*917*
	fexA	phenicol	Tn*558*
	spc	aminoglycoside	Tn*554*, Tn*6133*
	vanA	glycopeptide	Tn*1546*
	vga(A)	streptogramin, oxazolidinone, tetracycline, lincosamide, phenicol, pleuromutilin, macrolide	Tn*5406*
	vga(E)	streptogramin, oxazolidinone, tetracycline, lincosamide, phenicol, pleuromutilin, macrolide	Tn*6133*

1.4.3.2 Miniature inverted-repeat transposable elements

Miniature inverted-repeat transposable elements (MITEs) are short transposable elements generally up to 300 base pairs (bp) in length that, like insertion sequences, contain terminal inverted repeats flanked by TSDs. Unlike insertion sequences, MITEs do not code for their own transposases, but they require a transposase *in situ* for transposition[127]. Instead, they can carry structural motifs or promoter sequences[128]. Only relatively recently have MITEs been discovered to mobilise between bacteria[129] and have been found capable of carrying ARGs by transposition *in vivo* (experiments conducted in live isolated cells)[120].

1.4.4 Integrative conjugative elements

The most complex transposons are the conjugative transposons, also known as integrative conjugative elements (ICEs)[130]. These genetic elements encode their own

conjugation functions and can transfer between bacteria, usually using a similar mechanism as that employed by conjugative plasmids. Unlike plasmids, ICEs are usually integrated into the host chromosome. Their broad-functioning genotype enables them to carry both gene cassettes and transposable elements associating with ARGs[131]. ICEs are largely overlooked compared to plasmids but are known to associate with numerous types of ARGs, such as those conferring resistance to carbapenems, macrolides, phenicol, tetracyclines and vancomycins in bacterial pathogens[131].

1.4.5 Gene cassettes/integrons

Other MGEs include gene cassettes that are commonly part of integrons[132,133]. Integrons are genetic elements that contain a promoter, which direct transcription of genes within the gene cassette, and an integrase and a proximal recombination site where gene cassettes can insert by site-specific recombination. Mobilisable gene cassettes are between half to hundreds of kilobases in length[134] and are associated with a variety of functions, including the spread of ARGs that are usually associated with virulence factors, enzymes and genes coding for heavy metal resistance[106]. Gene cassettes are moved between integrons as an intermediate circular DNA molecule and integrated into the same genome by site-specific recombination (where sequences are exchanged between positions in the DNA using site-specific recombinases). Integrons themselves can be mobilised between different genomes within composite transposons, conjugative elements, plasmids or by transformation. Gene cassettes frequently contain many different ARGs and are named after the ARG they carry[135].

1.4.6 Integrative and mobilisable genetic elements

There are highly heterogeneous elements of less than a kilobase to a megabase in length that transfer through conjugation. They do not contain enough genetic information for independent conjugative transfer, so they utilise the transfer functions of conjugative plasmids or ICEs[136]. They can exist as plasmids or as integrative elements; the latter are sometimes called integrative and mobilisable elements (IMEs)[137]. They commonly contain cargo DNA including ARGs and virulence factors.

1.4.7 The mosaic of mobile genetic elements

Although some MGEs can be categorised by their structural characteristics and mechanism of transfer, many other MGEs consist of a spectrum of mobile genetic structures and mechanisms. For instance, some plasmids are non-conjugative and are transferred horizontally by exploiting the MPF pore previously provided by a conjugative plasmid of the same cell[138]. Thus, they do not have to bear the large genetic load required to encode conjugation functions[139]. This has challenged whether conjugative traits are necessary to determine a mobilisable plasmid. Again, there are examples of plasmids in both Gram-positive and Gram-negative bacteria that are still mobilisable without a relaxase gene[140]. As well as non-conjugative plasmids, IMEs rely on the conjugation of other MGEs. These elements can exist as either plasmids or integrative elements after excision from the genome.

It is common to find MGEs integrated within other types of MGEs, meaning it can be challenging to characterise an HGT event. For example, it is typical for plasmids to

contain gene cassettes/integrons and transposable elements. This provides an alternative route for MGEs to transfer to a chromosome or other resident plasmids of the same cell at times when its plasmid carrier cannot replicate and transfer to a recipient cell. In addition, more complex transposons, such as those of the Tn*3* family, transpose via the formation of a co-integrate, where an adjacent region to one of the insertion sequences forms a transposition unit flanked by the TSD. This allows them to pick up adjacent regions of various sizes during transposition and simultaneously collect sequences from different genomes, such as class 1 integrons[126].

1.4.8 The mobilome in the microbiome

There are a huge variety of MGEs that are an integral part of microbial DNA and the microbiome as a whole. The mobilome is a collective term to describe the profile of MGEs in the microbiome. As well as acting as a reservoir for ARGs, the close proximity of microbes within dense microbial communities provides an ideal environment for the exchange of MGEs between resident microbes. It is estimated over 13,500 genes related to HGT occurred in over 300 species across human body sites[141]. However, this does not decouple whether HGT events occur historically before or after colonisation of these body sites. Nonetheless, HGT is highly frequent in dense human microbial communities, with a rate of 25 times more than in diverse soil ecosystems[142].

1.4.9 The impact of horizontal gene transfer on antimicrobial resistance

The mobilome enables the transfer of the resistome within microbial communities. However, there is limited understanding of how the mobilome collectively impacts risk to undesirable outcomes of AMR. These outcomes include alternative and prolonged treatments to overcome AMR infections, fatalities caused by AMR infections, and other complications, such as secondary infections. Here, I address how the mobilome may influence the clinical outcomes of AMR.

The first consideration is which MGEs can transfer ARGs and what ARGs they carry. MGEs that have the potential to transfer ARGs to a human pathogen pose a greater risk to clinical outcomes of AMR than those that do not[143,144]. Some ARGs that have been acquired by a microbe via HGT due to selection pressures by anthropogenic antimicrobials have a greater propensity to transfer to other microbes, including pathogens, of the same ecological niche[145]. Transfer of naturally occurring and intrinsic ARGs, even without antimicrobial drug selection pressures, has been shown to potentiate resistant phenotypes[146] that can also lead to undesirable clinical outcomes.

Secondly, once the ARGs are transferred in the recipient host, they may or may not function to produce a resistance phenotype upon exposure to antimicrobial treatments. In some cases, the acceptor organism may continue to recapitulate a similar or greater resistance phenotype than in the donor, but in others it may have reduced or no functionality. Once an ARG is acquired via HGT by a human pathogen, the pathogen

itself may not necessarily pose a greater risk to clinical outcomes than without the ARG. It is even probable that some acquired ARGs already residing in pathogens do not lead to clinical cases of AMR. An ARG in one genome may not function in the same way after it is integrated into another genome. The acceptor host may have alternative regulatory or metabolic networks that interact differently to ARGs and other resistance determinants, like promoter sequences, from the donor[147]. Acquiring an ARG may sometimes compromise the fitness of the host, but compensatory mutations may ameliorate its effect[148]. Alternatively, the function of an ARG against an antimicrobial may be redundant if transferred to a pathogen that already has a more efficient mechanism of being resistant to the same antimicrobial[149].

The third consideration is whether the frequency of MGEs transferring ARGs influence the risk of AMR outcome. An increased frequency could be considered to have a greater risk, but given the appropriate context, a single, rare HGT event of an ARG, can have as much potential to cause AMR as more common events[150]. For example, an ARG acquired by an HGT event in a pathogen that is expressed during exposure to antimicrobial drugs, is a greater threat than an MGE that disseminates ARGs through microbial colonies across divergent species, with a fitness cost or inability to persist with the host machinery[151].

Finally, once a genetic resistance determinant is integrated and functioning within the genome of a pathogen, it may or may not lead to an AMR infection. Resistant pathogens in the human body and the environment can only become detrimental to clinical outcome of antimicrobial treatment if they spread and replicate at sites of infection.

In order to estimate the impact of the mobilome, analytical and computational models can be developed to predict clinical outcomes of AMR infections. To model how the mobilome can cause clinical AMR, the factors and events that lead up to its emergence need to be deconstructed and evaluated for their relative impact. Relevant ground truths and immutable frameworks, such as characteristics and mechanisms that define MGEs, can be incorporated in a model's representation of the mobilome. For example, a resistance plasmid always carries a promoter region leading to ARG expression, but a transposon carrying an ARG does not always integrate downstream of a promoter region meaning it may not be expressible. However, knowledge about how the mobilome spreads the resistome is still lacking. In the following discussion, I set out to evaluate what experimental approaches have been applied to date to profile the resistome and mobilome, and the interactions between them.

1.5 Surveillance of antimicrobial resistance

Current methods of antimicrobial resistance surveillance in the clinic rely on culture-based methods of determining a resistance phenotype from a colony of an isolated strain and subsequent sequencing of the its genome for genetic AMR determinants. Molecular approaches, such as polymerase chain reaction (PCR)-based methods and metagenomics, can be applied to discover genetic AMR determinants in microbial communities without relying on culture-based methods. The relative merits and contributions these approaches have to the surveillance of AMR globally is discussed hereafter.

1.5.1 Culture-based methods

1.5.1.1 Phenotypic testing

The phenotype of a microorganism with AMR is described by how resistant it is to a certain concentration of antimicrobial drug. There are two measurements that are commonly used to define this: minimum inhibitory concentration (MIC) and epidemiological cut-off values (ECOFFs) for resistance. The MIC is the lowest concentration of an antimicrobial that inhibits visible growth of a single microbial strain. Different concentrations of antimicrobials are added to agar or broth dilutions[152]. These are then inoculated with a standard concentration of an isolate. The lowest concentration of antimicrobial that inhibits the growth of the inoculate is recorded as the MIC **(Fig. 1.4a)**. Alternatively, a disk diffusion test can be used to test the susceptibility of a microbe to antimicrobials and its MIC[153]. Typically, filter paper disks soaked in a standard concentration of antimicrobial drug are applied to a microbial culture on an agar plate **(Fig. 1.4b)**. As the antimicrobial diffuses away from the disk, the concentration decreases. Lower concentrations that are effective against the strain will produce a wider ring of no microbial growth. In contrast, those disks with no visible rings indicate persistent growth of microbes that are resistant to a high concentration of the drug. In the clinic, breakpoint values can be used to determine whether the MIC is greater or less than clinically relevant levels of AMR. Historically, breakpoints values are chosen concentrations of an antibiotic which define whether a strain is susceptible or resistant to an antibiotic. If an MIC is less than or equal to the susceptibility breakpoint then it is considered to be susceptible to that antibiotic.

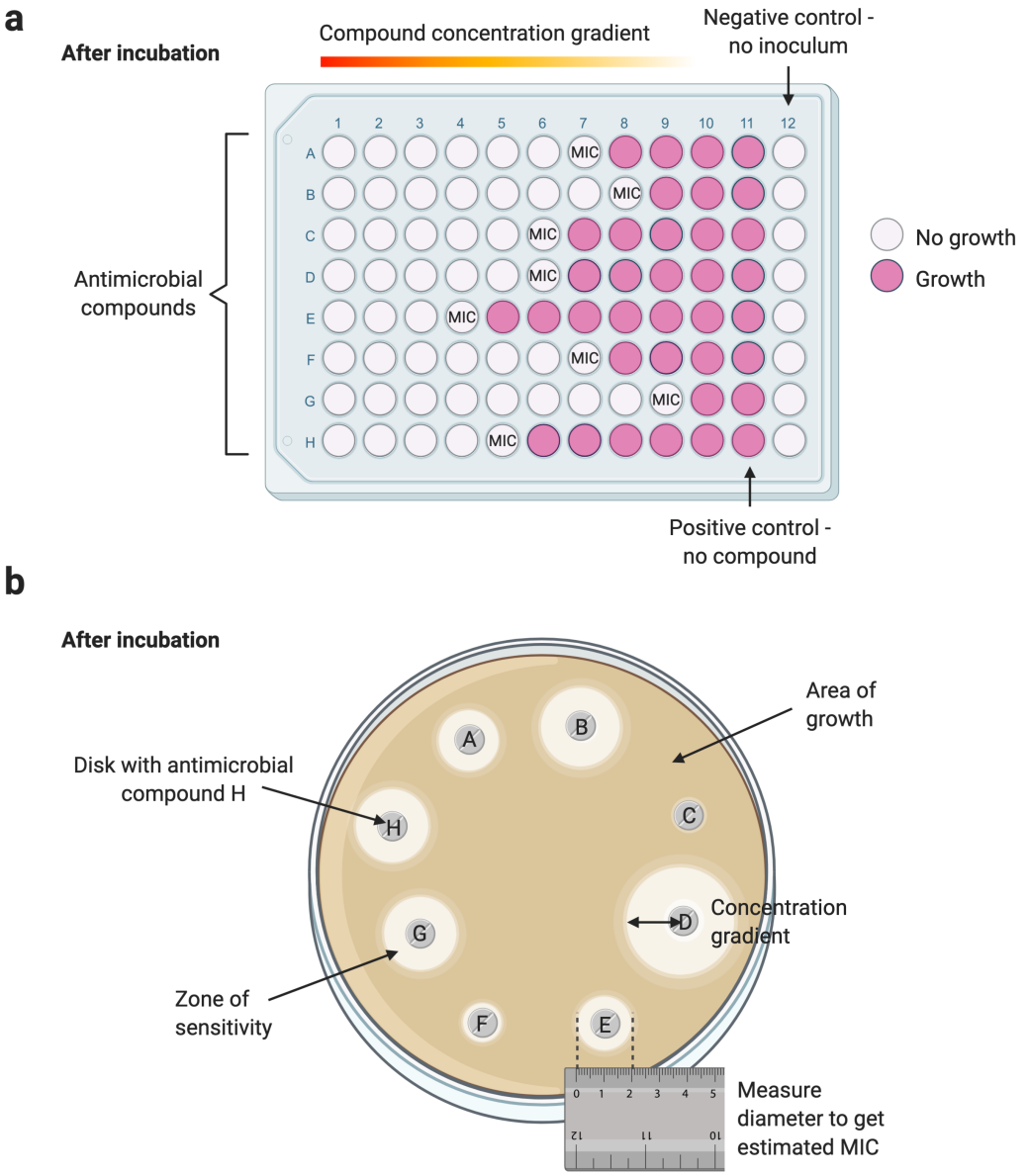

Figure 1.4. Measuring MICs using a) broth dilutions and b) disk diffusion.

If breakpoint data are unavailable, ECOFF values are an alternative[154]. An ECOFF value is an MIC value that is the upper limit of an MIC distribution of multiple wild type strains rather than a single strain[155]. A microbe is defined as a wild type if it does not contain resistance, and it has an MIC below an MIC phenotype established from other

studies of unrelated species. Strains presenting MICs above this ECOFF value are considered resistant. An advantage of using an ECOFF over a breakpoint value is a more continuous scale allows more sensitivity for detecting lower level and potentially emerging resistance. Phenotypic testing has provided insights into the geographical differences in antimicrobial susceptibility. For example, a study used the disk diffusion technique to show differences in susceptibility to 12 antimicrobial agents of *E. coli* isolates causing urinary tract infections across regions in the USA[156].

1.5.1.2 Whole genome sequencing

Culture-based methods alone provide information on how AMR is facilitated on a genotypic level. Whole genome sequencing (WGS) can determine the complete or nearly complete genomes of these susceptible and resistant isolates. These genomes can be compared to other sequenced genomes of related strains to identify novel genetic resistance determinants and whether they were acquired by HGT. Even without susceptibility measures, sequencing pathogens known to have AMR from multiple sources can be compared to track how their genomes evolve as they spread across multiple locations[157]. This is useful in locating origins of outbreaks where differences in genomes of strains converge[64]. Sequence comparison between species could also locate the MGE responsible for ARG transfer. However, resolving the exact timing of an HGT event would be challenging given how frequently microbes replicate. It is possible to predict antibiotic drug resistance and susceptibility using sequence comparison from WGS only, which is faster than phenotypic testing. More accurate predictions can be made from pathogens where HGT rarely occurs and acquired genetic resistant determinants are mostly through mutations, such as *Mycobacterium tuberculosis*[158].

One major limitation with using culture-based methods is that it cannot be applied to unculturable microorganisms, such as intestinal TM7 species that have been shown to harbour genetic resistance determinants[159]. In addition, growing a colony can take time (up to several days of culturing). Molecular approaches, including PCR and metagenomics, can identify known and predict alternative genetic resistance determinants without relying on culturing the microbes.

1.5.2 PCR-based methods

PCR amplifies a target DNA sequence using a pair of oligonucleotide primers that are both complementary to each end of the sequence. Thermostable DNA polymerase extends the oligonucleotides towards each other in a three-step reaction cycle of denaturing, primer annealing and polymerisation. The use of PCR and sequencing has been used extensively to monitor the spread of ESBL-producing *Enterobacteriaceae*[160–162]. PCR has also been able to uncover HGT mechanisms of ARG transfer. For example, class 1 integrons and *Salmonella* genomic island 1 in carrier *Salmonella* species were detected using PCR in *Salmonella* species that were previously non-carriers[163]. Although PCR is highly sensitive, it can only detect specific ARGs that have been previously identified and ignores homologous sequences that could also contribute to similar resistance phenotypes. To alleviate this bias, a less targeted approach is needed to profile all known and unknown ARGs, i.e. the resistome. Metagenomics sequencing provides a non-culture-based solution to profiling the resistome.

1.5.3 Metagenomics

Metagenomics is the study of DNA from all organisms present in a samples or microbial community, referred to as metagenomes. For the last decade, metagenomic sequencing has been more commonly used to study microbial ecology in humans, animals, plants, food and the environment. More recently, metagenomics is being applied to profile the resistome in microbial communities. There are several types of metagenomics that have been instrumental in the understanding of the microbiome and its interactions: targeted amplicon, whole and functional metagenomics. The first step for all methods is to extract the DNA from the sample. Metagenomic DNA extraction will be described first followed by the types of metagenomic methods and how they have been applied to profile the resistome.

1.5.3.1 Metagenomic DNA extraction

Before the DNA is extracted, samples may be frozen to keep them viable. It is important to record the length of time between sample collection and freezing, and the number of freeze-thaw cycles, as these factors could affect the diversity of genomes recovered[164,165]. One key objective for DNA extraction is to minimise contamination, which could profoundly impact findings from which microorganisms are scarce[19]. During extraction, microbes have to be lysed to isolate the DNA, which involves either mechanical lysis, such as bead beating, chemical lysis or a combination of both. Again, the choice of extraction protocol can significantly alter the metagenomic composition of the microbial community[166]. Most metagenomic DNA extraction protocols have focussed on isolating bacterial DNA[167]. However, any microbial niche consists of a

melting pot of DNA from viruses, eukaryotes, protists and archaea. Extraction protocols that are optimised for bacterial DNA can also easily isolate viral DNA as well[165]. However, some kits have been less successful at obtaining fungal DNA[168]. Fungi have a complex cell wall structure that includes chitin, β-glucans and sometimes other structural components, such as melanin. This makes the fungal cell wall more robust and therefore more difficult to lyse than bacterial cell walls and viral particulates. Ultimately, fungal DNA has been estimated to make up ~0.1% of the human gut microbiome[4]. As well as being a likely underestimate, this also undermines its contribution to the microbiota, given its large cell size and unique eukaryotic metabolic capabilities. Mechanical bead beating extraction protocols usually have more success than chemical approaches to lysing fungal cells[168]. However, this can result in more shearing of DNA and shortened fragments, leading to loss of DNA when fragments are selected by size during library preparation for sequencing.

1.5.3.2 Targeted amplicon metagenomics

Targeted amplicon metagenomics target a ubiquitous DNA sequence, most typically a DNA region containing 16S ribosomal RNA (rRNA) genes for sequencing bacteria and archaea or the 18S/internal transcribed spacer (ITS) rRNA gene for sequencing fungi. This approach means only different types of microorganisms can be identified and also mitigates sequencing genomes of other contaminants, such as human DNA. DNA is first extracted from the sample, then PCR is performed to amplify one or more selected variable regions of the rRNA gene. The 16S rRNA gene contains nine, short variable regions (V1 – V9) that are part of the secondary structure[vi] of the small ribosomal

vi The secondary structure is the interaction between base pairs of a nucleic acid molecule or between two nucleic acid molecules

subunit interspersed with conserved regions critical for core functionality[169]. These highly conserved sequences can be used as sites where universal primers can be designed for PCR amplification. More conserved regions are shared between higher level taxa, while less conserved regions are common to lower level taxa, such as genus and species[170]. Many studies using short-read sequencing technologies (Section 1.5.4.1) rely on sequencing smaller parts of the variable regions, meaning it can be challenging to fully characterise complex bacterial communities down to species level[171]. It is possible to sequence the entire region with short-read sequencing. However, this becomes more expensive and less high throughput[172]. Like 16S rRNA for prokaryotes, 18S rRNA provides the secondary structure of the small ribosomal subunit of eukaryotes and also contains both variable and conserved regions for taxonomic classification and specific primer amplification, respectively. The ITS is the spacer region between the small-subunit rRNA and the large-subunit rRNA genes. Part of its sequence is universal for fungi making it ideal for acting as barcode to amplify fungi specifically[173]. Again, the ITS1 and ITS2 subregions of the ITS are commonly used in short-read sequencing technologies as they are smaller and the most universal for fungi. As 18S variable regions tend to be more specific to taxa than ITS regions, ITS sequencing is usually applied to study fungal diversity, whereas 18S rRNA sequencing is preferred for higher resolution taxonomic studies of fungi[174]. It is also possible to target the full 16S/18S/ITS region for greater taxonomic resolution using long-read sequencing (Section 1.5.4.2), which is becoming cheaper to do[175]. Given the remote locations of many ARGs and MGEs relative to rRNA genes, amplified fragments of 16S/18S/ITS regions from amplicon metagenomics would not include the vast majority

of ARGs and MGEs. However, a software tool, called PICRUSt, can predict functional gene composition using reference sequences, including ARGs, from 16S rRNA data[176].

1.5.3.3 Whole metagenomic sequencing

In whole metagenomics (also known as shotgun metagenomics), all DNA is extracted from a sample, as with targeted amplicon sequencing. However, after the DNA is extracted and before sequencing, it is fragmented randomly into a DNA library and these fragments are ligated with adapters that are recognisable by the sequencing platform (Section 1.5.4). In principle, whole metagenomic sequencing allows genomes to be resolved at lower level taxa (such as species and strain) than amplicon metagenomics. This is because taxonomic profiling is limited to the specificity of 16S/18S/ITS regions in targeted metagenomics, whereas whole metagenomic sequencing can include strain-level variability from other sequences. As a result, whole metagenomics can profile multiple domains, including bacterial, archaeal, eukaryotic and viral sequences, together. This means bacteriophages, which are a type of MGE, can also be sequenced. However, the accuracy of taxonomic composition is heavily dependent on the sequence read depth[vii] [177]. Targeted amplicon metagenomic sequencing is more accurate at profiling taxonomic composition and abundance based on previously catalogued operational taxonomic units (OTUs)[viii] [178]. The major advantage of using whole over targeted amplicon metagenomics is that it can profile the resistome and mobilome (or indeed any gene profile) without relying on outdated or incomplete information from reference strain genomes.

vii The read depth (also known as coverage depth) is the number of unique reads that contain a particular nucleotide of the represented sequence. (Included in Glossary)

viii OTUs are a cluster of similar sequence variants of 16S/18S/ITS marker gene sequences, where each cluster represents a taxonomic unit of a genus or species

1.5.3.4 Functional metagenomics

Instead of applying bioinformatic predictions, functional metagenomics is a culture-based approach that aims to define the gene function, including the mechanisms of novel resistance genes[179], based on phenotypic expression. Microbial DNA is extracted, fragmented and inserted into a plasmid library. The plasmids are then introduced into a culturable host microbe, such as a *E. coli*, to create a metagenomic DNA library. An environmental stressor, such as an antibiotic, is applied to this library. Those isolates that survive, or produce another phenotypically positive response, are then sequenced to identify potential genetic determinants of that response. The plasmids of strains that survive are sequenced, which enables both known and completely novel ARGs to be discovered. In fact, a study found 79% of newly discovered ARGs in their functional metagenomic libraries were not previously classified as ARGs in databases[180]. There are flaws to this technique, however. Depending on the fragment size and locations of genetic resistance determinants, an ARG may be truncated or multiple ARGs coding for multiple regulatory elements and promoters could be omitted in a plasmid[181]. In addition, an ARG that may be expressed in one organism may not have a discernible phenotype in the surrogate host or is not expressible under *in vitro* growth conditions[182].

1.5.4 Sequencing technologies

Sequencing is the process of determining the arrangement of nucleotides in DNA or RNA. Sanger sequencing in 1977 became one of the first and most successful commercially viable DNA sequencing techniques. Sanger sequencing exploits the

function of DNA polymerase to incorporate chain-terminating dideoxynucleotides (ddNTPs) during DNA replication *in vitro*[183]. The ddNTPs are labelled with radioisotope or fluorescent dye. Each radiolabelled ddNTP (ddATP, ddCTP, ddGTP and ddTTP) is separated into four lanes of a gel, or if using fluorescent dye different wavelengths are used for each type of ddNTP. When one of these ddNTPs form a phosphodiester bond with a nucleotide from the original sequence, it prevents the DNA polymerase from extending the DNA. If using fluorescent dyes, the fluorescence is then captured by a detector in computer automated instruments based on capillary electrophoresis. If using radiolabelling, once DNA synthesis is complete, dark bands produced by the radioisotopes on photographic film reveal the order of the nucleotide sequence. The resulting DNA fragments after DNA sequencing are heat denatured and separated by size using electrophoresis. Largely driven by efforts to sequence the human genome in the Human Genome Project, next-generation sequencing (NGS) superseded Sanger sequencing for large-scale, automated sequencing projects, and became more commercially available from the mid 2000s[184]. However, Sanger sequencing still remains in use today for accurate sequencing of smaller DNA molecules. Also known as short-read sequencing, NGS is based on sequencing spatially separated fragments (or reads) of DNA molecules in parallel. The next breakthrough came by the end of the ‘00s by third-generation sequencing, which allowed longer reads to be sequenced[185]. High-throughput sequencing technologies made it easier and faster for more DNA to be sequenced, meaning metagenomes could be sequenced by short-read and long-read sequencing technologies. However, the quality of the metagenomic data can depend on the type of technologies used. In the following sections, short-read and long-read

sequencing technologies are evaluated for their merits and challenges in sequencing metagenomes for their resistomes.

1.5.4.1 Short-read sequencing

The current "gold-standard" sequencing technology is short-read Illumina sequencing, which is commonly used both for its affordability as well as its high accuracy. For example, the latest NovaSeq 6000 System has a 0.1% error rate in base calling (the process of assigning nucleotide bases from signals). DNA fragments are attached on one end to a flow cell and are amplified into clusters of the same fragments. After amplification, the fragments are read using sequence by synthesis. Four nucleotides (A, C, G and T), that are modified with a reversible fluorescent blocker, are washed over the surface of the flow cell separately. A nucleotide that complements a nucleotide of the fragment is added and the corresponding fluorescent dye is recorded. The fluorescent blocker only allows the DNA polymerase to add one nucleotide at a time. Once a recording is made, the fluorescent blockers are then removed from the newly synthesised nucleotides to allow the next cycle of extension. During sequencing, the recorded images are processed into much smaller text files of nucleotide characters, which can be interpreted and analysed by bioinformatics software.

Illumina sequencing can be conducted with single-read or paired-end sequencing. Single-read sequencing involves sequencing fragments from only one end, whereas paired-end sequencing allows sequencing of both ends of the fragments.

Another alternative short-read sequencing technology is Ion Torrent, which is sometimes preferred for its affordability but compromises on accuracy[186]. Instead of using a solid surface or fluorescent imaging, Ion Torrent uses emulsion PCR to amplify fragments on micro-sized diameter beads. Natural nucleotides are applied in a similar step-wise process to Illumina's cycle, but instead change the pH of the solution which are detected by electronic sensors.

1.5.4.2 Long-read sequencing

Alternative long-read sequencing technologies are available for metagenomic sequencing. Pacific Biosciences (PacBio) technologies use single-molecule real-time (SMRT) sequencing and can sequence much longer strands of between 10-60 kb compared to Illumina machines with a typical read length of up to 250 bp[187]. DNA polymerase molecules are attached to the bottom of 50 nm-wide wells called zero-mode waveguides (ZMWs)[188]. There are typically tens of thousands of ZMWs on a SMRT Cell. Each ZMW is illuminated from below, but the wavelength is too large for it to efficiently pass through and instead attenuated light penetrates the lower part of each ZMW. A DNA polymerase performs DNA synthesis on the stand that is immobilised to the bottom of the ZMW, where it is exposed to the light. Modified nucleotides enter the chamber. A fluorescent dye is attached to the phosphate chain of each nucleotide having four different colours depending on the nucleotide (A, C, G or T). When a nucleotide is added to the DNA at the bottom of the ZMW during DNA synthesis, a light pulse is produced of a particular wavelength depending on the fluorescent dye, and is detected. Once the nucleotide is incorporated, the phosphate chain is cleaved, and the fluorescent dye is released. This process repeats across these ZMWs in parallel. The longer read

lengths allows researchers to assemble more complete genomes *de novo* (Section 1.6.4.1)[187]. However, PacBio technologies are more costly than short-read sequencing. Extracted DNA libraries are not amplified so PacBio requires a greater amount of extracted DNA. Thus, metagenomic DNA needs to be pooled together from multiple samples[189]. This is far more challenging for metagenomic samples with lower levels of microbial DNA, such as saliva.

Alternatively, Oxford Nanopore technology can routinely sequence fragments of hundreds kb long. It is generally more affordable than both short-read and PacBio technologies, but has a higher nucleotide error rate[190]. Unlike other sequencing platforms, Nanopore technology monitors changes to the electrical current as nucleic acids are passed through arrays of protein nanopores. Like PacBio, DNA does not require amplification, eliminating PCR bias, and sequencing data can be streamed in real time. Due to the length of its reads, as well as its speed, affordability and portability, Nanopore technology is having ground-breaking success in large-scale WGS of organisms and pathogens anywhere. For example, it has been applied clinically as a real-time diagnostic tool for AMR pathogens from bloodstream infections[191] and the surveillance of the Ebola virus[192]. Like PacBio, long read lengths allow for longer *de novo* genome assemblies that are able to resolve repeated regions of sequences (Section 1.6.4.1). However, in order to achieve a higher accuracy, more reads are required from a single organism to provide enough read depth to overcome the error rate[190]. This can be achieved by using amplification or DNA concentration techniques inspired from Illumina sequencing that can be integrated into the Nanopore workflow before sequencing[190]. Accuracy can also be improved by sequencing 2D reads, where both

strands of the template strand and its complement strand are sequenced, instead of sequencing 1D reads, where only the template strand is sequenced. However, different protocols can influence the yield dramatically and should be optimised depending on the sample or desired outcome. Although flow cells can be reused to increase throughput with the aim of increasing yield, they are prone to degradation following washing and carryover of DNA from previous runs, which can actually reduce yield and accuracy of coverage depth. In order to profile resistomes from complex metagenomic samples, such as faecal samples, DNA amplification or DNA concentration techniques are still required to detect ARGs prior to sequencing[193]. Generally, resistome studies have relied on metagenomic data generated from short-read technologies. However, long-read technologies are increasingly useful for linking ARGs with MGEs and their host species (discussed in Section 1.7.2).

1.5.4.3 Computational processing

After sequencing, reads are recorded in text files that are then processed for further analysis. Firstly, any sequences of adapters that were added in the library preparation are removed and sequences trimmed. Reads are then quality controlled using automated software to measure metrics like base quality, GC content, sequence length distribution and adapter content[194]. In the context of metagenomics of microbial communities, human DNA contamination is removed using mapping software[195]. From hereafter, bioinformatics tools are applied to profile the resistome from whole metagenomes.

1.5.5 Bioinformatic methods of profiling the resistome

Bioinformatic software tools can be applied to identify genes including genetic determinants of AMR, such as ARGs, from files of whole metagenomic reads. There are two approaches to identifying ARGs from metagenomic data: finding known ARGs that have been previously recorded in reference databases or predicting new ARGs *de novo*. Identifying known ARGs may be most relevant for studies that are interested in ARGs known to cause phenotypic resistance relevant in clinic outcomes. In other cases, making predictions of candidate ARGs from whole metagenomes may be preferred over finding known ARGs from reference catalogues. For instance, predicting what ARGs may be missing from reference databases may need to be validated using targeted laboratory approaches, like PCR[196].

1.5.5.1 Reference-based detection

To identify known ARGs, reads are mapped to ARG reference databases such as the Comprehensive Antibiotic Resistance Database (CARD)[197], ResFinder[198], ARG-ANNOT[199] and MEGAres[200], or general databases that contain ARGs like NCBI[201]. Reads can be mapped using an alignment-based algorithm, commonly Bowtie2[195] or BWA (Burrow-Wheeler Alignment)[202], or using a *k*-mer[ix] counting-based algorithm, such as KMA[203], which matches the coverage of *k*-mer frequencies between query and reference sequences. *k*-mer-based mapping is more precise at distinguishing between ARGs from databases that contain ARGs with sequence similarity and redundancy, as it finds exact matches between sequences[203]. However, alignment-based algorithms have the advantage of being able to estimate the absolute abundance of an ARG in a sample

ix A *k*-mer is a small sequence of length *k*. (Included in Glossary)

based on the raw number of reads that are aligned, which can be applied to calculating the difference in abundances between samples[204]. Instead, the *k*-mer-based alignment algorithm KMA calculates the total number of *k*-mers mapped to each nucleotide of an ARG divided by the number of base pairs of that ARG, which is a less accurate estimate for ARG abundance.

Alternatively, metagenomic reads may be assembled into longer contiguous sequences using assembly tools like metaSPAdes[205] or MEGAHIT[206] before ARG identification. This allows other referenced-based methods to be used which rely on longer query sequences. BLAST[207] and DIAMOND[208] are very commonly used alignment tools for this purpose. Hidden Markov Models (HMMs)[209] are probabilistic models of multiple sequence alignments of proteins and are commonly used to detect remote homologies between protein sequences. Searches of HMMs are sometimes used for identifying proteins that have structural or functional similarities that are difficult to detect from sequencing alignment alone. Before using HMMs, the nucleotide sequence must be translated into amino acid sequences using tools like Prodigal[210]. Unlike mapping reads, it is not possible to quantify the abundance of ARGs from assemblies, as a metagenomic assembly algorithm will aim to generate a single representation for each genome that are distinct from each other. However, longer assemblies allows annotation of other genetic elements surrounding ARGs to inform its genetic context, for example, whether an ARG is part of an MGE (Section 1.7.2).

One major challenge with using reference databases is that they are incomplete. This is particularly pertinent for antifungal resistance genes, more so than for antibacterial

resistance genes. Although the most serious fungal infections tend to be more opportunistic, occurring when the human immune system is compromised, research into fungal infections has lagged behind bacterial infections[211,212]. Consequently, there has been a bias towards characterising and curating antibacterial over antifungal resistance genes in microbial ecology. Although there are a plethora of antibacterial resistance gene databases available, only one exists for antifungal resistant genes. MARDy is an antifungal resistance database that consists of 36 genes with 232 amino acid substitutions causes by nonsynonymous polymorphisms[x] [213]. However, as mentioned above, many fungal sequences tend to be underrepresented in whole metagenomes that have been isolated using extraction kits attuned to isolating bacterial and viral DNA, meaning antifungal resistance genes are difficult to identify from whole metagenomes[168]. Given the lack of curated antifungal resistance gene databases, there is a motivation to characterise antifungal resistance genes by sequencing genomes of phenotypically resistant strains from culture[214].

1.5.5.2 De novo discovery

It is possible that ARGs in microbiomes, particularly in unculturable microbes, have distant homologies to known ARGs in some databases, like CARD. One study used a three-dimensional (3D) structure-based method to predict over 6,000 genetic AMR determinants in the intestinal microbiome[196]. The authors developed 3D structures of known ARPs and used homology comparative modelling of the structures to predict related ARPs. They then attempted to validate their predictions using pairwise

x Nucleotide substitutions in genes that lead to a change in amino acid

comparative modelling. This is based on the concept that their active sites[xi] would be more conserved between proteins that are functionally similar. Although they were related by their 3D structure, these ARPs were not closely related to known ARPs by their amino acid identity. Although they could not validate whether these ARPs were functional *in vivo*, this allowed them to hypothesise the existence of unknown ARGs.

1.5.6 Future trends

There has been a paradigm shift towards using short-read and long-read sequencing for: 1) detecting the presence of genetic AMR determinants, like ARGs; 2) monitoring how resistant pathogens spread; and 3) how these strains evolve, which is beginning to be translated from single isolates to metagenomes. Although WGS of cultured isolates remain the gold-standard for characterising novel, putative genetic resistance determinants, metagenomic approaches are beginning to predict ARGs with notable accuracy. For example, a recent study used a Nanopore MinION to sequence metagenomes to determine genetic determinants of resistance and susceptibility of the resident pathogen *Streptococcus pneumoniae* using genomic neighbour typing (a computational method to predict the phenotype of a pathogen by their closest relatives using *k*-mer content of reads)[215].

Metatranscriptomics is an alternative method of predicting the phenotype based on measuring the ARG expression[216]. Instead of sequencing DNA, mRNA transcripts that are generated from DNA transcription of the cell are sequenced from whole

xi An active site of an ARP is the region which binds to the substrate as part of a chemical reaction that leads to an AMR phenotype.

metagenomes without relying on culture-based methods. However, the expression profiles only give a snapshot of the current phenotypic state. Many ARGs are only expressed or increase in expression under certain conditions, notably while being exposed to their target antimicrobial. Also the phenotype may be altered by post-translational modifications. Some ARGs that are acquired may not be expressed at all as they confer a loss of fitness to the host, such as having to synthesise a novel protein under a less effective metabolic pathway[151].

Molecular approaches, like metagenomics, cannot supersede phenotypic testing to measure the susceptibility of pathogens to antimicrobials. Nevertheless, with rapidly developing sequencing technologies and computational capabilities, whole metagenomics is a promising avenue for predicting potential ARGs, which can be verified with functional metagenomics[179]. In addition, more genomic datasets are becoming publicly available. These datasets are accessible for researchers on limited budgets to undertake data-driven work to generate novel, but plausible, hypotheses that can persuade funders to invest in validation studies with tangible results.

1.6 Profiling the mobilome

Text, tables and figures have been copied directly from "Probing the mobilome: Discoveries in the dynamic microbiome", Carr et al., 2020[217]. Section, figure and citation numbering have been modified and abbreviations included to agree with thesis format. To conform with the structure of this thesis, the section titled "Mobilome composition", "Box 1" and "Outstanding Questions", have been removed as they are

covered in Section 1.3. Technical terminology is defined in the footnotes and Glossary. The references for this paper are included in the full reference list of this thesis. The published paper is available online: https://doi.org/10.1016/j.tim.2020.05.003

1.6.1 Abstract

There has been an explosion of metagenomic data representing human, animal and environmental microbiomes. This provides an unprecedented opportunity for comparative and longitudinal studies of many functional aspects of the microbiome that go beyond taxonomic classification, such as profiling genetic determinants of AMR, interactions with the host, potentially clinically relevant functions and the role of MGEs. One of the most important but least studied of these aspects are the MGEs, collectively referred to as the mobilome. Here we elaborate on the benefits and limitations of using different metagenomic protocols, discuss the relative merits of various sequencing technologies, and highlight relevant bioinformatics tools and pipelines to predict the presence of MGEs and their microbial hosts.

1.6.2 Introduction

The shift to high-throughput sequencing technologies in microbial genomics has radically changed our understanding of microbial communities in different habitats. The appreciation of the complexity of these communities is now undergoing a further shift as more publicly available microbiome datasets based on shotgun metagenomic sequencing are becoming available. As well as establishing the taxonomy and relative abundance of microbial populations, these datasets are allowing individual genes and

their variants to be characterised, including ARGs. MGEs are critical to our understanding of how genes (and their associate functions) move via HGT within a community[218]. These elements can have a lasting impact on the composition of microbial communities, affecting their diversity and density, as well as their interaction with the environment[219]. The profile of these MGEs (mobilome) is thus likely to be a key player in influencing selection pressure-driven changes in the composition of microbial communities and their impact on the host organism or tissue. MGEs are also responsible for the movement of genetic AMR determinants and virulence factors between microbes[181]. For example, the use of antimicrobials can increase the prevalence of MGEs carrying functioning ARGs that are integrated in microbial genomes[220]. Profiling the mobilome and its associated ARGs can provide insights into how ARGs move across multiple genomes within the microbiome. To characterise the mobilome in a microbial community, all MGEs sequences need to be identified from metagenomic data and ideally would be assigned to a microbial host. Although detecting MGEs from single isolates using WGS is a common approach that is significantly more straightforward, metagenomic sequencing is increasingly being used to detect and classify multiple MGEs from microbial communities.

1.6.3 Targeted metagenomic approaches and challenges in extracting MGEs

Despite having to overcome significant hurdles, metagenomic sequencing of microbial samples is increasingly being used to identify novel MGEs. Both targeted and whole metagenomic methods are now being used to identify and discover novel as well as

known MGEs **(Fig. 1.5.)**. In contrast to whole metagenomic methods where all DNA extracts are sequenced, targeted metagenomics include a step that specifically selects a type of MGE prior to sequencing.

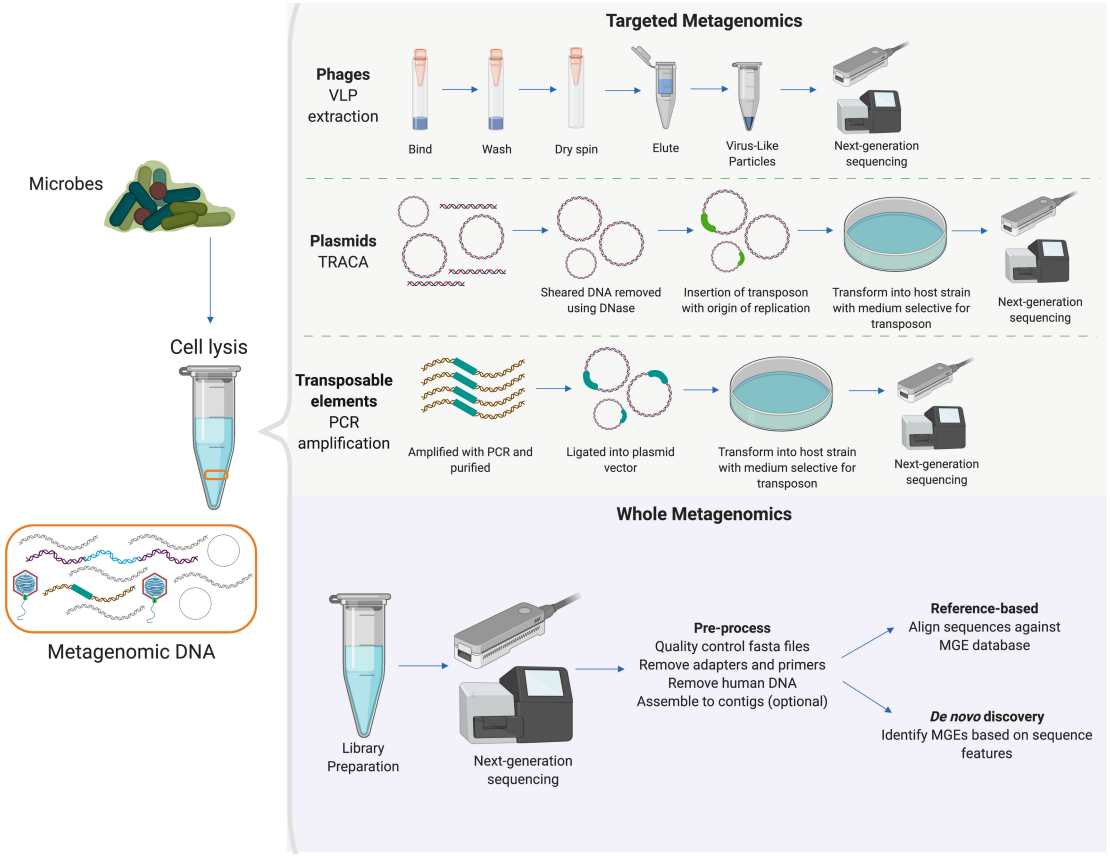

Figure 1.5. Targeted and whole metagenomic technologies for extracting MGEs.

Targeted metagenomic methods currently include purifying MGEs prior to shotgun sequencing. For example, free phage particles, along with other virus-like particles (VLPs), are purified in several stages of physical and/or enzymatic treatments[221–223]. Nucleic acids extracted from VLPs are then sequenced and assembled into contiguous sequences for further annotation[223–225]. Circular plasmids are isolated using high-throughput transposon-aided capture (TRACA) from metagenomic DNA, which are then typically transformed into *Escherichia coli* for cloning[226], followed by shotgun sequencing and PCR-based approaches to close gaps in sequences[227]. However, these

targeted approaches may misjudge the potential MGE load. Inefficiencies in the elution of VLPs from faecal samples have been shown to result in an underestimation of the viral load, and inconsistencies between protocols have led to discrepancies in results between studies[222]. Size-fractionation is an alternative technique involving enrichment of extracted DNA for novel viral particles by filtering the samples through a size exclusion membrane that has been applied to the cow rumen virome[228]. Of 148 viral genera enriched from the cow rumen, 75% had no counterpart in existing viral databases, highlighting the power of this technique to recover phages.

For plasmids, TRACA enriches metagenomic DNA for circular plasmids by using a DNAse that selectively removes linear DNA. Plasmids are subsequently "captured" by inserting a transposon (in an *in vitro* transposition reaction) with an origin of replication and selection marker before transforming them into typically *Escherichia coli* for cloning[226]. This is followed by shotgun sequencing, with additional PCR to close gaps in sequences[227]. However, TRACA has a bias towards capturing smaller plasmids between 3-10 kb, excludes linearised plasmids, and potentially inactivates plasmid genes as a result of transposon insertion[229]. Alternatively, inverse-PCR together with multiple displacement amplification (another DNA amplification technique) has also been applied to identify small circular plasmids in metagenomic samples[230].

Finally, a targeted metagenomic approach using PCR amplification can be used to identify transposable elements by targeting the repeat regions[231]. Metagenomic DNA is amplified by PCR primers targeting transposable elements, purified and ligated into

plasmid vectors, then transformed into host strains. After clonal expansion, the plasmids are isolated, sequenced and annotated for transposable elements.

Targeted metagenomic approaches are highly specific and therefore useful for extracting MGEs with distinct features, such as sequence composition. Given non-targeted MGEs would be excluded, these approaches would not be suitable for determining a more complete representation of MGEs within the whole metagenome. However, recent advances in sequencing technology and data storage mean that whole metagenomic DNA sequencing is now a viable option for investigating the wider pool of MGEs, giving us a better representative picture of the mobilome[232–235].

1.6.4 Whole metagenomics

Whole metagenomic DNA sequencing has great potential for both identifying known and unknown MGEs and also for predicting the MGE hosts. However, there are several limiting factors, specifically with current NGS technologies and bioinformatic software tools, that need to be considered.

1.6.4.1 Challenges in sequencing technologies

The current gold-standard for metagenome sequencing is using short-read sequencing methodologies, specifically Illumina and Ion Torrent technologies. Since short-read metagenomic sequencing produces reads that are too short to allow the identification of plasmids, phages and transposable elements, many bioinformatic pipelines involve assembling the metagenomic reads into longer contiguous sequences called contigs.

However, assembling metagenomes is computationally intensive, and the choice of assembly tool has a significant impact on the accuracy of identifying MGEs[236–238]. Dealing with the microbial complexity of a metagenome with limited read depth and repeated regions is a challenge for current assembly algorithms. These tools are prone to generate erroneous inter-species chimeric contigs when processing complex metagenomic sequence datasets. Thus, plasmid and transposon contigs are often inaccurate or incomplete. Different plasmids often contain similar replication and conjugative elements[102], whilst transposable elements contain repeated regions[239]. For phages, assembly of short reads has further challenges including a high incidence of repeat regions and/or hypervariable regions[240], genetic diversity[241], frequent modular structures[242], and heterogeneity at strain level[240,243]. To circumvent these issues, many metagenomic assemblers attempt to produce shorter, less complete but more accurate contigs rather than longer, inaccurate ones. A direct consequence of this is that metagenomic contigs are often too short to accurately predict large MGEs.

Long-read sequencing technologies (such as Oxford Nanopore and PacBio's SMRT sequencing), produce longer sequence reads, meaning it is possible to more accurately assemble much longer scaffolds and even complete genomes. Nanopore technology, for example, has been used to successfully recapitulate complete viral genomes from metagenomes[244,245]. However, the sequences generated contain more erroneous bases than short-read technology sequences due to technical defects in base calling[190,246]. PacBio has a higher accuracy rate in single-nucleotide and structural variants, but produces shorter reads than Nanopore and is more costly[187,247]. In addition, the limits in coverage depth from a run on a single Nanopore flowcell is a bottleneck for identifying

lower abundant MGEs in metagenomes with high microbial diversity[190]. However, it is possible to improve and even complete the assembly of MGEs from complex whole metagenomes using an ensemble of short-read and long-read sequencing technologies[248].

1.6.4.2 Bioinformatic methods in MGE sequence annotation

When analysing microbiome composition, isolation and sequencing of DNA forms only part of the story – the subsequent computational analysis is every bit as important. This is also the case when mining sequencing data for MGEs and other genetic elements. Although advances in technology have markedly improved the accuracy of whole metagenomic sequencing, accurate and efficient bioinformatics software is required to resolve MGEs from a complex pool of fragmented microbial genomes.

Typically, genomic sequence features are identified broadly either by reference-based or *de novo* methods, or a combination of both. Reference-based methods generally use alignment algorithms, such as BLAST[207], to align query nucleotide or amino acid assemblies against a reference database or search tools against probability sequence models, such as HMMER for HMMs[249]. Non-MGE-specific nucleotide sequence databases, such as RefSeq[250], and protein sequence databases, like Pfam[251] and UniProt[252], have been applied to detect HGT events in metagenomes[253,254]. Virus-specific sequence databases have more recently been established, such as the Prokaryotic Virus Orthologous Groups (pVOGs)[255], curated viral databases from RefSeq, PATRIC[256] and IMG/VR[257]. Databases suited for searching transposable elements in metagenomic assemblies include ISfinder for insertion sequences[258], and ICEberg for ICEs and

IMEs[259]. PlasmidFinder is a popular database for identifying plasmids that contains plasmid replicon sequences from *Enterobacteriaceae* and gram positive bacteria[103]. In all cases, MGE-containing databases contain a very narrow representation of the mobilome with incomplete coverage of element types, and do not reflect the actual MGE diversity. For instance, transposable elements are one of the most ubiquitous and genetically diverse elements in the microbiome[239,260], making cataloguing all of them an intractable task. Despite this obvious limitation, well-curated reference databases can be useful for discovering novel MGEs as they are often used in benchmarking new *de novo* bioinformatics tools[261].

Despite their utility, MGE reference databases do not include all MGEs in existence. Further, it is difficult to find novel MGEs that are dissimilar in sequence and structure to known MGEs. Finding these novel MGEs requires the use of *de novo* bioinformatics methods and tools to make predictions based on sequence data. There is a plethora of different algorithms used for discovering putative phages in assembled metagenomes, such as VirSorter[262], VirFinder[263], MARVEL[264], VirMiner[265] and ViraMiner[266] **(Table 1.3)**. Apart from VirSorter that uses primarily HMMs, all these tools apply machine learning to identify viral-like domains. A handful of tools have been developed for identifying plasmid sequences from metagenomes, including cBar[267], PlasFlow[238], Recycler[268] and metaplasmidSPAdes[269] **(Table 1.3)**. Machine learning approaches are also used in cBar and PlasFlow to predict linear and circular plasmids. Other non-machine learning-based tools, Recycler and metaplasmidSPAdes, identify plasmids using De Bruijn graph assembly of *k*-mers. metaplasmidSPAdes also includes a naïve Bayesian classifier on custom plasmid-specific profile-HMMs to improve its accuracy.

For discovery of insertion sequences, only two *de novo* pipelines have been developed using existing algorithms to identify direct repeats and palindromic inverted terminal repeats **(Table 1.3)**[270].

When designing and building bioinformatic tools, it is valuable to benchmark them for specificity and sensitivity. For MGE identification tools applied to metagenomes, the ideal dataset for benchmarking predictions would include labels of known MGEs within real metagenomic sequences. Aside from VirMiner and metaplasmidSPAdes, these tools have not been benchmarked using representative metagenomes. Since these ground truth datasets are difficult to obtain, many of these tools were benchmarked using simulated metagenomic sequences generated from a representative set of genomes from the most abundant species of a microbial community. Therefore, it is likely that when these tools are applied to complex whole metagenomic samples, they would not perform as well as their stated accuracy would suggest.

Table 1.3. Published tools for *de novo* MGE discovery intended for whole metagenomes.

MGE	Tool	Authors and Year	Data Type	Search algorithm	Advantages	Disadvantages
Insertion sequence	Pipelines: Two *de novo* and one profile HMM search	Kamoun et al., 2013[270]	Raw fragments	*De novo* "Repeat search": RepeatScout algorithm [271] *De novo* "inverted repeat search": palindrome software of the EMBOSS package[272] Profile HMM: MUSCLE[273] and HMMER2 package[209]	*De novo* methods do not rely on incomplete ISfinder database Profile HMM search performs significantly better than BLAST on simulated and real metagenomic datasets	Repeat search had high false positive rate inverted repeat search has lower true positive rate Repeat search and inverted repeat search not tested on metagenomic datasets
Bacteriophage	MARVEL	Amgarten et al., 2018[264]	Raw fragments in metagenomic bins	Random forest machine learning	Better sensitivity and similar specificity to VirSorter and VirFinder	No option in software to retrain on alternative training data Only tests algorithm on simulated metagenomic bins Does not consider prophages
	VirSorter	Roux et al., 2015[262]	Contigs	Prediction of circular sequences[274] Gene predicting using MetaGeneAnnotator[275] HMMER3 for pHMMs and BLASTP for unclustered proteins	Prediction of novel prophages from reference-independent prediction of viral domains	Not tested on metagenomics of whole microbial communities, only viral metagenomes Does not have complete prophage prediction, as optimised for assemblies of fragments
	VirFinder	Ren et al., 2017[263]	Raw fragments	*k*-mer-based Logistic regression model with lasso regularisation machine learning	Outperforms VirSorter Do not need to assemble metagenomes before using tool	Model limited to learning from training data before 1st January 2014 so may not be appropriate for recently discovered viral sequences, and no option in software to retrain on alternative training data, Only tests algorithm on simulated metagenomes Need to filter out eukaryotic host sequences, as may mis-classify as viral
	VirMiner	Zheng et al., 2019[265]	Raw fragments	Random forest machine learning on phage contigs	Validates algorithm and compares with VirSorter and VirFinder using metagenomic data from human gut samples. Better sensitivity than and similar specificity to VirSorter and VirFinder Also extends the pipeline to include raw read processing and assembly, sequence and functional annotation of phage contigs, and phage-host prediction using CRISPR-spacer recognition, and two-group comparison (e.g. case and control) User-friendly website	Does not have a command-line or API tool, making it difficult to analyse multiple metagenomes No option in software to use alternative tools in pipeline or retrain random forest on alternative training data
	ViraMiner	Tampuu et al., 2019[266]	Contigs	Deep Learning using Convolutional Neural Networks	Model can be retrained on alternative data unlike MARVEL or VirFinder	Does not directly compare performance against other tools The accuracy of the model on human metagenomic contigs is likely to be an overestimate because reference-based alignment is used to benchmark these contigs that would likely contain false negatives
Plasmid	Recycler	Rozov et al., 2016[268]	Raw fragments	Circular de Bruijn graphs with coverage filters	Even though lack of metagenome benchmark, tool compares plasmid prediction from cow rumen metagenomic data[276] with plasmids extracted using PCR validation from a previous study[230]	Ignores linear plasmids, and those integrated in chromosomes Performance metrics, i.e. precision and recall, only calculated from applying to a simulated plasmidome Only 35% of plasmid predictions from metagenomes matched plasmids reported in PCR validation

Continues next page

cBar	Zhou and Xu, 2010[267]	Contigs	Sequential minimal optimization-based model on pentamer frequencies	First tool that attempts to distinguish plasmids from chromosomal DNA from whole metagenomes	Achieves 88.29% accuracy with independent test setbut does not describe how the independent test set was generated Does not attempt to bin plasmids
PlasFlow	Krawczyk et al., 2018[238]	Contigs	Machine learning model trained using a deep neural network on genome signatures	Outperforms cBar on plasmidome data	Compares PlasFlow to cBar, Recycler and PlasmidFinder on whole metagenomes, but could not evaluate performance Assemblies required to be longer than 1 kb
metaplasmidSPAdes	Antipov et al., 2019[269]	Raw fragments	Circular assembly graphs with coverage filters. Includes a verification tool, plasmidVerifty, which uses a naive Bayesian classifier on plasmid-specific profile-HMMs	plasmidVerify outperforms cBar and PlasFlow annotation of custom plasmid and non-plasmid sequences from RefSeq Generally identifies more plasmids than Recycler using metagenomic data, mock data, multiple genomic isolates and plasmidome data	Ignores linear plasmids

1.6.5 Technological challenges in host prediction of MGEs

Identifying the microbial hosts of different MGEs will be central to developing our understanding of how MGEs shape microbial communities and *vice versa*. However, this is problematic for a variety of reasons, not least of which is our limited ability to find the specific microbial origin of MGEs in metagenomic samples. As technologies move forward, additional approaches such as wet-lab protocols and bioinformatics tools are being applied with both short and long-read metagenomic sequencing to link MGEs with their host microbe.

1.6.5.1 Wet-lab technologies for microbial host prediction

Although associating genetic elements with individual organisms within a community initially seems insurmountable, there are promising laboratory-based techniques that can be exploited. Some of these can make use of features of different sequencing technologies, whilst other methods require pre-processing of samples prior to sequencing. Binning reads into groups prior to computational assembly is probably the simplest of these techniques. As SMRT sequencing can be applied to identify the methylation status of a nucleotide **(Fig. 1.6a)**, metagenomic reads can be binned into species or subspecies based on methylation motifs[277]. SMRT sequencing can be applied to identify the methylation status of a nucleotide **(Fig. 1.6a)**. Sequences are then clustered into groups based on the similarity of multiple methylation motifs. These motifs are usually shared by both chromosomes and plasmids within a microbe but are often unique to a microbial strain. However, as microbial communities become more

complex, the methylation motifs become less unique as it becomes more likely that more than one strain or species contains the same motif.

An alternative approach is the use of proximity ligation methodologies, specifically Hi-C **(Fig. 1.6b)**[278]. DNA molecules in close proximity in the genome's three-dimensional structure are covalently bonded together. Thus MGEs that are in close proximity to their host genome are covalently bonded to the host genome. These connected sequences are then digested around the bond and ligated to form a continuous strand with ligation junctions. After this proximity ligation, the DNA is fragmented and sequenced as usual. Sequence information regarding these ligation junctions is used in downstream computational analysis pipelines to assign assembled metagenomic reads to their host microbe species. Hi-C has been used alongside short-read metagenomic sequencing to link plasmids to their hosts with strain-level resolution in synthetic metagenomes[279] and species-level resolution in real metagenomic communities[280,281]. However, Hi-C has limited resolution capabilities for closely related organisms due to their high sequence similarity and uneven Hi-C link densities[282]. Proximity ligation has also been used to link phages to species from cattle rumen metagenomes[283]. Since proximity ligation relies on the three-dimensional structure of the host genome only, phages that do not integrate into the genome as prophages are largely undetected by this process. However, single-cell viral tagging with short-read metagenomic sequencing is an alternative approach specifically for predicting the hosts of both lytic and lysogenic phages[284].

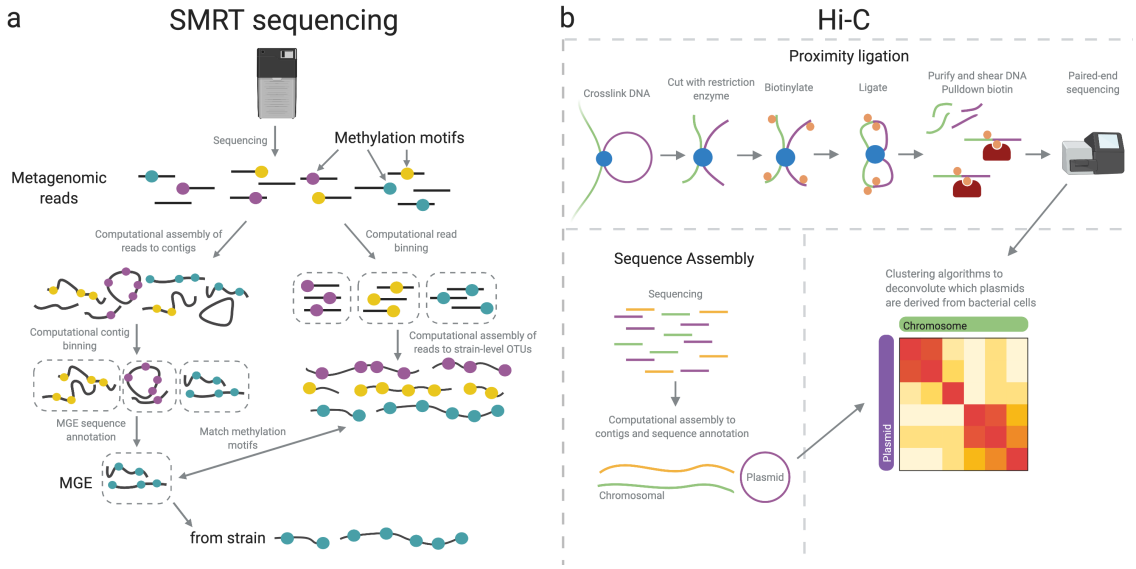

Figure 1.6. Wet-lab protocols for microbial host identification of MGEs (applicable to plasmids and prophages) using a) SMRT sequencing and b) Hi-C.

1.6.5.2 Bioinformatic methods in microbial host prediction

Metagenomic reads and contigs containing MGEs and host genomes can be binned into groups using computational as well as wet-lab methods, allowing for two levels of identification and discrimination. There are many different algorithms for metagenomic binning, including analysing sequence composition features and coverage, sequence signature properties, *k*-mer frequencies and gene co-abundance across samples[235,285–291]. However, these binning algorithms, particularly gene co-abundance, can be computationally intensive.

An approach that can link MGEs with their hosts relies on distinct MGE sequences also found in microbial genomes[292–294]. When an MGE enters a bacterium, the bacterium uses a defence mechanism of Clustered Regularly Interspaced Palindromic Repeats (CRISPR). Fragments of the MGE sequence, known as spacers, are integrated between CRISPR loci in the bacterial genome. These spacers are transcribed into small RNA

molecules and processed into a ribo-protein complex which targets and destroys invading genomes. The hosts of these MGEs can then be predicted by aligning the predicted MGE contigs against a reference database of candidate host genomes containing CRISPR spacers. This method has been previously used to identify phage and plasmid hosts in human gut metagenomes[93,295]. However, since many of these reference databases are incomplete, it may only be possible to assign a small proportion of MGE contigs to a host[292].

1.6.6 Conclusions and further perspectives

In general, there is currently no single sequencing, wet-lab or bioinformatics technique for whole metagenomes that can efficiently profile the entire mobilome and its microbial context. As we have shown here, employing a combination of approaches is the best solution to classifying novel MGEs and assigning these and known MGEs to their host microbes. In order to resolve longer MGEs such as plasmids and phages whilst maintaining accuracy, the ideal approach is to use a combination of short-read and long-read sequencing. Highly accurate short metagenomic reads can be assembled and scaffolded against more complete but less accurate contiguous sequences from long-read sequencing. Identifying the microbial hosts of the MGEs presents further problems. However, SMRT long-read sequencing used in combination with proximity ligation on short-read sequencing is a complementary approach that can be applied to all MGE types and will allow for association of these elements to host genomes with a reasonably high degree of certainty.

Having generated these sequences, many different bioinformatic methods can be highly effective at identifying and classifying MGEs in these sequences accurately, or binning MGEs with host sequences from the acquired metagenomic data. The bioinformatic tools listed are not evaluated computationally in this review, but cited reviews and papers have done so for tools identifying phages and plasmids[265,296]. Due to a rapid software developments, it is likely some tools outlined here will already be superseded by the time of publication, with one or a few tools that have been iterated and become standard. Popular approaches, such as machine learning, will still be important tools. However, a tool that has a high accuracy on simulated metagenomes may not perform well on real metagenomes and could be computationally expensive. Therefore, researchers will need to critically evaluate which tool is most suitable for their particular requirements.

There is no single correct solution for characterising the mobilome. The performance of bioinformatics tools for *de novo* discovery is limited by the data quality which is dependent on the sequencing platform. Current sequencing technologies for whole metagenomes fall short of the levels required for a truly accurate and fully representative analysis of the mobilome. However, there is cause for optimism. The recent development of new methodologies, such as proximity ligation and SMRT sequencing technologies, means that we are rapidly evolving our ability to not only identify potential MGEs, but also to associate them with their host genomes. As these technologies improve, so too will bioinformatic tools be developed to make full use of these new datasets, and thus provide us with a more complete picture of the mobilome and how it spreads genetic elements through microbial communities.

1.7 The mobile resistome of the human gastrointestinal tract using whole metagenomics

The human GIT consists of the human tissues of the mouth, oesophagus, stomach and intestines, and host microbiomes that carry ARGs and MGEs[297,298]. The different sites of the gut (stomach and intestines) and oral cavity (mouth) are inhabited by quite disparate microbial communities[15,299,300]. As a consequence of being anatomically connected and being part of the same food processing pathway, the GIT sites can influence each other's microbial compositions[13,301]. Thus, ARGs and MGEs can mix between these sites as well. However, the gut and oral cavity have different environments with different exposure to ingested factors, including anthropogenic antimicrobial exposure. Although it is widely accepted that the gut and oral cavity harbour reservoirs of ARGs and MGEs[298,302], there has been very little research into the compositions of the oral resistome and mobilome using whole metagenomics in particular, and how these compare to the gut. In addition, there has been very little discourse about what impact the GIT mobile resistome has on AMR infections and clinical outcomes.

1.7.1 The resistome

There has been a recent flurry of metagenomic resistome studies focussing on the human gut niches (stomach and intestines) of the GIT. In fact, the first whole metagenomic study to profile the resistome of healthy individuals worldwide was conducted on stool samples[303]. A total of 507 distinct ARG types conferring resistance to 20 antibiotic classes were detected across 180 stool samples from 11 different countries,

raising the status of the gut as an AMR reservoir. As a result, there has been more emphasis on whole metagenomic resistome surveys in the gut than in the oral cavity, even though the first functional metagenomics study to query ARGs was in the oral microbiome, which led to the discovery of *tet(37)*[304]. Larger functional metagenomic studies of human oral microbiota revealed more ARGs conferring resistance against tetracycline (including a novel ARG *tet(32)*) gentamycin and amoxicillin[305,306]. This thesis builds upon this work to include the first large-scale study to profile the oral resistome using whole metagenomics (Chapter 2)[307]. However, other whole metagenomic studies of the gut resistome have provided insights into the dynamics of the GIT resistome in early and adult life, and during different exposures to various types of antimicrobial drugs.

1.7.1.1 The GIT resistome in early life

Profiling the gut resistome in preterm infants found the presence of clinically relevant ARGs suggesting that pioneering microbes of early human microbiome already have capacity for AMR[308]. These microorganisms may originate from environments rich in microbes that have been selected for AMR by anthropogenic antimicrobial exposure, in particular hospitals where many neonates are born, and potentially kept in intensive care units[309]. In addition, resistant microbes can transfer from other humans in close contact where their resistomes are more established (Section 1.7.1.3). For instance, there is a transfer of microbes containing ARGs from maternal milk to the infant's gut that drives diversity of the gut resistome in early life[310]. No metagenomic studies have been conducted on infant oral resistomes, but given their exposure to the environment, the oral cavities of newborns are likely colonised by ARG-carrying microbes as well.

1.7.1.2 Adult GIT resistome

Once the microbiome stabilises, the resistome then remains relatively stable, both in the gut and the oral cavity[307,311]. The first study to profile gut resistomes worldwide also found that differences in microbial phylogeny between individual guts also co-localised with differences in resistome composition[303]. The microbial composition varies considerably across the GIT, which is also influenced by factors discussed earlier in Section 1.1.1[15]. Certain microorgansisms may harbour particular ARGs that benefit their survival, meaning different GIT sites may have different resistomes. One study comparing the oral and gut resistomes of uncontacted Amerindians found all ARGs, apart from tetracycline resistant ARGs coding for ribosomal protection proteins, were exclusively found in either oral or gut whole metagenomes, but not both[312]. However, these ARGs were chosen by their phenotypic resistance to antimicrobials by functional metagenomics and not by profiling the entire metagenomic dataset. Early metagenomic studies using targeted microarray probes found ARGs conferring resistance to tetracycline, with *tet(M)* and *tet(W)* most prevalent in saliva and stool samples, respectively[313].

1.7.1.3 The GIT resistome after antimicrobial intervention

Direct anthropogenic use of antimicrobials in hospitalised infants reveals an increase in the diversity of ARGs in the gut resistome, carriage of multidrug-resistant *Enterobacteriaceae* and changes in the microbiota composition that all persist long-term[314]. Exposure to antimicrobials may have a profound impact in infants with

developing microbiota communities that have not yet reached fruition, particularly for infants on rudimentary diets. In the adult gut the effects of lower chronic and higher acute doses of antimicrobials can have a varied impact on the microbiota and resistome composition in the short- and long-term. For instance, a cocktail of three last-resort antimicrobials (meropenem, gentamicin and vancomycin) were orally administered to healthy adult men for four days leading to nine common species being undetected in gut metagenomes after six months[43]. Species harbouring β-lactam, glycopeptide and aminoglycoside ARGs colonised the gut in the long-term after the exposure. Although these findings reflect mild long-term effects in the microbiome and resistome of healthy adults, this study used a short course of multiple broad-spectrum antimicrobials, which are not typically administered clinically or as a complete course of treatment. Lower levels of chronic exposure to antimicrobials in livestock farming can perturb the microbiota and shape the resistome in guts of temporary farm workers[311]. These effects on the microbiota are predicted to partially reverse after four months, but more ARGs were still detectable after six months. Compared to the gut microbiota, a week course of broad-spectrum antibiotics (clindamycin, ciprofloxacin, minocycline or amoxicillin) in 66 individuals revealed a greater recovery of the salivary microbiota after 12 months[315]. Narrow-spectrum as well as broad-spectrum antimicrobials can also drive changes of the resistome in the gut more than the oral cavity. A study looking into the longitudinal resistome of the stool samples and the oral swabs from young children found an increase in ARG diversity in stool samples a month after the administration of narrow-spectrum penicillin V administration, but not in the oral cavity[42]. Authors studying the effect of the saliva and gut resistomes from broad-spectrum antibiotics suggest that the oral microbiota has better intrinsic resistance to antimicrobial exposure than the gut[315].

This may be the case for oral biofilms that consist of extracellular polymeric substances (EPSs), which can exclude antibiotics from readily targeting bacteria residing deep within the biofilm[316]. Further, high concentrations of antimicrobial compounds naturally produced by microorganisms may accumulate within the EPS, selecting for localised resistance. However, the authors of this study make a peculiar assumption that the exposure to antimicrobials in both sites are the same. In fact, different body sites have variations in amount and time of exposure to orally administered drugs (the conventional route of administration for the majority of antibiotic treatments) as governed by pharmacodynamics, with the gut having a greater bioavailability than the oral cavity. One reason why the gut microbiome and resistome may be less resilient to antimicrobials than the oral cavity, is that the gut is exposed more to orally administered antimicrobial drugs than the oral cavity. These drugs pass rapidly through the oral cavity before entering the stomach and intestines where they remain for much longer periods at high concentrations before being absorbed into the bloodstream, metabolised or excreted. This extra exposure time and greater antimicrobial bioavailability in the gut, particularly in the small intestine, can lead to increased selection of antimicrobial resistant microorganisms which may acquire resistance through mutation or via HGT. Selection then presents a prime opportunity for resistant strains to proliferate as part of the microbiota during recolonisation, unless a fitness cost prevents this[151], explaining the greater resistome diversity in the gut compared to the oral cavity.

1.7.2 The mobile resistome

The major types of MGEs that capture, accumulate and disseminate ARGs are insertion sequences and composite transposons, gene cassettes/integrons, plasmids, ICEs and

bacteriophages, as mentioned above. There are very few studies that use whole metagenomic sequencing to identify ARGs on MGEs. Most whole metagenomic studies focussing on associations between the resistome and mobilome have been done in microbiomes from the human gut or polluted/contaminated water[310,317–320]. A handful of studies have applied whole metagenomics to identify ARG-MGE associations in the oral cavity of the human GIT.

Studies that have attempted to profile mobile resistomes from whole metagenomes have done so in several different ways. One study working with short-read metagenomic sequences from a polluted lake inferred potential ARG-MGE associations by identifying ARGs with putative MGE proteins (rather than a whole MGE) on the same contig, such as identifying the *RepC* plasmid protein with fluoroquinolone resistance gene *qnrS*[320]. Another study reported a transposon carrying an integron with a *sul1* gene, an *aadA* gene cassette (resistant to sulphonamides, spectinomycin and streptomycin) and a *dfrA17* gene cassette (resistance to trimethoprim) encompassed in a plasmid from assemblies of whole metagenomes from an infant stool sample[319]. Although it is possible to ascertain ARG-MGE associations from metagenomic assemblies from a single sample, it is challenging to quantify the prevalence of ARG-MGE associations with short-read whole metagenomics alone. To address this, a study implemented correlation analysis of abundances between ARGs and MGEs across multiple samples in whole short-read metagenomic sequences of planktonic microbial communities in the river Han[317]. The authors found ARG density was significantly correlated with the abundance of integrases and verified ARGs were located on the same contigs as *intI1* integrons.

One major challenge for all these studies is the difficulty in assembling contigs from short-read sequences that are long enough to cover an MGE or resolve repeated sequences in an MGE, especially ICEs, transposable elements and plasmids. A solution to this as mentioned previously is use long-read sequences that resolve repeated regions as scaffolds for short-read sequences that retain accuracy. For instance, one study combined reads from Illumina short-read and Oxford Nanopore MinION long-read sequencing of wastewater microbiomes[318]. This increased the accuracy and length of metagenomic contigs, allowing more plasmids and ICEs containing ARGs to be detected with high confidence and for ARG-carrying ICEs to be tracked across the wastewater treatment process. In addition, the contigs were long enough that the bacterial hosts could be identified for the ICEs, although this could not be done for plasmids. Proximity ligation methods, such as Hi-C, can be applied to assign ARG-carrying MGEs to microbial hosts from metagenomes[321]. Hi-C has been used to monitor the dynamics of MGEs between bacteria in longitudinal stool metagenomes from neutropenic patients (with abnormally low levels of neutrophils) taking multiple courses of antibiotics[105]. Here, this technique improved the accuracy of associations between taxa and ARG-carrying MGE as compared to metagenomic assemblies alone and showed HGT is more frequent in neutropenic patients than in controls taking no antibiotics.

In general, the lack of comprehensive mobilome and associated resistome profiles in whole metagenomes are down to the challenges of resolving large MGEs, incomplete MGE reference databases for reference-based MGE identification and limited

bioinformatic tools for *de novo*-based MGE sequence annotation as discussed previously[217]. In addition, many studies focus on profiling one or two types of MGEs. For example, studies focussing on ARG-bacteriophage associations rarely consider other MGEs[97,322,323]. As a consequence, the relative associations between ARGs and different types of MGEs are not fully understood. However, there is a huge scope to combine and develop bioinformatic methods to profile the entire mobilome, even from short-read metagenomic data. This would help us understand what proportion of the resistome is associated with different types of MGEs, or at a more granular level, how different types of ARGs are associated with particular MGEs.

1.7.3 Opportunities and challenges in using whole metagenomics

It is difficult to give an accurate description of the mobile resistome across the human GIT, especially in healthy adults, since very few whole metagenomic studies have been conducted and compared across GIT sites. Although there are a vast number of whole metagenome studies profiling faecal samples, it is contested whether faecal samples give an accurate representation of the stomach or intestines of the GIT. The microbial DNA in stool samples are over-represented by microbiomes in the lumen of the intestines and not the mucosal surface, failing to capture structural resolution[324]. As well as in the gut, the oral cavity consists of many different ecological niches, including hard, non-shedding surface of the teeth, and shedding mucosal surfaces, such as tongue, cheek and the soft and hard palate[300].

As these studies in the gut resistome and mobile resistome flood the academic literature, many scientists starting research in the human microbiome may see more potential to work with the gut than other body sites. It is the case that researchers with limited funds to initiate a new human microbiome study would be more inclined to use freely available metagenomic datasets or collaborate with other researchers where funding is available. Given that most human metagenomic datasets and studies focus on stool samples, this may only serve to perpetuate the idea of the gut being central to discussions on the human microbiome. However, the microbiome and metagenomics fields are still in their infancy compared to other disciplines in genomics and there is hope that innovative approaches can alleviate this bias.

1.8 Objectives

There is an unmet need to address the associations between the resistome and the mobilome across GIT sites and how these can affect clinical outcomes of AMR. Therefore the specific aims of this thesis were:

I. **Profiling and comparing resistomes between oral sites and stool whole metagenomes (Chapter 2)**

II. **Profiling and comparing mobilomes and mobile resistomes between oral sites and stool whole metagenomes for bacteriophages (Chapter 3), plasmids (Chapter 4) and transposable elements (Chapter 5)**

III. **Evaluating the contribution of bacteriophages, plasmids and transposable elements to the HGT of ARGs (Chapter 6)**

Firstly, I compare the resistome composition between oral cavity sites and gut using publicly available whole metagenomic data of buccal mucosa (cheek), dental plaque, dorsum of the tongue (top surface of the tongue), saliva and stool from different countries around the world.

I then compare the mobilome and mobile resistomes between GIT sites from the same publicly available whole metagenomic data. Given the huge variety of MGEs but the very limited research in whole metagenomes, this study focusses on profiling common

MGEs: bacteriophages (Chapter 3), plasmids (Chapter 4) and transposable elements (Chapter 5) from whole metagenomic sequences. Metagenomic analyses of these other ARG-carrying MGEs, like gene cassettes/integrons, are beyond the time afforded in this thesis, but nonetheless, have important roles to play in sequestering of ARGs in microbial genomes. In Chapter 3, I compare the composition of bacteriophages from whole metagenomes across GIT sites. I then analyse the diversity of phages and their bacterial hosts, phage stability, and function and phylogeny of large, rare jumbo phages across GIT sites. The prevalence of ARG-carrying bacteriophages is compared across GIT sites. In Chapter 4, whole metagenomic data are assembled into circular plasmid contigs to make a catalogue of plasmids using an existing software tool. I then construct a bioinformatics pipeline to query the metagenomes for these plasmids. Similar to phages and transposable elements, the composition of plasmids and ARGs identified from these plasmids is compared between GIT sites. Chapter 5 describes the development of a new software tool and bioinformatics pipeline to identify transposable elements (in particular insertion sequences, composite and unit transposons) *de novo* from whole metagenomic data. The composition of transposable elements and ARGs identified from them are compared across GIT sites.

Finally, in Chapter 6, I collate the results from the previous four chapters to compare and contrast how each type of MGE (bacteriophages, plasmids and transposable elements) associates with the resistome. I then discuss how the HGT of these ARG-carrying MGEs might affect AMR infections and clinical outcomes. Lastly, I consider how this research can be continued and expanded to profile and characterise the resistome and mobile resistome.

Chapter 2: The Resistome

2 The Resistome

2.1 Introduction to study

Whole metagenomic data has been successfully applied to characterise resistomes in humans, particularly of human gut resistomes[87,308,310,325–327], but comparatively little whole metagenomic research has been conducted for oral resistomes and how this relates to the gut. This is the first study profiling oral resistomes with a direct comparison to gut resistomes from the same individuals using whole metagenomes. It is known from functional metagenomic studies that the oral cavity contains a diversity of ARGs, such as those conferring resistance to tetracycline, amoxycillin and gentamicin in saliva and plaque samples[304,328]. The presence of ARGs from smaller whole metagenomic studies of oral cavities from isolated Amerindian communities and ancient humans (with little exposure to anthropogenic antimicrobials), indicates the presence of ARGs is in fact an inherent characteristic of the oral microbiome[312,329]. Yet these ARGs can have a serious impact on the clinical outcomes of AMR infections. ARG-carrying oral streptococci have been shown to spread to other body sites and cause AMR infections, such as infective endocarditis[330].

Both the gut and oral cavity contain a reservoir of ARGs. However, little is known about how the resistome composition differs between each other. In this study, resistomes were profiled from 788 oral and 386 paired gut whole metagenomes in healthy individuals taken from pre-existing studies in China[331], Fiji[332], the Philippines[333], Western Europe[301,334,335] and the USA[336]. The gut resistome was profiled from stool

metagenomes, whereas the oral resistome was characterised from multiple oral sites, including buccal mucosa, dental plaque, dorsum of the tongue and saliva. The main findings are that the oral resistome contains the highest and lowest abundances of several ARGs compared to gut, and the gut and the surface of the tongue contained the highest ARG diversity. Although no metadata on antimicrobial use was available for each individual, it is hypothesised that the differences in resistome composition between GIT sites are driven by variations in microbial composition and differences in antimicrobial exposure. Finally, it is proposed that ARGs that are more abundant in the oral cavity compared to the gut could be highly persistent and potentially important in clinical outcomes of AMR infections in other body sites.

2.2 Published paper

Text, tables and figures have been copied directly from "Abundance and diversity of resistomes differ between healthy human oral cavities and gut", Carr et al., 2020. Section, figure and citation numbering have been modified and abbreviations have been included to agree with thesis format. The Methods section has been moved before the Results section for clarity. More detailed explanations of methodology and additional methodology have been included in the Methods, footnotes and Appendix 2A-C. Supplementary Materials of this paper are in Appendix 2D-M. The references for this paper are included in the full reference list of this thesis. The published paper is available online: https://doi.org/10.1038/s41467-020-14422-w

2.2.1 Abstract

The global threat of AMR has driven the use of high-throughput sequencing techniques to monitor the profile of resistance genes, known as the resistome, in microbial populations. The human oral cavity contains a poorly explored reservoir of these genes. Here we analyse and compare the resistome profiles of 788 oral cavities worldwide with paired stool metagenomes. We find country and body site-specific differences in the prevalence of ARGs, classes and mechanisms in oral and stool samples. Within individuals, the highest abundances of ARGs are found in the oral cavity, but the oral cavity contains a lower diversity of resistance genes compared to the gut. Additionally, co-occurrence analysis shows contrasting ARG-species associations between saliva and stool samples. Maintenance and persistence of AMR is likely to vary across different body sites. Thus, we highlight the importance of characterising the resistome across body sites to uncover the AMR potential in the human body.

2.2.2 Introduction

In recent years, AMR has been highlighted as one of the biggest threats to global health, food production and economic development[52]. Given this rapidly developing global crisis, it is imperative that the current gaps in our understanding of the distribution, spread and associations of all AMR factors are filled. AMR is most often conferred through the expression of ARGs that reduce a microbe's susceptibility to the effects of an antimicrobial compound. As such, monitoring the abundance and diversity of these ARG profiles, or the resistome, has huge potential to increase our understanding of the spread and persistence of AMR within a population. High-throughput NGS technologies

are beginning to be used as tools for screening ARGs for potential surveillance of AMR worldwide. Shotgun metagenomic data mapped against dedicated ARG reference databases are providing a wealth of insight into the resistomes of human[87,308,310,325–327] and animal guts[85,337] as well as the wider environment[179,320,338,339]. However, no large studies have, to date, attempted to characterise the resistome profiles of the human oral cavity. Commensal microbes from the oral cavity harbouring ARGs have potential to lead to antimicrobial resistant infections at other body sites. For example, ß-lactam, clindamycin, and erythromycin resistant strains of oral streptococci have caused infections at distal body sites such as infective endocarditis[330].

Metagenomic studies of the oral cavity indicate that this site potentially contains a diverse range of ARGs, including those encoding resistance to tetracycline, amoxycillin and gentamicin in saliva and plaque samples[304,328]. Thus, oral ARGs appear to be natural features of the human oral cavity. The presence of an oral resistome containing aminoglycoside, ß-lactam, macrolide, phenicol and tetracycline ARGs in isolated Amerindian communities and ancient humans, indicates that the presence of these genes is not dependent on antibiotic exposure and is an inherent feature of the oral microbiome[312,329].

The oral microbial community faces unique ecological pressures, such as mechanical force, nutritional availability, pH levels, oxidative stress and redox potential. Despite these continually changing conditions, these communities have been shown to be relatively stable even after short-term antibiotic exposure. HGT has been documented as an important mechanism for the transfer and acquisition of ARGs within and between

oral bacterial species[139,340]. The erythromycin resistance *mefA* and *mefE* genes have been found on the *MEGA* MGE associated with Tn*916*-like conjugative transposons (also ICEs), and this has been implicated in conjugative transfer between viridans group streptococci (VGS) and other streptococci[341]. Thus, the oral microbiome contains a long-standing and mobile population of ARGs and is a significant reservoir for ARGs to be transferred to pathogenic microbes.

Here, we derive and compare the oral and the gut resistomes from 788 and 386 shotgun metagenomes, respectively, from healthy individuals from China[331], Fiji[332], the Philippines[333], Western Europe[301,334,335] and the US[336]. We found country-specific differences in the proportion of saliva, dental plaque and stool samples containing ARGs, ARG classes and mechanisms. We made up to 415 comparisons of oral resistomes with paired gut resistomes derived from stool shotgun metagenomes from the same individuals, showing the oral resistome contains the highest and lowest abundances of ARGs, but a lower diversity of ARGs than the gut resistome. Overall, these results demonstrate the requirement for wider AMR surveillance studies at different body sites, including the oral cavity, to understand the composition of the resistome across different human microbial habitats.

2.2.3 Methods

2.2.3.1 Metagenomic sequence data

A total of 1,174 publicly available metagenomic samples covering the USA, China, Fiji, the Philippines and Western Europe (France and Germany), all sequenced using

Illumina HiSeq 2000, were analysed. Longitudinal USA samples were excluded from the majority of the study after the first time point to ensure each sample was independent, unless specified otherwise. All metagenomes passed over half the quality control metrics in FastQC 0.11.3 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) with these pass rates calculated in MultiQC[194]. These samples include 1) longitudinal data across two years with various timepoints from the Human Microbiome Project 1 (referred to as USA)[336] containing buccal mucosa (n = 87: 32 with one, 36 with two, 18 with three and 1 with six timepoints); dorsum of tongue (n = 91: 22 with one, 43 with two, 24 with three and 2 with four timepoints); dental plaque (n = 90: 23 with one, 43 with two, 20 with three, 1 with four and 3 with six timepoints); stool (n = 70: 13 with one, 33 with two, 21 with three, 2 with four and 1 with six timepoints), 2) healthy control samples from a Chinese rheumatoid arthritis study[331] containing dental plaque (n = 32); saliva (n = 33); stool (n = 72), 3) saliva (n = 136) and stool (n = 137) samples from Fiji[332], 4) saliva samples (n = 23) from healthy hunter-gatherers and traditional farmers from the Philippines[333], and 5) saliva (n = 21) and stool (n = 21) samples from Western Europe (5 saliva and 5 stool samples from Germany[301,334], and 16 saliva and 16 stool samples from France[301,335]).

Raw paired-end metagenomic reads from Chinese and Philippines samples were downloaded from EMBL-EBI (https://www.ebi.ac.uk/metagenomics/). Paired-end metagenomic samples from USA were downloaded from https://portal.hmpdacc.org/. Raw paired-end metagenomic reads from Fiji (project accession PRJNA217052 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA217052/]), France and Germany (project accession PRJEB28422 [https://www.ncbi.nlm.nih.gov/bioproject/?

term=PRJEB28422]) were downloaded from the NCBI. All US, China, Fiji and Philippines samples, and stool samples from France and Germany, were collected and sequenced as described in the following cited studies[331–336]. Saliva samples from France and Germany were collected and sequenced as described in the following cited study[301]. Metadata for the samples can be found in Supplementary Data 2.1 available from https://tinyurl.com/y3yxng3q.

2.2.3.2 Processing metagenomic data

The raw reads for all samples were trimmed using AlienTrimmer 0.4.0[342] using default parameters and Illumina contaminant oligonucleotides (https://gitlab.pasteur.fr/aghozlan/shaman_bioblend/blob/18a17dbb44cece4a8320cce8184adb9966583aaa/alienTrimmerPF8contaminants.fasta). Human contaminant sequences were removed from all samples by discarding reads that mapped against a human reference genome (downloaded from Human Genome Resources at NCBI on 27th February 2017) using Bowtie2 2.2.3[195] with parameters *-N 1 -k 1 --end-to-end --very-sensitive –phred33 --no-discordant*[xii]. The quality of the raw reads and the filtered reads of each sample was evaluated using the FastQC 0.11.3 (https://github.com/s-andrews/FastQC).

2.2.3.3 Identifying ARGs

All processed metagenomes were mapped against *nucleotide_fasta_protein_homolog_model* from the ARG database CARD 3.0.0[197] using

---

xii
N is the number of mismatches allowed in a seed alignment during multiseed alignment. Multiseed alignment is where Bowtie2 aligns substrings (or seeds) from the read and its reverse complement.

k is the number of distinct, valid alignments for each read. The search terminates when the algorithm cannot detect anymore than k alignments.

--*end-to-end mode* is where Bowtie2 aligns the read from one end to the other, without trimming characters off either end of the read.

--*very-sensitive* is a preset of default parameters which makes the algorithm more sensitive and accurate but slower. The only value change is parameter N.

--*phred33* is an option that specifies the encoded quality scores used in Illumina sequences.

--*no-discordant* specifies that Bowtie2 allows both mate pairs to align uniquely but that they do not need to satisfy paired-end constraints.

KMA 1.2.6. The selection of an ARG database is described in **Appendix 2A**. Hits were identified where the template coverage was greater than 90%. The metagenomes were mapped against these hits using Bowtie2 2.2.5 with parameter *--very-sensitive-local*[xiii]. The choice of mapping tools is addressed in **Appendix 2B** and the choice of coverage threshold is explained in **Appendix 2C**. Mapped reads were filtered from unmapped reads, sorted and indexed using Samtools 1.9[343]. Statistics for the number of reads mapped for each ARG were identified using Bedtools 2.28.0[344].

2.2.3.4 Abundance of ARGs

The reads per kilobase of read per million (RPKM) was calculated for every sample as the number of reads divided by the total number of library reads per million, then divided by the gene length in kilobases. The relative abundance of ARGs for each country and sample type was calculated by dividing the RPKM by the sum of RPKM for each country and sample type. The relative abundance of ARGs for each sample and sample type was calculated by dividing the RPKM by the sum of the RPKM for each sample. Differential abundance of ARGs between paired sample types from each country were calculated using the DESeq2 1.20.0 package[204] as recommended by Jonsson et al.[345]. ARGs that were significantly differentially abundant (adjusted p-value < 0.05) across study cohorts for paired sample comparisons were identified using a meta-analysis random effects model with the metafor 2.1-0 package[346].

xiii *--very-sensitive-local* is the preset of default parameters in local mode where Bowtie2 is not required to align the entire read from one end to the other (unlike end-to-end mode).

2.2.3.5 ARG class abundance and antibiotic prescription rate

The mean RPKM and standard error for every ARG class was compared against the antibiotic prescription rate measured as the DDD Per 1,000 individuals in 2015 from China, the Philippines, Western Europe (France and Germany) and the USA. This data was derived from ResistanceMap (https://resistancemap.cddep.org/) accessed on 19th February 2019. No antibiotic use data was available for Fiji. Linear regression was conducted on the log transformed mean RPKM versus the DDD Per 1,000 for all ARG classes.

2.2.3.6 Percentage of samples with ARGs, ARG classes and mechanisms

To show whether the percentages of samples containing an ARG class were consistent across the same number of reads, metagenomes were first subsampled using seqtk 1.2 (https://github.com/lh3/seqtk) with parameter seed *-s100.* 6.9 million reads were subsampled from 18 saliva samples with the lowest number of reads greater than 6.9 million reads, from each cohort: China, Fiji, the Philippines and Western Europe. 18 million reads were subsampled from 18 dental plaque with the lowest number of reads greater than 18 million reads from both China and USA cohorts. 16.9 million reads were subsampled from 18 stool samples with the lowest number of reads greater than 16.9 million reads from China, Fiji, the USA and Western Europe cohorts. These were mapped to CARD 3.0.0 as described in *Identifying ARGs.* R 3.5.1 was used for all downstream analysis. Each ARG was annotated with Drug Class and Resistance Mechanism using CARD 3.0.0 metadata. Percentages of samples containing an ARG,

ARG class and mechanism were calculated from these samples. 95% confidence intervals (CIs) were evaluated from percentages identified from bootstrapping samples 20 times for each cohort and sample type[xiv].

2.2.3.7 Principal Coordinates Analysis

790 metagenomes that contained at least 1 millions reads and were not longitudinal USA samples were first subsampled to 1 million reads using seqtk 1.2 (https://github.com/lh3/seqtk) with parameter seed *-s100.* These were mapped to CARD 3.0.0 as described in *Identifying ARGs*. Principal Coordinates Analysis was applied to the binary distance between ARG presence or absence profiles for each sample (excluding longitudinal USA samples) using the vegan 2.5-2 package (https://cran.r-project.org/web/packages/vegan/index.html). Principal Coordinates Analysis represents and visualises dissimilarity between data points between a set of uncorrelated axes in lower dimensional, Euclidean space. Resistotypes were identified using hierarchical clustering of the Euclidean distance between principal coordinates with eigenvalues above zero. An eigenvalue of an axis has a magnitude representing the amount of variation captured in that axis. Silhouette analysis was used to determine the optimum number of resistotypes using the cluster 2.0.7.1 package (https://cran.r-project.org/web/packages/cluster/index.html). The number of resistotypes is defined by the number of clusters with the largest average silhouette width. Silhouette analysis optimises for a number of clusters by measuring how well an observation is clustered.

xiv Bootstrapping is a resampling method where smaller subsets of samples of the same size are repeatedly drawn (20 times in this case) from the original dataset

The silhouette width of an observation *i* is defined as:

(eq. 2.1) $$S_i = \frac{(b_i - a_i)}{max(a_i, b_i)}$$

where a_i is the dissimilarity between *i* and all other points of the cluster it belongs to, b_i is the average dissimilarity between *i* and all other observations in the nearest cluster it does not belong to. The average silhouette width across observations is calculated for each number of clusters.

2.2.3.8 ARG diversity

To ensure the ARG richness could be compared statistically across different sample types from the same individuals[347], the metagenomes (excluding longitudinal USA samples) were subsampled using seqtk 1.2 with seed *-s100*. Paired samples from the same individuals in each of the following groups containing two sample types were subsampled to a number rounded down by two significant figures from the lowest number of reads in the group. *China dental plaque vs. saliva:* 3.5 million reads were sampled from China dental plaque (n = 31) and paired saliva (n = 31) samples. *China stool vs. saliva:* 3.5 million reads were sampled from China stool (n = 31) and paired saliva (n = 31) samples. *China stool vs. dental plaque:* 14 million reads were sampled from China stool (n = 30) and paired dental plaque (n = 30) samples. USA *buccal mucosa vs. dental plaque:* 1 million reads were sampled from USA buccal mucosa (n = 78) and paired dental plaque (n = 78) samples. *USA buccal mucosa vs. dorsum of tongue:* 1 million reads were sampled from USA buccal mucosa (n = 86) and paired dorsum of tongue (n = 86) samples. *USA buccal mucosa vs. stool:* 1 million reads were sampled from and USA buccal mucosa (n = 64) and paired stool (n = 64) samples. *USA*

dental plaque vs. dorsum of tongue: 4.2 million reads were sampled from USA dental plaque (n = 89) and paired dorsum of tongue (n = 89) samples. *USA dental plaque vs. stool:* 4.2 million reads were sampled from USA dental plaque (n = 68) and paired stool (n = 68) samples. *USA dorsum of tongue vs. stool:* 14 million reads were sampled from USA dorsum of tongue (n = 67) and paired stool (n = 67) samples. *Fiji saliva vs. stool:* 1.2 million reads were sampled from Fiji saliva (n = 132) and paired stool (n = 132) samples. Fiji samples containing fewer than 1.2 million reads were excluded from the analysis. *Western Europe saliva vs. stool:* 3.1 million reads were sampled from Western Europe saliva (n = 21: 5 from Germany and 16 from France) and paired stool (n = 20: 5 from Germany and 16 from France) samples. These were mapped to the CARD database as described in Methods Identifying *ARGs*.

Once the metagenomes were subsampled, ARGs identified and filtering by coverage, the ARG diversity per sample was measured as the ARG richness, recommended previously by Bengtsson-Palme et al.[60]. For every sample, the ARG richness was calculated as the number of unique ARGs. To account for multiple ARGs coding for an efflux pump complex, the ARG richness was calculated excluding ARGs that regulate or are part of an efflux pump complex. The ARG richness between samples in each group was tested for statistical significance with a Mann-Whitney, paired, two-sided test.

2.2.3.9 Correlation analysis

MetaPhlAn2 2.6.0[349] was used to identify taxonomic composition from all samples. Spearman's correlation was applied to relative abundances of reads mapped to ARG and MetaPhlAn2 species profiles for paired samples. ARGs and species that were not found in more than half of samples for each country were removed, to alleviate the bias from potential joint ranking of zero values by Spearman's rank. The rho[xv] and p-values were calculated using the *stats* package in R and the p-values were adjusted with Benjamini-Hochberg where FDR < 5%[xvi]. Correlations were found from China saliva and paired stool samples, and Philippines saliva samples. No significant correlations could be found from Fiji, Western Europe or USA samples.

2.2.3.10 Data availability

ARG data, figures and tables are available at https://github.com/blue-moon22/resistomeData. Data underlying Figures 2.1-2.5 and Appendix 2D-L (Supplementary Figures 2.1-2.9 in paper) are provided in the Source Data file (https://tinyurl.com/y4nbpne4).

xv rho is defined as:

$$rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where d_i is the difference between two ranks of each observation and n is the number of observations.

xvi The Benjamini-Hochberg procedure is described as follows.
Rank the p-values in ascending order, from smallest to largest. The largest p value that satisfies the inequality below is significant, and *all other p* values smaller than it are also significant.

$$p < \frac{i}{m} a$$

where i is the rank, m is the total number of tests, and a is the FDR (in this case 0.05). The FDR is the false discovery rate which is the proportion of significant results that are false positives.

2.2.3.11 Code availability

R package for resistome analysis is available at https://github.com/blue-moon22/resistomeAnalysis. The script to run the analysis is available at https://github.com/blue-moon22/resistomeData.

2.2.4 Results

2.2.4.1 Country and body site-specific differences in resistomes

To establish the incidence of ARGs in oral as well as stool metagenomes collected from various regions, metagenomes were mapped and quantified against CARD[197]. Saliva samples were only available from China, Fiji, the Philippines and Western Europe. To account for the differences in read depths across different datasets, the samples were subsampled to the same number of reads across cohorts for absolute ARG incidence measures. The percentages of saliva samples that contain at least one ARG for each class and mechanism from these cohorts were evaluated. To account for varying read depth across samples, the samples were subsampled to the same number of sequences. Saliva samples from China, Fiji, the Philippines and Western Europe contain 20, 14, 23 and 17 ARG classes, respectively **(Fig. 2.1a)**. Further ARG classes are found in Philippines saliva samples, but most of this variability originates from one individual: a farmer from Zambal who has carbapenem, fosfomycin, rifamycin, and triclosan ARG classes[333]. All or almost all saliva samples from every cohort contain cephamycin, fluoroquinolone, lincosamide, macrolide, streptogramin and/or tetracycline ARGs, and a high percentage (above 50%) of saliva samples from all cohorts contain pleuromutilin ARGs. Unlike most cohorts, all saliva samples from China contain aminoglycoside

ARGs represented by one ARG, *APH(3')-Ia,* and also a high proportion of these samples contain glycylcycline represented by one ARG, *tet(A)* **(Appendix 2: Fig. 2Da)**. The peptide ARG class is only found in saliva from Chinese and Philippines individuals. Mechanisms of AMR including antibiotic efflux, inactivation, target alteration and target protection are present in all saliva samples across all cohorts **(Fig. 2.1b)**, whilst the antibiotic target replacement mechanism is found in China, Philippines and Western Europe but not in Fiji. Reduced permeability to antibiotics is only found in saliva from the same farmer in Zambal.

Dental plaque metagenomic data were only available from China and the USA. The percentages of the China and USA plaque samples containing at least one ARG class and mechanism were compared and found to consist of 16 and 18 ARG classes, respectively **(Fig. 2.1c)**. A greater percentage of Chinese compared to USA plaque samples contain pleuromutilin and/or sulfonamide/sulfone ARGs. Similarly to saliva, all or almost all plaque samples from both cohorts contain lincosamide, macrolide, streptogramin and/or tetracycline ARGs, with a high percentage (above 50%) of these containing cephamycin, fluoroquinolone, glycylcycline and/or pleuromutilin ARGs. Notably, fluoroquinolone and tetracycline ARG classes in dental plaque are comprised of fewer ARGs compared to saliva samples **(Appendix 2: Fig. 2Db)**. Antibiotic efflux, inactivation, target alteration, target protection and antibiotic target replacement mechanisms are present in all samples across both cohorts **(Fig. 2.1d)**.

Stool samples were available from all locations, apart from the Philippines. Stool samples from China, Fiji, the USA and Western Europe were found to contain 31, 30, 17

and 30 ARG classes, respectively **(Fig. 2.1e)**. All or almost all stool samples contain cephalosporin, cephamycin, diaminopyrimidine, lincosamide, macrolide, streptogramin and/or tetracycline ARGs, although most of these ARG classes are found in lower percentages of Fiji stool samples. Compared to oral samples, stool samples contain a lower proportion of the fluoroquinolone ARG class but higher proportions of cephalosporin and/or diaminopyrimidine ARG classes exclusively consisting of *CblA-1* and *dfrF* ARGs, respectively **(Appendix 2: Fig. 2Dc)**. In addition to the USA stool samples containing the fewest number of ARG classes, they also contain a low proportion (less than 50%) of fluoroquinolones, penams (β-lactam with saturated five-membered ring such as penicillin), penems (β-lactam with unsaturated five-membered ring), peptide and/or phenicols ARG classes compared to China, Fiji and Western Europe. Stool samples from China, Fiji and Western Europe contain resistance to triclosan, an antimicrobial that can be found in many household cleaning products, with the highest proportion found in China. Antibiotic efflux, inactivation, target alteration, target protection and antibiotic target replacement mechanisms are present in all samples across all cohorts, but reduced permeability to antibiotics is not found in the USA **(Fig. 2.1f)**.

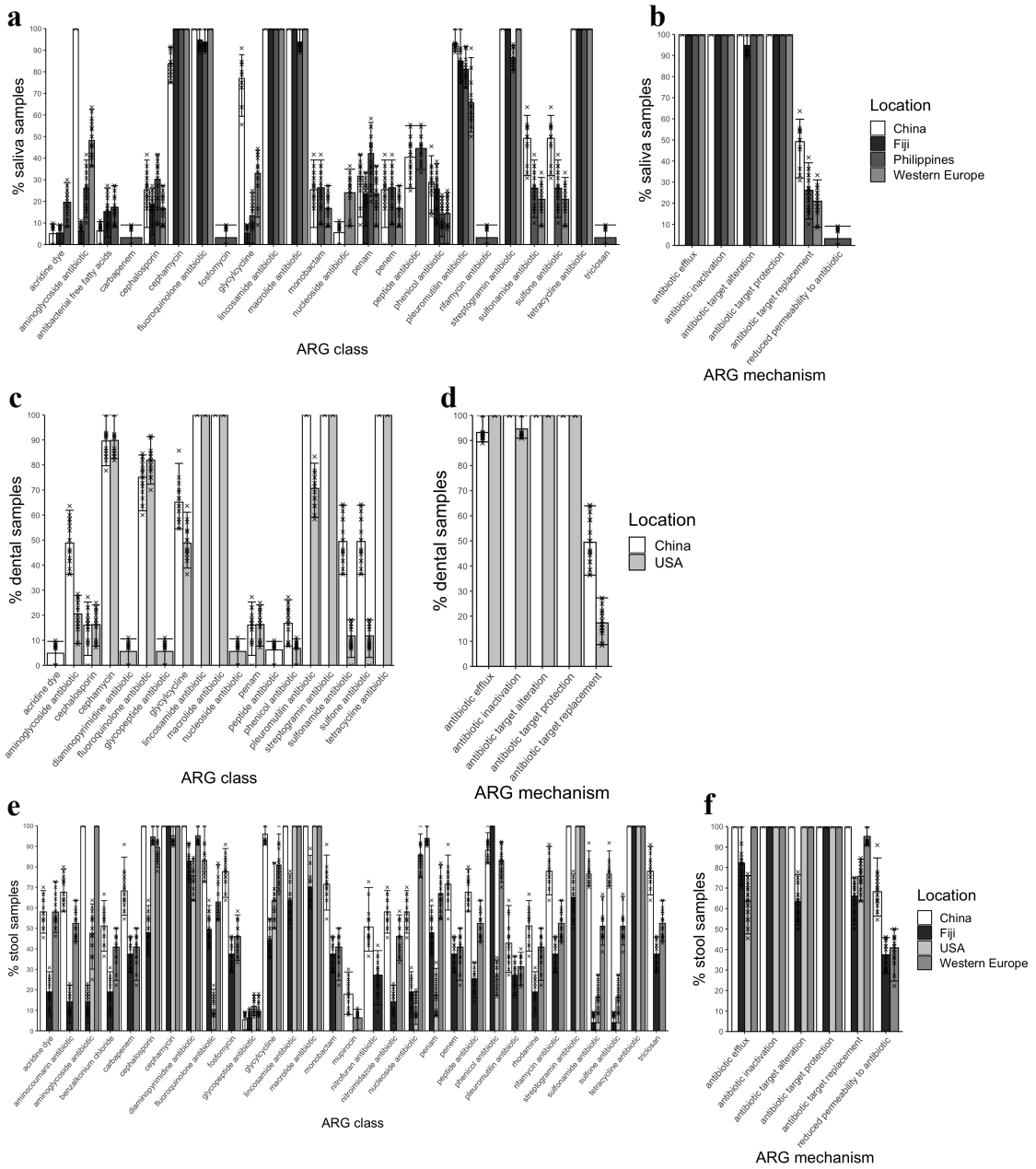

Figure 2.1. Percentage of individuals that contain ARG classes and ARG mechanisms.

Percentage of saliva samples that contain **a)** an ARG class and **b)** an ARG mechanism, of individuals from China (n = 18), Fiji (n = 18), the Philippines (n = 18) and Western Europe (n = 18). Percentage of dental plaque samples that contain **c)** an ARG class and **d)** an ARG mechanism, of individuals from China (n = 18) and the USA (n = 18). Percentage of stool samples that contain, **e)** an ARG class and **f)** an ARG mechanism, of individuals from China (n = 18), Fiji (n = 18), the USA (n = 18) and Western Europe (n = 18). The height of bars are the means and the error bars are 95% CIs of percentages extracted from bootstrapping samples 20 times shown by points.

To determine whether there are differences in ARG composition between oral and gut samples as well as between countries, the ARG incidence profiles for every sample were summarised using Principal Coordinates Analysis and clustered into distinct groups termed resistotypes. Resistotypes were identified using hierarchical clustering and silhouette analysis[350]. Four resistotypes in total were identified **(Fig. 2.2a, b)**. Oral samples are mainly found in two major resistotypes, R2 and R4. R2 is comprised of mainly buccal mucosa and saliva, and R4 contains mainly dental plaque and dorsum of tongue. All stool samples are found in only two major resistotypes, R1 and R3. R1 consists of mostly stool and saliva from Fiji, whereas R3 contains mainly stool from China, Fiji, USA and Western Europe.

To evaluate whether the resistome is related to antibiotic prescription rates, the abundance of ARGs for every ARG class was compared to the Defined Daily Doses Per 1,000 individuals of equivalent drug classes for each region. Prescription data were derived from ResistanceMap (https://resistancemap.cddep.org/). This comparison indicates that overall ARG class abundance does not follow a significant linear relationship with antibiotic prescriptions for any country and body site **(Fig. 2.2c, Appendix 2: Table 2M)**.

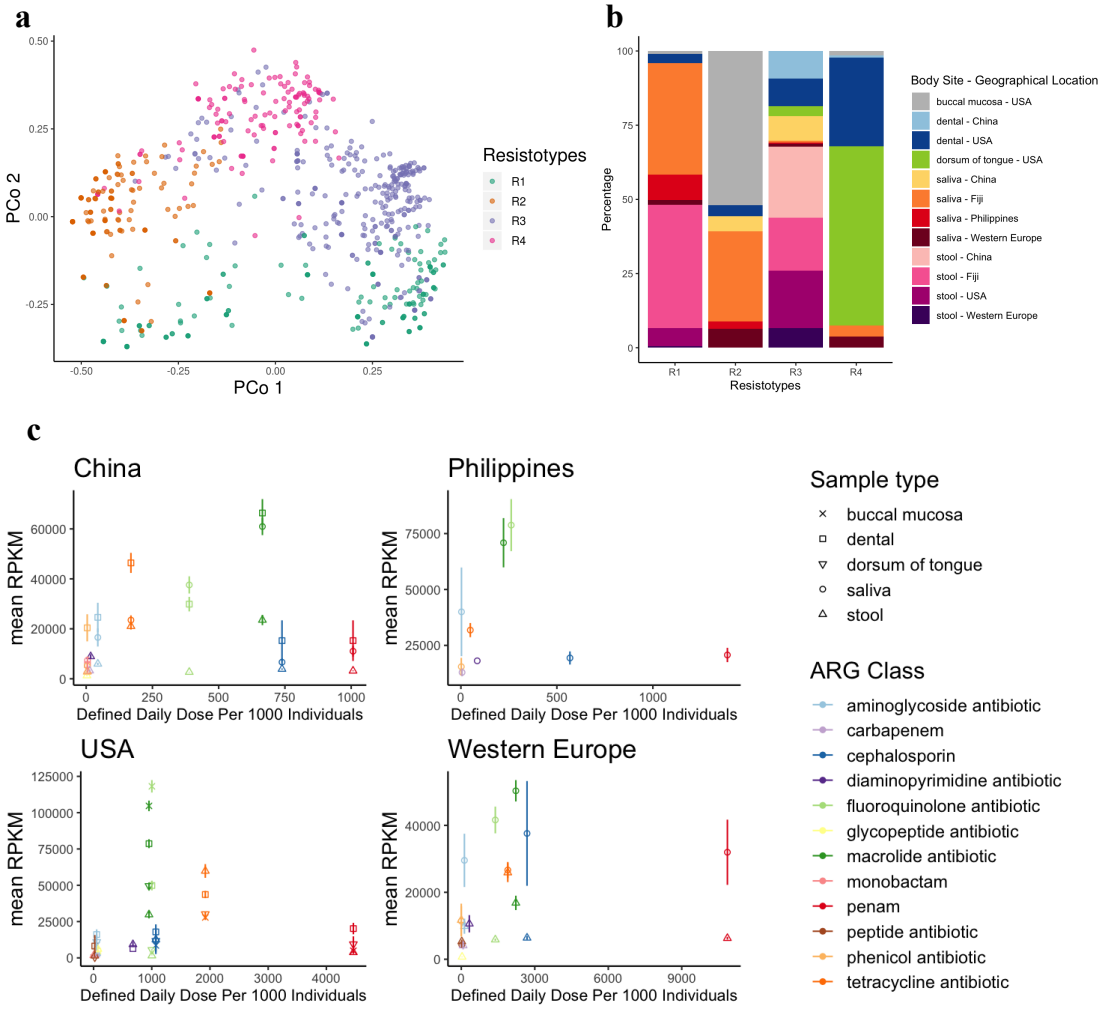

Figure 2.2. Clustering of ARG incidence profiles into distinct groups, and comparing ARG abundance to antibiotic use.

a) Principal Coordinates Analysis of the incidence (presence/absence) of ARGs for all samples where each sample is represented by a point. Samples are labelled as Resistotype clusters, evaluated from hierarchical clustering of binary distance between ARG incidence profiles. Number of clusters was selected with the highest average silhouette width using silhouette analysis. Samples from individuals from China (dental plaque: n = 29, saliva: n = 33, stool: n = 72), Fiji (saliva: n = 129, stool: n = 136), the Philippines (saliva: n = 22), the USA (buccal mucosa: n = 86, dental plaque: n = 80, dorsum of tongue: n = 91, stool: n = 70) and Western Europe (saliva: n = 21, stool: n = 21). **b)** Percentage of Resistotypes that contain samples from a body site and geographical location. **c)** Mean and standard error (error bars) of RPKM of ARGs for each ARG class against the Defined Daily Doses Per 1,000 individuals in 2015 from China, the Philippines, Western Europe (France and Germany) and the USA. (Fiji antibiotic use data unavailable.) Mean RPKM calculated from individuals from China (dental plaque: n = 32, saliva: n = 33, stool: n = 72), Philippines (saliva: n = 23), USA (buccal mucosa: n = 87, dental plaque: n = 90, dorsum of tongue: n = 91, stool: n = 70) and Western Europe (saliva: n = 21, stool: n = 21).

The availability of longitudinal oral and stool samples from USA individuals who had not taken antimicrobial agents over two years afforded us the ability to investigate the stability of resistomes without antibiotics. Hierarchical clustering reveals that the same individuals and body sites cluster together, verifying that resistomes at all sites remain stable over a prolonged period with no antimicrobial selection pressure **(Appendix 2: Fig. 2E)**.

2.2.4.2 ARG composition differs between the oral cavity and gut

To further investigate the differences between oral and gut resistome profiles, the abundance and diversity between oral and gut samples from China, Fiji, USA and Western Europe were evaluated and compared. The total abundance, measured as the total RPKM, of all ARGs in gut samples is lower than in oral (buccal mucosa, dental plaque, dorsum of tongue and saliva) samples across all pairwise comparisons in all cohorts **(Fig. 2.3a)**. The overall abundance is similar between paired USA oral sites, with buccal mucosa and dental plaque having a slightly higher abundance than the tongue dorsum samples **(Appendix 2: Fig. 2F)**. Oral samples contain a higher relative abundance of ARGs coding for fluoroquinolone efflux pumps, lincosamide/streptogramin/pleuromutilin resistance, macrolide efflux pumps and macrolide/lincosamide resistance than stool samples **(Fig. 2.3b, Appendix 2: Fig. 2G)**. These classes are mostly dominated by one of two ARGs across all cohorts, such as *patB* (coding for part of the PatA-PatB efflux pump) in the fluoroquinolone efflux pump class **(Appendix 2: Fig. 2Ha-d)**. Stool contains a higher proportion of tetracyclines and 'Other' ARGs across all cohorts **(Fig. 2.3b, Appendix 2: Fig. 2G).** These 'Other' ARGs are mostly found in aminocoumarins, aminoglycosides, cephalosporins,

diaminopyrimidines, penams, penems and peptides across all cohorts **(Appendix 2: Fig. 2Ha-d)**.

The abundance of individual ARGs were compared between sample types using differential analysis with DESeq2[204]. A meta-analysis strategy was implemented to combine results from all regions. Stool samples are enriched with more ARGs compared to oral samples, but oral samples have enriched ARGs of highest and lowest abundances compared to stool samples across all regions **(Fig. 2.3c, Appendix 2: Fig. 2Ia-k)**. *macB, OXA-85* and *tetA(60)* ARGs in dental plaque have the highest log fold changes (20.8, 21.0 and 21.1 respectively), whilst *patB* and *RlmA(II)* in plaque, and *RlmA(II)* in saliva have the lowest log fold changes (2.9, 2.6 and 2.3 respectively) compared to stool samples **(Fig. 2.3c)**. Highest log fold changes are seen in *cmlA6*, *Lactobacillus reuteri cat-TC*, *macB*, *TEM-1* and *tet32* from saliva (20.6, 20.5,19.8, 20.5 and 20.5 respectively), and *cmlA6*, *macB*, *OXA-85, pmrA, tetA(60)* and *tet(G)* from dental plaque (20.8, 20.6, 20.7, 20.7, 20.8 and 20.8 respectively) compared to stool samples in China that are not enriched across all cohorts **(Appendix 2: Fig. 2J)**. As well as differences between oral and gut, differentially abundant ARGs were found between different sites in the oral cavity **(Appendix 2: Fig. 2I)**. For example, between the USA dorsum of tongue and plaque samples, and between the USA dorsum of tongue and buccal mucosa, all ARGs are enriched in the dorsum of the tongue. Similarly, between dental plaque and buccal mucosa, all ARGs are enriched in dental plaque. Most of these ARGs confer resistance to cephamycin, fluoroquinolone, MLS and tetracycline antibiotics. From China, there are more significantly abundant ARGs in saliva than plaque with resistance to aminoglycoside, cephalosporin, fluoroquinolone (*pmrA* and *patB*), lincosamides,

macrolides (*macB* and *mefA* ARGs), MLS (in particular to the *Erm 235* ribosomal RNA methyltransferase family) and tetracycline antibiotics. Overall, stool samples are enriched with more alternative ARGs and ARG classes compared to oral samples, but with the highest and lowest enrichments of individual ARGs originating from oral samples.

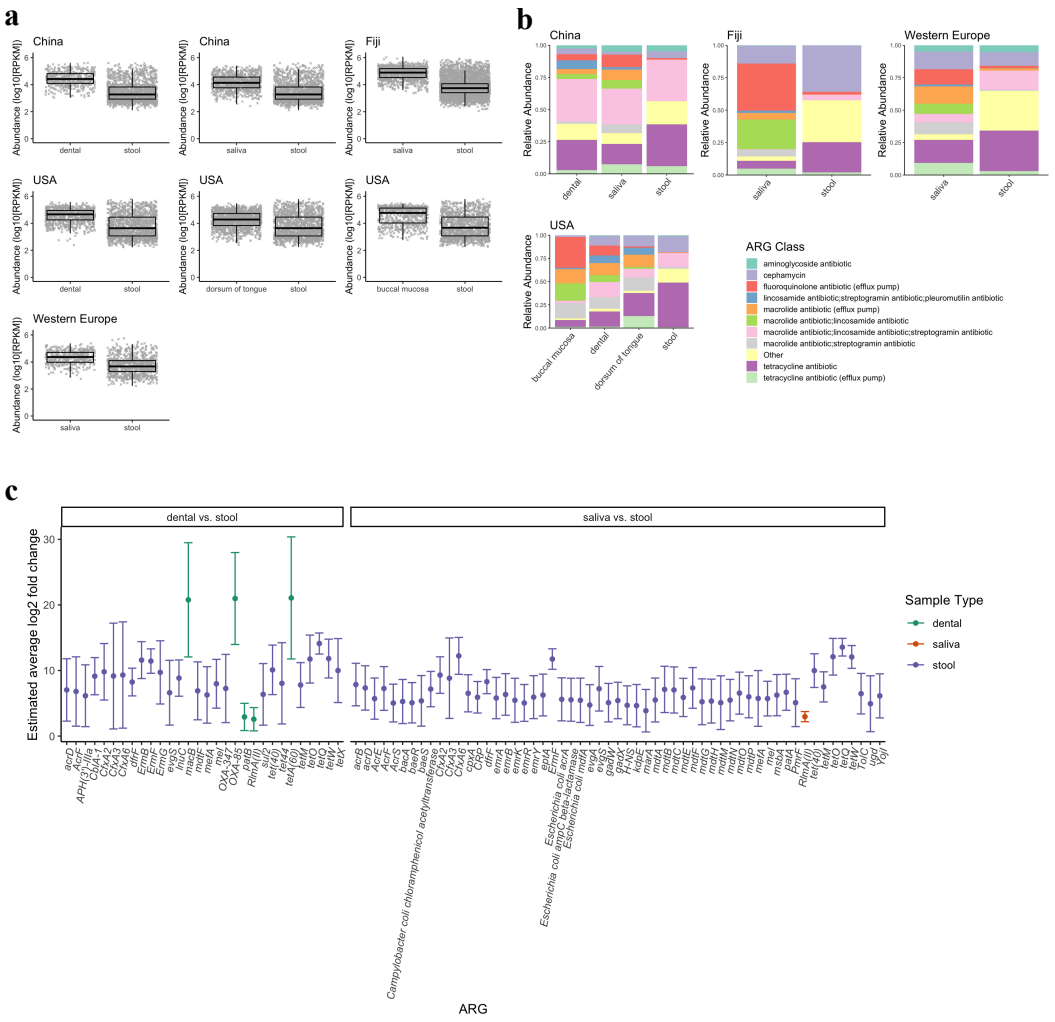

Figure 2.3. Comparing ARG abundance between the oral cavity and gut.

a) Absolute abundance in log10 of RPKM of ARGs for paired samples of individuals from China (stool and dental plaque: n = 30, stool and saliva: n = 31), Fiji (saliva and stool: n = 132), the USA (stool and dental plaque: n = 68, stool and dorsum of tongue: n = 69, stool and buccal mucosa: n = 64) and Western Europe (saliva and stool: n = 21). Centre line is median, box limits are upper and lower quartiles, whiskers are 1.5x interquartile ranges and points beyond whiskers are outliers. **b)** Relative abundance of reads labelled by the top ten most abundant ARG classes across all geographical locations or Other classes for each body site of individuals from China (dental plaque: n = 32, saliva: n = 33, stool: n = 72), Fiji (saliva: n = 136, stool: n = 137), the USA (buccal mucosa: n = 87, dental plaque: n = 90, dorsum of tongue: n = 91, stool: n = 70) and Western Europe (saliva: n = 21, stool: n = 21). **c)** Estimated average log2 fold change of ARGs between paired dental plaque and stool, and saliva and stool samples using random effects meta-analysis across study cohorts (p-value < 0.05). Error bars are 95% confident intervals from meta-analysis. ARGs selected for meta-analysis where adjusted p-value < 0.05 from differential abundance analysis between paired samples of individuals from China (stool and dental plaque: n = 30, stool and saliva: n = 31), Fiji (saliva and stool: n = 132), the USA (stool and dental plaque: n = 68) and Western Europe (saliva and stool: n = 21).

To investigate ARG diversity further, the ARG richness was evaluated between pairwise comparisons of sample types for each cohort with ARG richness defined as the number of unique ARGs per sample. Although there are no significant difference between saliva and stool samples from Fiji and Western Europe, both Chinese and USA samples have significant differences in ARG richness. China and USA stool samples have a significantly higher ARG richness than paired China plaque and saliva, and paired USA plaque and buccal mucosa (Mann-Whitney, paired, two-sided test, p-value < 0.05) **(Fig. 2.4)**. In contrast, the USA dorsum of tongue contains a significantly higher ARG richness than USA stool. Between oral sites, Chinese saliva has a greater ARG richness than paired dental plaque (Mann-Whitney, paired, two-sided test, p-value < 0.05). In addition, USA dorsum of tongue has a higher ARG richness than both plaque and buccal mucosa, whilst plaque has a greater ARG richness than buccal mucosa (Mann-Whitney, paired, two-sided test, p-value < 0.05). It is important to note that while ARG richness only measures the gene incidence regardless of expression, multiple ARGs have the potential to be involved in the expression of a single efflux pump complex, meaning ARG richness may overestimate this potential expression. Therefore, to determine the impact of this overestimation, the analysis was repeated to exclude ARGs that regulate or are part of an efflux pump complex. The differences in ARG richness have the same significance across all paired samples and countries **(Appendix 2: Fig. 2K)**.

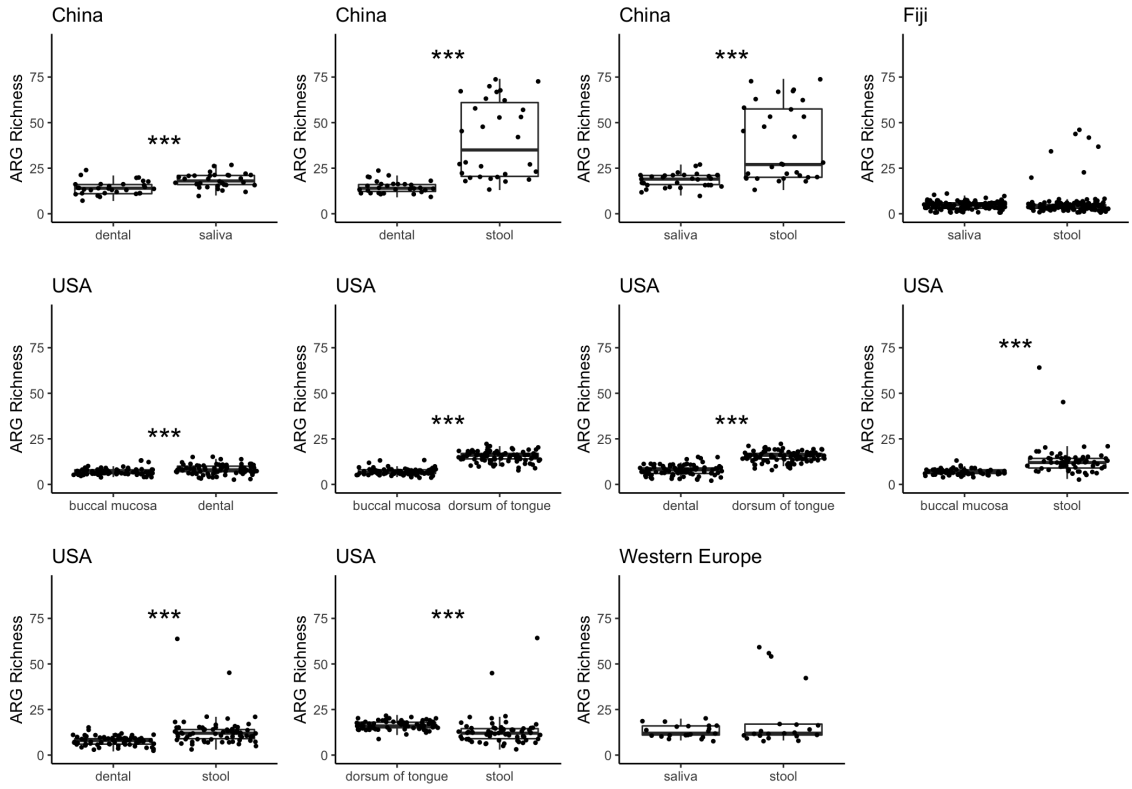

Figure 2.4. Comparing ARG richness between paired body sites.

ARG richness is defined as the number of unique ARGs for paired samples of individual from China (dental plaque and saliva: n = 31, stool and dental plaque: n = 30, stool and saliva: n = 31), Fiji (saliva and stool: n = 132), the USA (buccal mucosa and dental plaque: n = 78, buccal mucosa and dorsum of tongue: n = 86, dental plaque and dorsum of tongue: n = 89, stool and buccal mucosa: n = 64, stool and dental plaque: n = 68, stool and dorsum of tongue: n = 69) and Western Europe (saliva and stool: n = 21) with Mann-Whitney, paired, two-sided test (p-value < 0.05 as *, < 0.01 as **, < 0.005 as ***). Centre line is median, box limits are upper and lower quartiles, whiskers are 1.5x interquartile ranges and points beyond whiskers are outliers.

2.2.4.3 Oral and gut ARG profiles associate with species

Spearman's correlation analysis between ARG and species abundances were conducted to predict the origin of ARGs. Only significant correlations are found in saliva and stool from China, and saliva from the Philippines. The *CfxA* β-lactamase family, *RlmA(II), tetQ*, *tetA(46), pgpB, patB* and *pmrA* ARGs are all strongly correlated with specific species found in both countries **(Fig. 2.5a, b)**. The strongest co-occurrence can be found

in saliva samples from China between *APH(3')-Ia* and a *Komagataella pastoris* strain, *Lactococcus lactis*, *Enterococcus faecalis* and *Leuconostoc mesenteroides*, whilst *pgpB* correlates with *Porphyromonas gingivalis*. The highly abundant ARG *RlmA(II)* in Chinese saliva is associated with *Gemella haemolysans*, *Veillonella parvula* and *Streptococcus mitis/oralis/pneumoniae*. In contrast to saliva, there are fewer species associated with a greater number of ARGs in stool samples from China. *E. coli* in Chinese stool samples is co-associated with many ARGs that encode multidrug efflux pumps and ARGs from *E. coli including ampC β-lactamase*, *acrA* and *mdfA* **(Appendix 2: Fig. 2L)**. Thus, this analysis has the potential to be a predictive tool of ARG origin in metagenomes. However, it can only be applied where an ARG or taxon is found in a high proportion (in this case, at least half) of the samples to ensure Spearman's correlation does not falsely rank many zero-values.

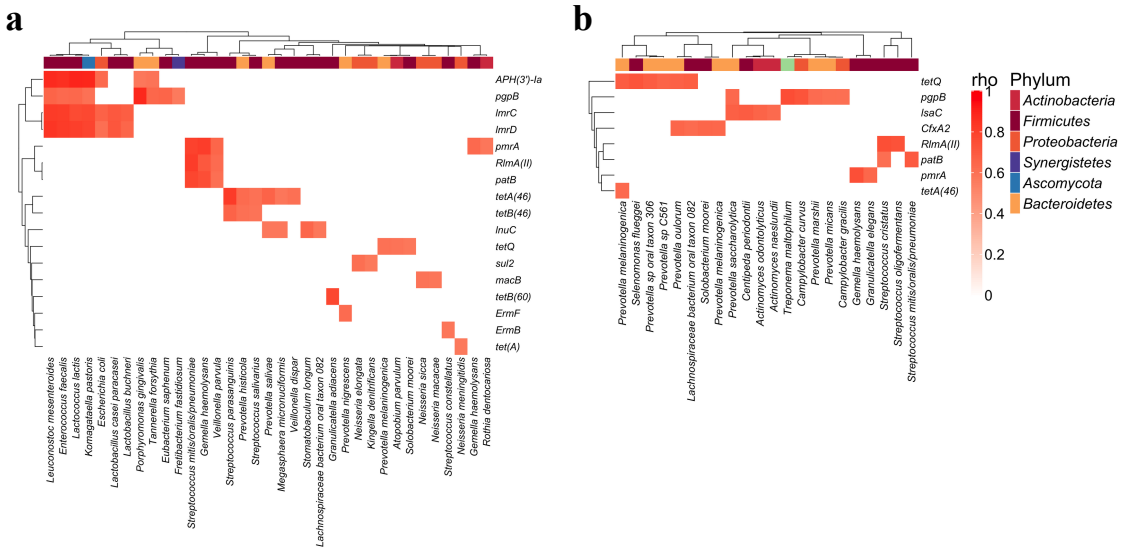

Figure 2.5. Spearman's correlation of ARG and species abundance from saliva samples.

Each heatmap represents correlations of individuals from **a)** China (n = 31) and **b)** Philippines (n = 23). Rows and columns are clustered by hierarchical clustering of Euclidean distance. Columns are coloured by phylum. P-values are adjusted by Benjamini-Hochberg multiple test correction. *Rho* shown only where adjusted p-value < 0.05.

2.2.5 Discussion

AMR is one of the most serious health problems of recent times. The advent of high-throughput sequencing technologies has enabled us to analyse resistomes throughout a microbiome. In this study, we provide key insights into the resistomes of different intraoral sites from healthy individuals across diverse geographical locations and compare their composition to paired gut resistomes. At a population level, there are both country and body site-specific differences in the prevalence of ARGs, ARG classes and resistance mechanisms. It is possible that differences in extraction protocols and batch effects may have a greater bias towards some ARGs over others. Therefore, we do not make direct statistical comparisons between cohorts. For China, the Philippines, USA and Western Europe, the abundance of ARG classes does not correlate with antibiotic

prescription rates. A possible reason for this is the prescription data does not include over-the-counter antimicrobial use, which is especially prevalent in China and the Philippines, and thus may underestimate antimicrobial use[351,352]. In addition, antibiotics are widely used in husbandry and the fishing industry with poorly understood impacts on AMR incidence and dynamics in humans[61,353,354]. Prescription levels for a particular antibiotic are unlikely to be of significant value in the surveillance of AMR in the regional community. Instead, determining the population resistome would be more informative[355].

The abundance and diversity of ARGs at different body sites is also of interest. Although, there are significantly more distinct ARGs in stool compared to oral samples, those ARGs present at the highest relative abundances exist in oral samples. There are several potential reasons why this may be. It is notable that many sites in the oral cavity (e.g. plaque and tongue dorsum) host highly complex and robust microbial biofilm structures. It has been posited that the compact structure of microbes within oral biofilms is a conducive environment for the aquistion of ARGs and their HGT within biofilms[298]. Likewise, the generally protective nature of biofilms against antimicrobial drugs may favour ARG acquisition. It is notable that the dorsum of tongue contains a higher diversity of these genes than other oral sites. This may be explained by the unique papillary structure on the dorsum of the tongue which acts as complex microbiological niche favouring the deposition of oral debris and microbes[356], thus giving rise to a richer microbial community with potentially greater numbers of transient microbes. Another reason could be the difference in species resident in the gut compared to the oral cavity. *E. coli* strains in Chinese stool samples are predicted to

contain a variety of ARGs, especially of the multidrug class, whereas species found in Chinese saliva were estimated to contain fewer ARGs. Due to stringent constraints of the correlation analysis, however, it was not possible to predict the origins for ARGs for all cohorts.

Antibiotic use leading to acquisition of ARGs is another potential factor. Pharmacokinetics of orally administered antibiotics suggests that the oral cavity and oesophagus would be only briefly exposed to the antibiotic during swallowing, whilst the gut is exposed over a more prolonged period. As antibiotics transit through the intestinal tract, they are gradually absorbed via the intestinal epithelium into the bloodstream. Therefore, microbes in the gut exposed to antibiotics for a longer period of time due to their increased bioavailability will receive higher antibiotic selection pressures than those in the oral cavity[43]. The incidences where the oral cavity is likely to acquire ARGs from selective pressures of local antibiotics are from topical antibiotics for periodontal infections or orally administered antibiotics being absorbed into the bloodstream.

The differences in ARG profiles across body sites has significant implications for the characterisation and interpretation of resistome studies. Previous shotgun metagenomic studies have focused almost exclusively on the resistome from the human gut[87,308,325,327]. While the gut may be a diverse reservoir of ARGs, whether these genes are particularly prevalent or have the potential for expression sufficient to drive resistant infections at other body sites is not clear[357]. It is therefore imperative that to test potential applications of non-culture based metagenomic AMR surveillance, we need to

characterise the resistome at different body sites with different pharmacokinetic exposures to antimicrobials. This information can then be integrated with culture-based susceptibility tests, culturomics[358] and functional metagenomic screens[359] to determine the expression potential of these ARGs. In doing so, we will obtain a more complete picture of the state of AMR within a population.

2.2.6 Further discussion

Although this study identified known ARGs from whole metagenomic data, it is likely to have underestimated the diversity of ARGs across body sites. This is because the number of identified ARGs was limited to the ARGs that are already defined in CARD. As discussed in Chapter 1, reference-free approaches, such as functional metagenomics can identify novel ARGs. Functional metagenomics of the gut resistome of preterm infants found 794 ARGs that only had a median of 25.4% amino acid identity with known ARGs in CARD (retrieved 20th October 2014)[180]. However, functional metagenomics is limited by the requirement for ARGs to be expressed and selected by specific antibiotics. Alternatively, predictive models of potential genetic AMR determinants can be elucidated from whole metagenomes[196,360].

This study also raises other questions including why the resistomes differs between body sites and how ARGs are acquired in different body sites and populations by HGT or mutation. Differences in resistomes between GIT sites in this study is thought to be influenced by differences in microbial composition and variable exposures to anthropogenic antimicrobials.

To investigate how the resistome is dependent on microbial composition, the hosts of these ARGs could be determined using long-read or proximity ligation approaches as well as short-read sequencing metagenomics to resolve genomes between species and strains more easily[105,318]. As microbial composition differs across the gut radially and longitudinally[324], biopsy samples of metagenomes from mucosal lining and lumen from distal to proximal intestines would be ideal. Since biopsies are invasive, alternative

technologies could be considered to sample different parts of the gut lumen, such as ingestible capsules[361]. In terms of understanding how differential antimicrobial exposures impact resistomes across body sites, a longitudinal metagenomics study across different body sites could be conducted between individuals before, during and after antibiotic treatment.

Although this study was able to identify homologous ARGs[xvii] from reference sequences, variants of ARGs, particularly single-nucleotide polymorphisms (SNPs)[xviii], were not identified. Profiling SNPs can inform how ARGs evolve and diverge into distinct groups across microbial communities[362], and potentially across human populations[327]. Resolving SNPs from whole metagenomic data using a reference-based approach that relies on calculating read depth is a major challenge[363]. As metagenomes consist of a complex mixture of genomes (with many being incomplete and at low abundance levels), it is not possible to estimate read depth accurately. SNP calling becomes a difficulty since many genomes of the same species are likely to be closely related and it is difficult to calculate significance of a small number of variants based on inaccurate read depth. As discussed in Section 1.5.3.4, functional metagenomics with antibiotic selection pressures could determine novel ARGs, including variants of known ARGs. Alternatively, a non-reference-based *de novo* method using de Bruijn graph structures of co-assembled multiple genomes can identify SNPs and indels (short insertions or deletions)[364,365], which could be developed for more complex, whole metagenomic data[366].

xvii Homologous sequences are similar sequences that are related by evolutionary changes from a common ancestral sequence. (Included in Glossary)

xviii A single-nucleotide polymorphism is a substitution of a single nucleotide at a specific position in a genome. (Included in Glossary)

As well as mutations, ARGs can be acquired by HGT and are carried by MGEs. In the following chapters, I investigate how ARGs may associate with bacteriophages (Chapter 3), plasmids (Chapter 4) and transposable elements (Chapter 5) using tools applied to the same short-read whole metagenomic data from the human gut and oral cavity. In contrast to profiling ARGs, bacteriophages, plasmids and transposable elements were discovered *de novo* without relying on reference databases. This is because MGE reference databases represent a very small proportion of the MGEs in existence and cannot capture the extent of MGE diversity. Once these MGE sequences were profiled, associated ARGs were identified by aligning these MGEs against CARD.

Chapter 3: Bacteriophages and their Association with the Resistome

3 Bacteriophages and their Association with the Resistome

3.1 Introduction

Bacteriophages are the most abundant viral components of the human microbiome. They are viruses that infect and replicate their genome within bacterial or archaeal cells and are likely to have significant effects on microbial composition and function[219,367]. Like eukaryotic viruses, they can have single or double stranded DNA or RNA genomes. They have two principal life cycles: virulent, which destroy bacterial cells immediately after replication; and temperate, during which the phage integrates its genome into the host genome (lysogeny). The latter is involved in the HGT of many genetic elements, including virulence factors. Although phages encode functional genes that can alter the cellular mechanisms of their hosts[99], it is rarer to find ARGs in phage genomes[96]. Nevertheless, they may have some contribution to the spread of ARGs across environmental and healthcare ecosystems[97].

Many studies that attempt to profile bacteriophages in human microbiomes use computational analysis of faecal metagenomic data, often following enrichment of VLPs[223,368]. Between 2016 and 2018, 750,000 uncultivated virus genomes were identified from metagenomic datasets, five times greater than the total number of genomes sequenced from virus isolates[369]. These include crAssphages, a highly abundant bacteriophage clade currently thought to play a special role in human and

primate gut microbiomes[370–372]; the discovery of bacteriophages with atypically large genomes greater than 200 kb in length, known as jumbo phages[373]; and the more recently identified megaphages with genomes larger than 500 kb[374]. A few studies have profiled bacteriophages using metagenomic data from the oral cavity[375,376]. The heterogeneity of bacteriophage genomes and the lack of correlation between phage phylogeny and that of their hosts makes classification and host assignment challenging tasks, leaving a relatively unexplored melting pot of "viral dark matter"[368,377].

Here, bacteriophage DNA was profiled from whole metagenomes of the human GIT, specifically comparing gut (represented by faecal samples) with paired saliva and dental plaque from China, and the dorsum of tongue, dental plaque and buccal mucosa from longitudinal samples from the USA. Novel bacteriophages, including jumbo phages, were identified from assembled metagenomic contigs using *de novo* bioinformatic pipelines, including viral motif recognition[262] and protein-coding gene-sharing networks[378], to identify and classify linear and circular viral contigs. Bacteriophage hosts were also predicted using CRISPR spacer matches with reference bacterial genomes[93].

In summary, a catalogue of 78,327 genetically distinct bacteriophage genomes was created from 854 oral and gut metagenomes. Bacteriophages and their host profiles are specific to GIT sites, with more differences found between the oral cavity and the gut than between different oral sites. The dorsum of the tongue contains a greater diversity of bacteriophages than paired stool samples, saliva and buccal mucosa. In addition, bacteriophages can persist in some individuals for months at a time, including phage

families that are common to both oral cavity and gut, or are specific to a GIT site, such as the crAss-like family of the gut, *Inoviridae* of the oral cavity, and *Microviridae* of the gut, dorsum of the tongue and saliva. Consistent with greater genotypic diversity, 37 unique circular jumbo phage genomes were found in oral cavities from China and the USA, most on the dorsum of the tongue from the USA, while none were identified in paired stool samples. Only 72 distinct prophages (i.e. integrated into the host chromosome) were found to contain ARGs, mainly encoding intrinsic rather than acquired resistance. None of these were jumbo phages. The oral cavity provides conducive environments, such as robust biofilms, that can harbour genetically diverse phages, but phages of the GIT rarely encode ARGs.

3.2 Methods

3.2.1 Metagenomic data for creating the phage catalogue

A total of 1,061 publicly available metagenomic samples covering the USA, China and the Philippines, all sequenced using Illumina HiSeq 2000, were used to create a reference phage contig catalogue. Longitudinal USA samples were excluded from the majority of the study after the first timepoint to ensure each sample was independent, unless specified otherwise. All metagenomes passed over half the quality control metrics in FastQC 0.11.3 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) with these pass rates calculated in MultiQC[194]. These samples include:

1) Longitudinal data across two years with various timepoints from the Human Microbiome Project 1 (referred to as USA) containing buccal mucosa (n = 87:

32 with one, 35 with two, 19 with three and 1 with six timepoints); dorsum of tongue (n = 91: 22 with one, 43 with two, 24 with three and 2 with four timepoints); dental plaque (n = 90: 24 with one, 41 with two, 21 with three, 1 with four and 3 with six timepoints); stool (n = 70: 13 with one, 33 with two, 21 with three, 2 with four and 1 with six timepoints)[336].

2) Healthy controls and rheumatoid arthritis patients from a Chinese study containing dental plaque (healthy: n = 32, rheumatoid arthritis: n = 76); saliva (healthy: n = 33, rheumatoid arthritis: n = 24); stool (healthy: n = 72, rheumatoid arthritis: n = 100)[331].

3) Saliva samples (n = 24) from healthy hunter-gatherers and traditional farmers from the Philippines[333].

Raw paired-end metagenomic reads from Chinese and Philippines samples were downloaded from EMBL-EBI (https://www.ebi.ac.uk/metagenomics/). Paired-end metagenomic samples from USA were downloaded from https://portal.hmpdacc.org/. All USA, China and Philippines samples were collected and sequenced as described in the following cited studies[331,333,336].

3.2.2 Phage contig catalogue

The metagenomic reads (processed from Chapter 2: 2.2.3.1 and 2.2.3.2) were assembled into contigs using SPAdes v3.9.0[205] with parameters *-k21,33,55 --only-assembler --meta*. Small contigs with a length of less than 3,000 bp were removed. Linear and circular phage contigs were identified by searching for viral signatures using VirSorter v1.0.5[262].

The methods in this paragraph were carried out by collaborator, Dr Andrey Shkoporov.

The phage contigs were clustered with multi-alignment using BLASTn v2.6.0[207] with parameters *-evalue 1e-20 -word_size 100 -max_target_seqs 10000* and redundant contigs were removed where identity and breadth coverage were both greater than and equal to 90%. Phage contigs were then clustered into viral clusters using vConTACT2 v0.9.10 with all default parameters[378]. Taxonomic classification of phage contigs to Order and Family level was predicted using Demovir (build 20th April 2018) on phage proteins with an e-value cut-off of 1e-5 (https://github.com/feargalr/Demovir). Non-phage viral families were labelled as "Other": *Alloherpesviridae*; *Ascoviridae*; *Baculoviridae*; *Flaviviridae*; *Herpesviridae*; *Iridoviridae*; *Marseilleviridae*; *Mimiviridae*; *Nudiviridae*; *Phycodnaviridae*; *Picornaviridae*; *Pithoviridae*; *Poxviridae*; and *Retroviridae*. Although some phage clusters were labelled as non-phage, the developers of Demovir recommend against using their classification to discriminate between bacterial, archaeal and eukaryotic sequences from metagenomic samples. The phage hosts were predicted by aligning phage contigs against a database of CRISPR spacers as described by Shkoporov et al., 2019[93]. Metadata for the phage catalogue were compiled using the helper script *create_catalogue_dataset.R*.

3.2.3 Sequence and functional annotation of jumbo phages

The methods in this paragraph were carried out by collaborator, Dr Andrey Shkoporov.

Circular phage genomes of length > 200 kbp and linear phage genomes of length > 200 kbp that were connected to these circular phages in vConTACT2's gene-sharing network were put forward as candidate jumbo phages using the helper script

get_candidate_jumbophages.R. The scaffold file containing their genomes was generated using the helper script *extract_jumbophage_contigs.py*.

The candidate jumbo phage genomes were annotated for functional proteins and tRNA genes. Protein prediction was conducted using Prodigal v2.6.3[210] with parameters *-p meta*. Protein sequences were searched against databases of HMMs: pVOGs (downloaded 1st November 2019)[255], pFAMs[379] (downloaded 2nd September 2019) and TIGRFAMs[380] (downloaded 3rd September 2019), using hmmsearch v3.2.1[209] with e-value cut-off of 1e-5. tRNA genes were identified from nucleotide sequences using ARAGORN v1.2.36[381] with parameters *-t -i -c -d -w* and with helper script *clean_aragorn_output.py*. The hit with the lowest e-value and domain e-value was selected for every query protein with candidate target protein hits for each database. Next, the hit with the highest bit score and domain bit score was selected for every query protein with candidate target protein hits from more than one database. The few remaining protein query sequences with the same e-values and hit scores were deduplicated. These steps were run using the helper script *collate_functional_annotations.R*. Circular candidate jumbo phages > 200 kbp that did not contain a major capsid protein were also excluded, leaving a total of 545 putative jumbo phages.

3.2.4 Phage annotation in metagenomes

The methods in this paragraph were carried out by collaborator, Dr Andrey Shkoporov.

854 metagenomes from healthy individuals were mapped against the non-redundant phage catalogue using Bowtie2 v2.3.4.1. Phage contigs (excluding spurious jumbo

phages) and phage clusters were quantified for each sample where contig breadth coverage was 75% or greater using helper script *phage_quantification.R*. Relative phage abundance profiles were calculated by scaling the depth coverage of phage contigs that were divided by total reads per sample. The metagenomes came from the USA with buccal mucosa (n = 87: 32 with one, 35 with two, 19 with three and 1 with six timepoints); dorsum of tongue (n = 90: 22 with one, 43 with two, 24 with three and 2 with four timepoints); dental plaque (n = 90: 24 with one, 41 with two, 21 with three, 1 with four and 3 with six timepoints); and stool samples (n = 70: 13 with one, 33 with two, 21 with three, 2 with four and 1 with six timepoints), China with dental plaque (n = 32); saliva (n = 33); and stool samples (n = 72), and the Philippines with saliva samples (n = 24). Metadata for the samples is available here: https://tinyurl.com/y6tzb6gz. The following analysis was conducted in script *phage_analysis.R*.

3.2.5 Phage diversity

The methods in this paragraph were carried out by collaborator, Dr Andrey Shkoporov.
In ecology, the β-diversity measures the variation of taxonomic composition between samples, i.e. the ratio between regional and local diversity. To find differences in β-diversity of phage cluster profiles between groups of individuals, the Bray-Curtis dissimilarities of phage cluster incidence (presence or absence) profiles were computed between individuals and visualised using non-metric multidimensional scaling (NMDS).

Silhouette analysis of *k*-medoids using the cluster package v2.1.0 was applied to select the number of distinct groups with the largest average silhouette width **(eq. 2.1)**. *k*-medoids is a clustering algorithm where data points are grouped into *k* clusters with a

specified *k* value. Clusters are partitioned to minimise the distance between points within a cluster and a designated data point as the centre of the cluster.

The Bray-Curtis dissimilarity is defined as:

(eq. 3.1) $$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

where C_{ij} is the number of phage clusters shared in both sites *i* and *j*. S_i is the total number of phage clusters in site *i*, and S_j is the total number of phage clusters in site *j*.

In contrast to the β-diversity, the α-diversity measures the variation of taxonomic composition within a sample. The α-diversity was calculated as the phage cluster richness which is the number of unique phage clusters for each sample. Only samples with greater than 100 phages were included. Individuals with samples containing less than or equal to three clusters were excluded from the group comparison. The phage cluster richness was compared between paired GIT sites from the same individuals in each of the following groups: China dental plaque vs. saliva (n = 30); China stool vs. saliva (n = 30); China stool vs. dental plaque (n = 30); USA buccal mucosa vs. dental plaque (n = 45); USA buccal mucosa vs. dorsum of tongue (n = 45); USA buccal mucosa vs. stool (n = 36); USA dental plaque vs. dorsum of tongue (n = 86); USA dental plaque vs. stool (n = 67); and USA dorsum of tongue vs. stool (n = 68). Since the number of phage contigs in each sample is significantly linearly correlated with Phage Cluster Richness ($p < 2.2 \times 10^{-16}$) **(Appendix 3: Fig. 3A)**, the number of phage clusters for each sample were subsampled to the smallest number of phages found in a sample for each paired comparison. The phage cluster richness between samples in each group was tested for statistical significance with a Two-sided Wilcoxon Rank Sum Test.

3.2.6 Microbial composition

MetaPhlAn2 v2.6.0[349] was used to identify the composition of bacteria and archaea from samples apart from longitudinal USA samples (as described in Chapter 2, Section 2.2.3.9). One dorsum of tongue USA sample did not have bacterial nor archaeal microbial predictions.

The methods in this paragraph were carried out by collaborator, Dr Andrey Shkoporov. Procrustes analysis was applied to visualise the superposition of NMDS dimensions of phage incidence profiles on microbial genera incidence profiles using the protest function in the vegan package v2.5.6 in R. Procrustean randomisation test (PROTEST) was performed with 999 permutations to a significance of $p = 0.001$.

3.2.7 Longitudinal analysis of phages

The stability of the phage community within the microbiome was investigated by computing the number of timepoints each phage cluster is found from each individual and GIT site in the longitudinal USA data. The proportion of phage clusters and reads mapped against these clusters were calculated for each number of sampled timepoints available for each individual and GIT site: buccal mucosa ($n = 20$), dental plaque ($n = 24$), dorsum of the tongue ($n = 26$) and stool ($n = 24$). Persistent phage clusters are defined as being found in three or more timepoints, whereas transient phage clusters are defined as being found in less than three timepoints. The Bray-Curtis dissimilarities were computed between persistent and transient phage cluster incidence profiles from

GIT sites of individuals containing both persistent and transient phage clusters: buccal mucosa (n = 18), dental plaque (n = 23), dorsum of the tongue (n = 25) and stool (n = 24). NMDS was applied to scale the dissimilarity into a two-dimensional ordination using the metaMDS function in the vegan package v2.5.6 in R. Permutational multivariate analysis of variance (PERMANOVA) analysis was performed using the adonis function in the vegan package.

3.2.8 ARG annotation of phages

Phage contigs were annotated for ARGs by mapping against CARD v3.0.0 using BLASTn v2.10.0 with parameters *-evalue 1e-5*. Hits were filtered by 90% identity.

3.2.9 Code availability

The code for all analysis (including contributions from Dr Andrey Shkoporov) is available from https://github.com/APC-Microbiome-Ireland/phageome_analysis

3.3 Results

3.3.1 Phage composition and diversity differs between GIT sites

A catalogue of 139,929 phage contigs were identified from 854 oral and gut metagenomes. Phage contigs were grouped into clusters using vConTACT2[378], a software based on gene-sharing profiles that has been applied to stratify uncultured viral populations in a variety of environments, including human gut metagenomes[93,368].

75,680 phage contigs were clustered into 14,078 phage clusters, while the remaining 64,249 remain as singletons, making 78,327 distinct genotypes. It is more common to find phage contigs that are unique to one sample than phage clusters found, meaning phage contig profiles are more variable than phage cluster profiles **(Appendix 3: Fig. 3B)**. β-diversity was evaluated with the Bray-Curtis dissimilarity metric from phage cluster abundance profiles to investigate differences between metagenomes. Samples were clustered using Silhouette analysis on *k*-medoids of NMDS dimensions. Four groups were generated **(Fig. 3.1a)**. Most phage clusters are exclusively found in either Group 2, 3 or 4 or are shared between Group 2 and 3 **(Fig. 3.1b)**. The composition of these groups is dominated by samples from a particular GIT site **(Fig. 3.1c)**. Groups 1 to 3 contain all oral sites: Group 1 consists of mostly buccal mucosa, Group 2 contains mostly dorsum of tongue and saliva, and Group 3 has mostly dental samples. In contrast, Group 4 contains exclusively stool samples. These results reveal that many phage clusters are solely found in either dorsum of tongue (Group 2), dental (Group 3) or stool samples (Group 4), but most phage clusters that are shared occur in both dorsum of tongue and dental sites. Hierarchical clustering of the relative abundance of phage clusters also shows GIT sites can be differentiated by phage cluster profiles, especially between dental, dorsum of tongue and stool samples **(Appendix 3: Fig. 3C)**. GIT sites have distinct phage cluster profiles, and phage clusters that are shared are mostly found in oral sites. Incomplete metadata on sex and age meant that individual variation could not be considered **(Appendix 3: Fig. 3D)**.

Additionally, the α-diversity of the phage cluster composition was compared between paired GIT sites from China and the USA. The Phage Cluster Richness, i.e. the number

of unique phage clusters, was evaluated for every paired sample. Dental plaque from the USA ($p = 2.84x10^{-4}$) and from China ($p = 0.0476$), and dorsum of the tongue samples from the USA ($p = 1.89x10^{-7}$) have a significantly greater Phage Cluster Richness than stool samples (Two-sided Wilcoxon Rank Sum Test) **(Fig. 3.1d)**. USA dorsum of the tongue samples also have a significantly higher richness than dental plaque ($p = 0.0469$). Buccal mucosa from the USA and saliva from China have the lowest richness compared to all other paired GIT sites (saliva vs. dental plaque from China: $p = 2.94x10^{-6}$, saliva vs. stool from China: $p = 1.93x10^{-4}$, buccal mucosa vs. dental plaque from the USA: $p = 3.91x10^{-7}$, buccal mucosa vs. dorsum of the tongue from the USA: $p = 1.26x10^{-8}$, and buccal mucosa vs. stool from the USA: $p = 0.00175$).

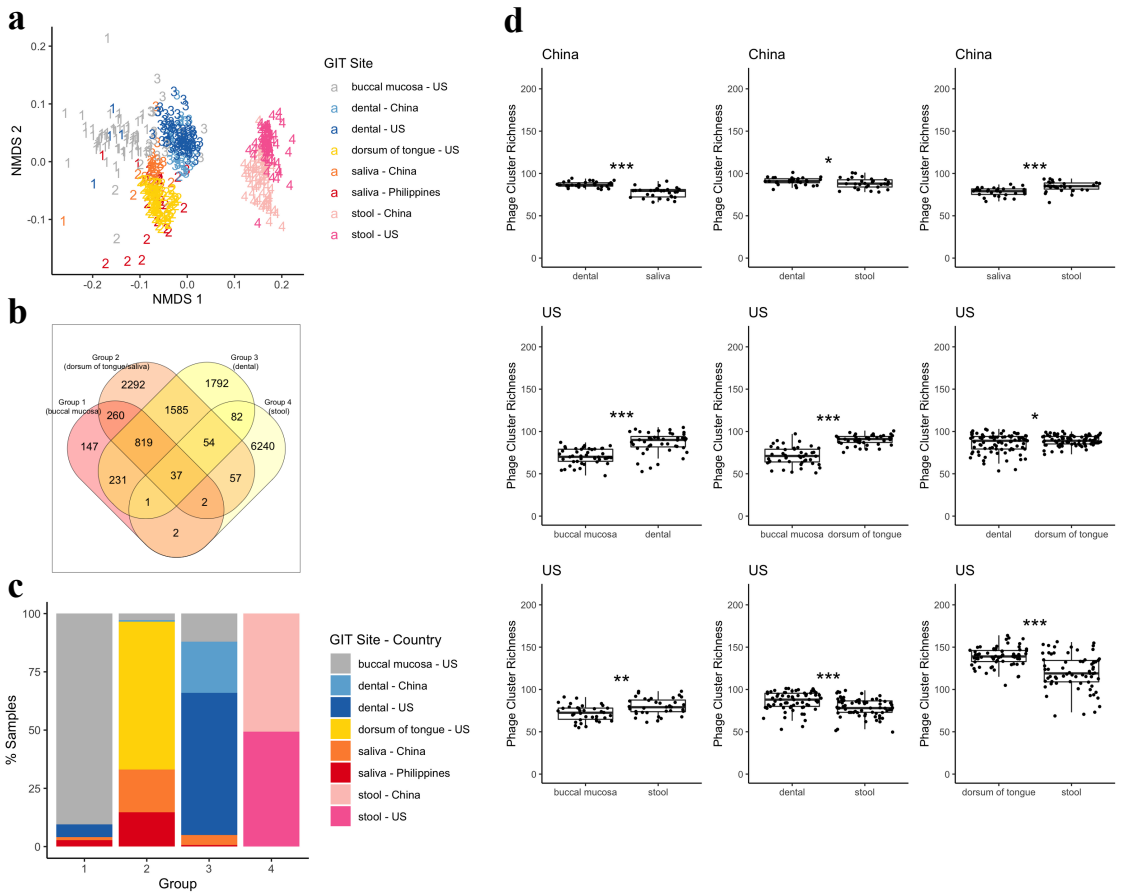

Figure 3.1. Phage incidence and abundance profiles.

a) NMDS of Bray-Curtis dissimilarities between phage incidence profiles of samples (excluding longitudinal USA). Ordination coordinates are grouped by *k*-medoids clustering, where number of groups, *k*, has the largest Silhouette width. USA buccal mucosa (n = 87), dorsum of tongue (n = 90), dental plaque (n = 90) and stool (n = 70); China dental plaque (n = 32), saliva (n = 33) and stool (n = 72); and Philippines saliva (n = 24). **b)** Number of viral clusters in each group. **c)** Percentage of samples in each group from 1a, labelled by GIT site and country. **d)** Phage Cluster Richness between paired GIT sites. Phage Cluster Richness is defined as the number of unique viral clusters in a sample that is subsampled to the smallest number of non-unique clusters. Phage Cluster Richness is calculated for samples of individuals from China (dental plaque and saliva: n = 30, stool and saliva: n = 30, stool and dental plaque: n = 30) and the USA (buccal mucosa and dental plaque: n = 45, buccal mucosa and dorsum of tongue: n = 45, buccal mucosa and stool: n = 36, dental plaque and dorsum of tongue: n = 86, dental plaque and stool: n = 67, dorsum of tongue and stool: n = 68) (excluding longitudinal USA) with Two-sided Wilcoxon Rank Sum Test ($p < 0.05$ as *, < 0.01 as **, < 0.005 as ***). Centre line is median, box limits are upper and lower quartiles, whiskers are 1,5x interquartile ranges and points beyond whiskers are outliers.

Most oral sites and stool metagenomes contain a high abundance of *Siphoviridae*, *Myoviridae* and *Podoviridae* phage families but there are some distinctive differences in the less abundant phage families between GIT sites **(Appendix 3: Fig. 3E)**. Bacterial hosts for phages were predicted using Demovir (https://github.com/feargalr/Demovir). Although host bacteria could only be predicted for 11.8% (16,513/139,929) of phage contigs, these highly abundant phage families infect a range of bacterial genera **(Appendix 3: Fig. 3F)**. crAss-like phage clusters are only found in stool samples and in *Prevotella* spp. as host, apart from two saliva and two dorsum of the tongue samples. In contrast to the crAss-like family, *Inoviridae* are found almost exclusively in oral sites and predicted to infect *Neisseria* species. *Microviridae* are present in dorsum of the tongue, saliva and stool samples but are far less prevalent in buccal mucosa and dental plaque. *Prevotella* in oral sites and *Faecalibacterium* are the only genera predicted to contain *Microviridae*. *Bicaudaviridae* are only found in ten dorsum of the tongue samples and are represented by only one phage cluster. There are 64 crAss-like **(Appendix 3: Fig. 3Ga)**, 30 *Inoviridae* **(Appendix 3: Fig. 3Gb),** 56 *Microviridae* **(Appendix 3: Fig. 3Gc)**, 3,433 *Myoviridae* **(Appendix 3: Fig. 3Gd)**, 728 *Podoviridae* **(Appendix 3: Fig. 3Ge)** and 6,967 *Siphoviridae* **(Appendix 3: Fig. 3Gf)** phage clusters. Across highly abundant families (i.e. *Siphoviridae*, *Myoviridae* and *Podoviridae*), GIT sites are clustered by the incidence of phage clusters suggesting that phage clusters of the same family are also specific to GIT site. In terms of lower abundance families found across multiple GIT sites, there is a pronounced separation of oral sites and stool in *Microviridae* phage clusters, with some separation of oral sites by *Inoviridae* phage clusters. Notably, the *Microviridae* family is dominated by a few

phage clusters in oral sites in contrast to stool that contains many different phage clusters.

3.3.2 Phage hosts match varied microbial composition across GIT sites

Bacteriophages have previously been shown to modulate the microbiota composition (in particular the bacterial and archaeal composition) in the mouse gut via phage infection and lysis of specific host bacteria[219]. To investigate whether microbial composition is associated with distinctions in phageome profiles between GIT sites, bacterial and archaeal taxonomies were profiled using MetaPhlAn2[349]. Clustering of taxonomic composition is associated with GIT site, with greater separation between oral sites and gut **(Fig. 3.2a)**. Procrustes analysis was applied to match corresponding points between phageome and taxonomic profiles. PROTEST was then used to determine whether two profiles showed significant association. Microbial composition and phageome profiles correlate and co-locate by GIT site, especially between stool and dental samples **(Fig. 3.2b)** (0.70 to a significance of $p = 0.001$ in PROTEST). Predicted phage hosts group by GIT site with very little overlap between gut and oral sites **(Fig 3.2c, Appendix 3: Fig. 3H)**. *Actinomyces, Atopobium, Campylobacter, Fusobacterium, Neisseria* and *Rothia* genera, that are mostly found in the oral cavity **(Appendix 3: Fig. 3I)**, are predicted hosts of oral site phage **(Fig 3.2c)**. Likewise, *Bacteroides*, *Bifidobacterium, Clostridium*, *Roseburia*, *Ruminococcus* and *Parabacteroides* genera that are found more exclusively in stool samples are also potential phage hosts in the gut. However, *Eubacterium*, a bacterium mostly found in stool, and *Haemophilus, Prevotella*, *Streptococcus* and *Veillonella,* bacteria mainly residing in the oral sites, are prevalent across all oral sites

and the gut. Upon closer inspection of the abundance of these genera at a species level, there are some species that are more prevalent in either the oral cavity or the gut. For instance, *E. brachy* and *E. saphenum* are species of *Eubacterium* that are found almost exclusively in the oral cavity **(Appendix 3: Fig. 3J)**. Likewise, *P. copri* mostly represents *Prevotella*, and is found almost exclusively in the gut. Generally, phage host predictions match microbial composition at a genus level.

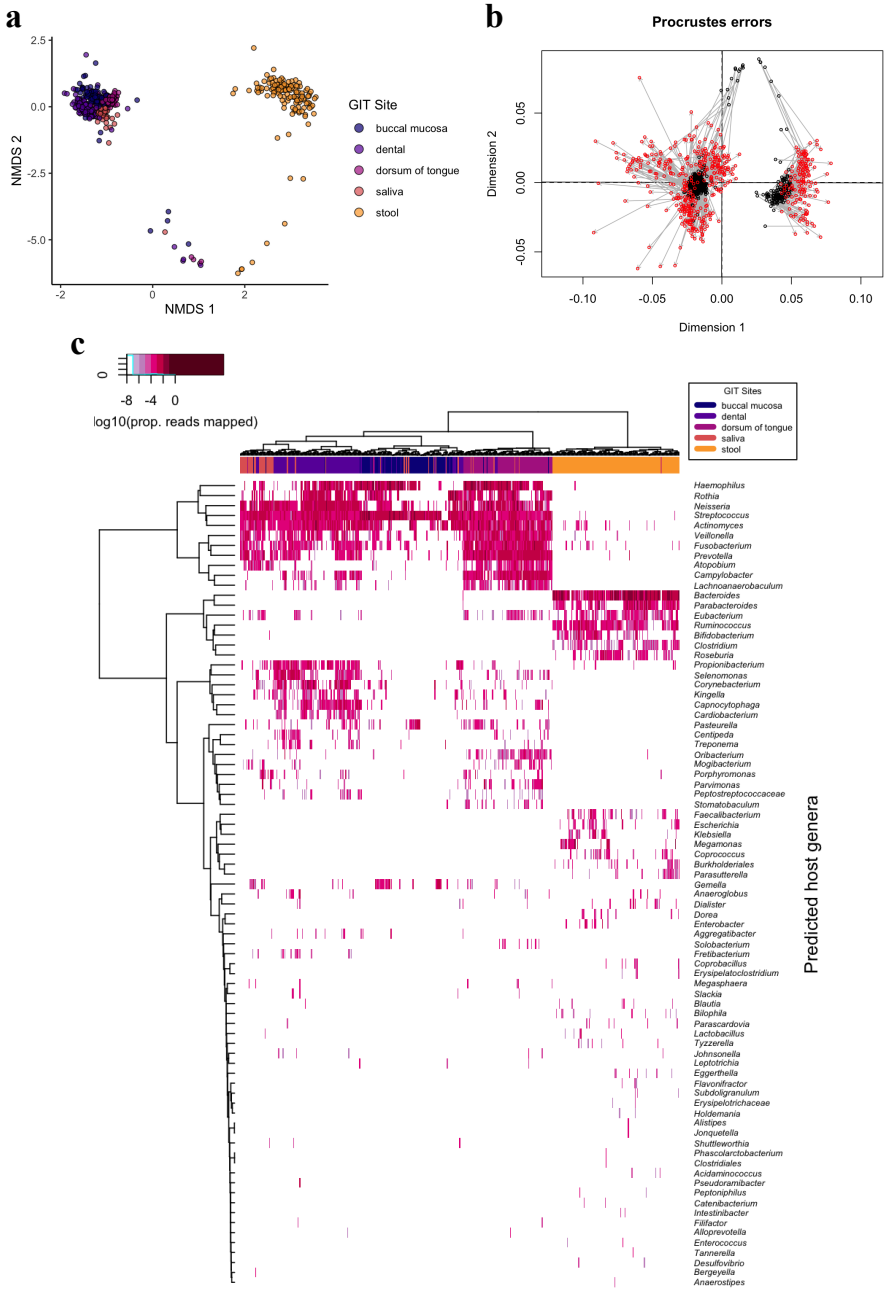

Figure 3.2. Relationship between phage profiles and microbial composition, and abundance of predicted phage hosts.

a) NMDS of Bray-Curtis dissimilarities between microbial taxa incidence profiles of samples (excluding longitudinal USA) labelled by GIT site. **b)** Procrustes rotation of NMDS coordinates between microbial genera profile from 2a (black) and phageome profile from 1a (red). Correlation in symmetric Procrustes rotation = 0.70 (p = 0.001; 999 permutations; PROTEST). **c)** Log10 of the proportion of reads mapped to phage contigs with predicted host for each sample, clustered by hierarchical clustering, x-axis coloured by GIT site and y-axis labelled by genus of predicted host. USA buccal mucosa (n = 87), dorsum of tongue (n = 90), dental plaque (n = 90) and stool (n = 70); China dental plaque (n = 32), saliva (n = 33) and stool (n = 72); and Philippines saliva (n = 24).

3.3.3 Stability of phage clusters across longitudinal metagenomes

To clarify whether phages that are associated with a GIT site are also stable over time, phage clusters were profiled in longitudinal samples from the USA taken over a two-year period with a minimum of two and maximum of six sampling timepoints. The proportions of total phage clusters **(Fig. 3.3a)** and total reads mapped to these phage clusters **(Fig. 3.3b)** drops over time across all GIT sites, but this is unclear from timepoints 4-6 in buccal mucosa, 4 in dorsum of tongue and 5-6 in stool due to a small number of samples.

To compare the association between stable and unstable phage clusters and GIT sites, the longitudinal phageome was separated empirically into persistent and transient phage cluster profiles. Persistent phage clusters are defined as being present in three or more timepoints in a given GIT site, whereas transient phage clusters are found in less than three timepoints. There are 194 (buccal mucosa), 1,377 (dental plaque), 1,423 (dorsum of the tongue) and 1,899 (stool) persistent phage clusters, and 1,323 (buccal mucosa), 2,795 (dental plaque), 3,246 (dorsum of the tongue) and 2,773 (stool) transient phage clusters. NMDS of persistent and transient phage cluster profiles in an ordination of two-dimensions appears to show that longitudinal samples are more clustered by individual in persistent **(Fig. 3.3c)** compared to transient profiles **(Fig. 3.3d)**. PERMANOVA was applied to persistent and transient phage cluster profiles for each GIT site to find which category has the highest individuality. A greater percentage of

variance of persistent than transient phage clusters can be explained by individual variability across GIT sites. 78.1% (buccal mucosa), 72.3% (dental plaque), 78.4% (dorsum of the tongue) and 85.5% (stool) variance of persistent phage cluster profiles and 43.9% (buccal mucosa), 41.9% (dental plaque), 47.0% (dorsum of the tongue) and 48.6% (stool) variance of transient phage cluster profiles can be explained by individual ($p < 0.001$, PERMANOVA). All viral families are represented in both persistent **(Fig. 3.3e)** and transient phage clusters **(Fig. 3.3f)**. Noticeably, the crAss-like family are prominently represented in persistent phage clusters of stool samples. Both persistent and transient phage clusters are found at various levels of prevalence in these individuals **(Fig. 3.3g)**. The percentages of persistent phage clusters from one individual also seen in another are 76.0 ± 21.7 (buccal mucosa), 80.8 ± 5.7 (dental plaque), 81.3 ± 11.1 (dorsum of the tongue) and 75.9 ± 13.4 (stool), and for transient clusters are 69.7 ± 11.0 (buccal mucosa), 83.4 ± 7.2 (dental plaque), 81.9 ± 5.3 (dorsum of the tongue) and 76.9 ± 7.1 (stool) (median ± interquartile range). There are no significant differences in sharing of persistent or transient phage clusters between individuals for each GIT site ($p = 0.477$, Wilcoxon Rank Sum Test).

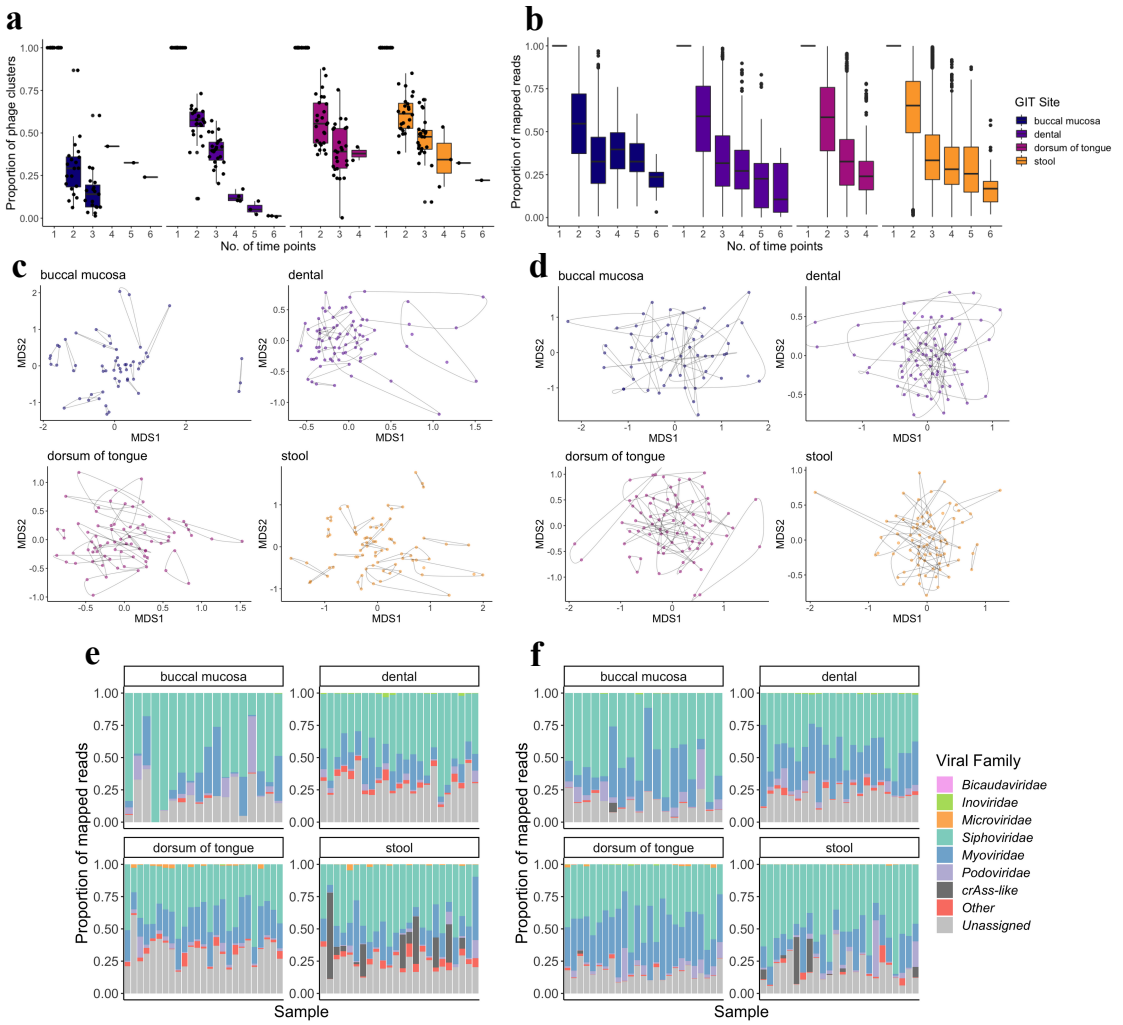

Figure 3.3. Phage cluster stability in longitudinal USA oral and gut samples.

Proportion of **a)** phage clusters and **b)** reads mapped to phage clusters, in one to six timepoints for USA individuals with at least three sampling timepoints (buccal mucosa: n = 20, dental plaque: n = 24, dorsum of the tongue: n = 26, stool: n = 24). Non-metric multidimensional scaling of the Bray-Curtis dissimilarities between phage clusters incidence profiles of samples with **c)** persistent and **d)** transient phage clusters. Points represent samples and lines joining points represent grouping samples from the same individual and GIT site (buccal mucosa: n = 18, dental plaque: n = 23, dorsum of the tongue: n = 25, stool: n = 24). Proportion of reads that were mapped to phage clusters, coloured by viral family, containing only. No convergent solutions were found. Proportion of reads mapped to **e)** persistent and **f)** transient phage clusters for individuals in c) and d). "Other" represents non-phage viral families, *Alloherpesviridae*, *Ascoviridae*, *Baculoviridae*, *Flaviviridae*, *Herpesviridae*, *Iridoviridae*, *Marseilleviridae*, *Mimiviridae*, *Nudiviridae*, *Phycodnaviridae*, *Picornaviridae*, *Pithoviridae*, *Poxviridae* and *Retroviridae*. **Figure 3.3. continued on next page.**

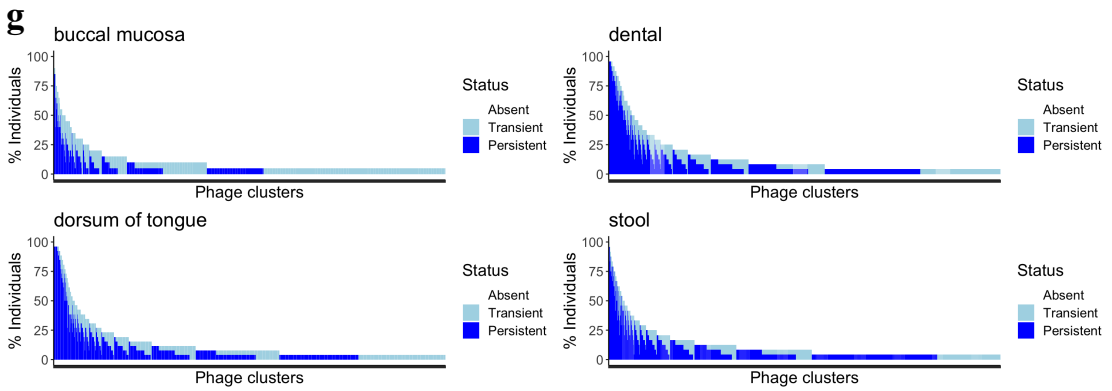

Figure 3.3. continued: g) Prevalence of transient and persistent phage clusters in the same individuals ordered by decreasing prevalence of total and persistent phage clusters.

3.3.4 Very few phage genomes contain ARGs

77 (0.055%) phages were found to associate with ARGs. Out of these, 24 were part of phage clusters while 48 were left as singletons, leaving 72 distinct genotypes. These phages carried one ARG each (an average of one ARG per phage cluster or singleton). All of their genomes are linear and 77.9% (60/77) carry integrase genes, indicating most of them are integrated as prophages in the host genome. This is in contrast to 25.8% (35,221/136,489) of all linear phages that have integrases. The distribution (i.e. the cumulative distribution function) of the genome sizes of phages that carry ARGs is significantly higher than that of phage genomes that do not (p-value < 2.2×10^{-16}, Kolmogorov-Smirnov test). Many of these phages originate from stool samples, while others are derived from buccal mucosa, dental plaque and dorsum of the tongue samples **(Fig. 3.4)**. Phages from the gut stool samples contain a broad range of ARGs, particularly those conferring resistance to multiple antimicrobials by efflux pump and non-efflux pump mechanisms. However, *RlmA(II)*, *mel*, *hmrM*, *ErmX* and the *CfxA* family are exclusively found in oral sites. In addition, the *hmrM* gene, coding for

fluoroquinolone and acridine dye efflux, is shared between oral sites from the same individuals.

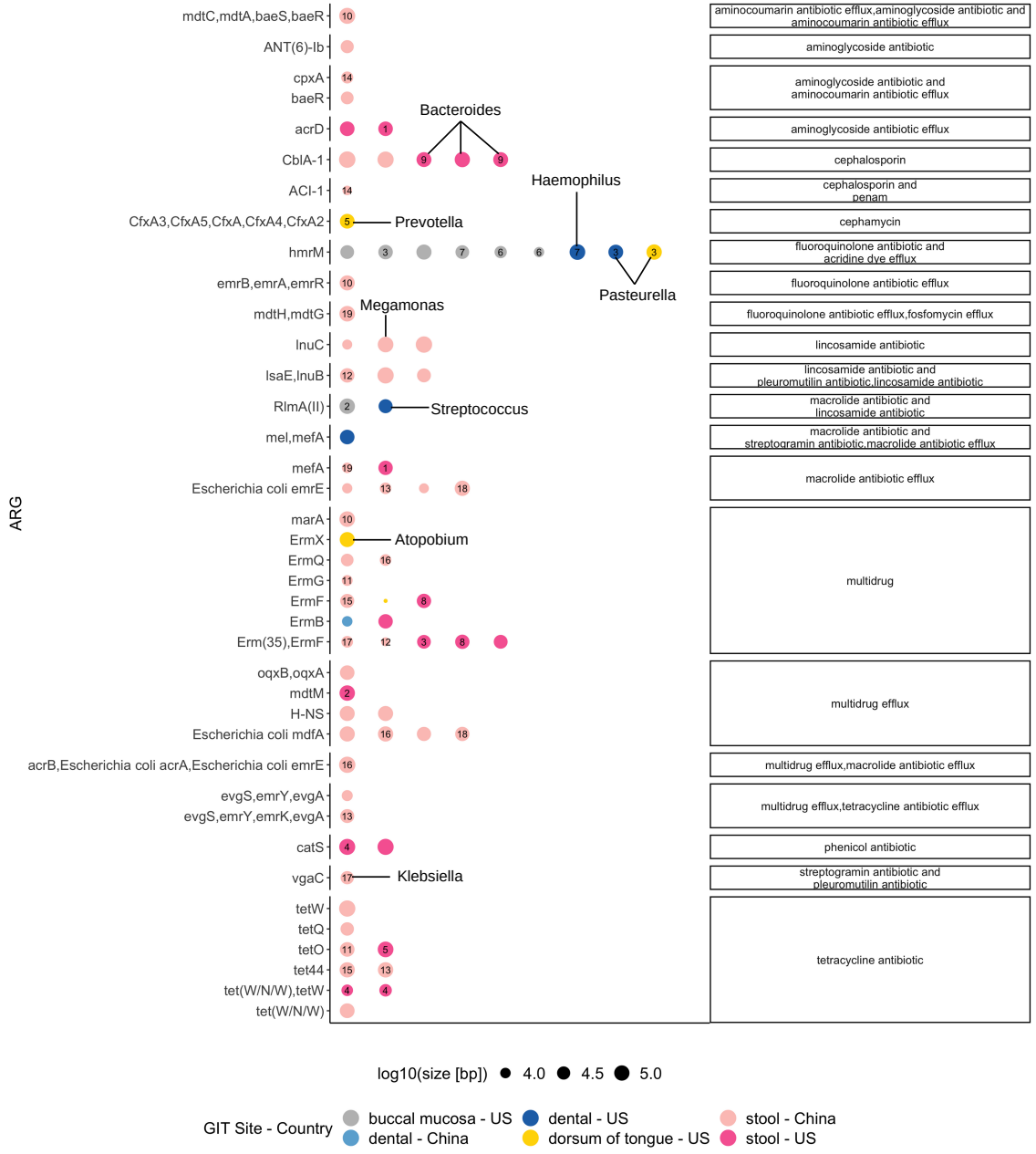

Figure 3.4. ARGs in 77 phages.

Each circle represents a unique phage where circle size is scaled by log10 of the genome size in bp. Circles are colour-coded by GIT site and country, and those with numbers represent phages originating from the same individual. The number marks the particular individual. n = 32 from China and n = 20 from the USA (including longitudinal USA samples). Predicted host genera where identifiable are stated in labels below the circles. ARG classes are described on the right hand side. ARGs that are part of the multidrug class confer resistance to three or more drug classes. Classes that include efflux are those ARGs that code for efflux pumps that pump antimicrobials out of the cell.

3.3.5 Circular jumbo phages are commonly found in the oral cavity but not in the gut

Most phages have a genome size of less than 200 kbp, but 551 phage contigs were found with genome sizes greater than 200 kbp, known as jumbo phages **(Fig. 3.5a)**. 368 of these jumbo phages belong to 108 unique phage clusters, while 183 jumbo phages do not belong to any cluster **(Supplementary Data 3.1: https://tinyurl.com/y4gnasxn)**. 96 genomes were circularised and are located in oral samples, in particular the dorsum of tongue, but not in stool samples **(Fig. 3.5b)**. This is despite the fact that stool samples have the highest proportion of contiguous metagenomic assemblies above 200 kbp compared to oral sites **(Appendix 3: Fig. 3K)**. Circular jumbo phage genomes are not present in saliva from the Philippines, but this could be due to fewer assemblies. Three of the largest circular jumbo phage genomes that are above 300 kbp are all located in dorsum of tongue samples from the USA **(Fig. 3.5c)**. Of the six linear megaphage genomes above 500 kbp, one is found in a stool sample from China, and the others from three dental plaque and two from dorsum of tongue samples from the USA.

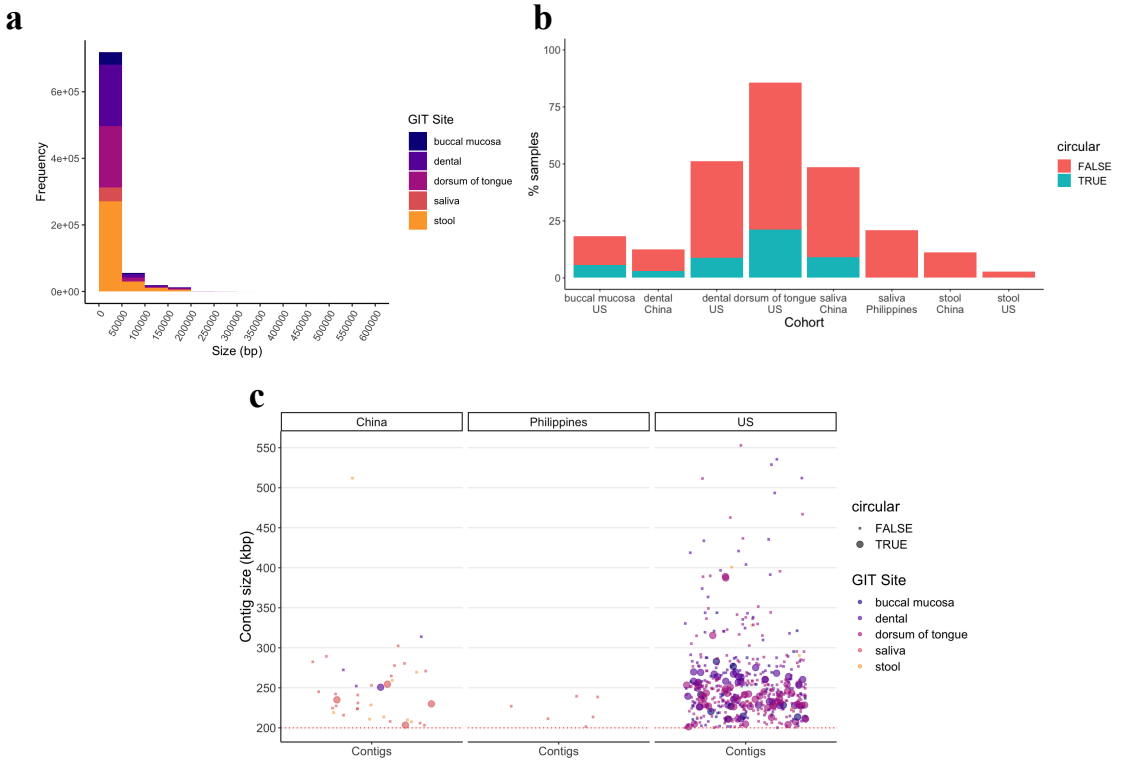

Figure 3.5. Prevalence of jumbo phages.

a) Histogram of phage genome sizes in bp. **b)** Percentage of samples (including USA longitudinal) that contain a jumbo phage (size >= 200 kb) from USA (n = 87 buccal mucosa, n = 90 dorsum of tongue, n = 90 dental plaque and n = 70 stool), China (n = 32 dental plaque, n = 33 saliva and n = 72 stool), and the Philippines (n = 24 saliva). **c)** Sizes of unique jumbo phage contigs found in cohort. Red dashed line represents 200 kb cut-off.

81 circular jumbo phage genomes are found in 22 phage clusters and the remaining 15 are singletons, meaning 37 distinct circular jumbo phage groupings were identified **(Supplementary Data 3.1: https://tinyurl.com/y4gnasxn)**. In some cases, both linear and circular jumbo phage genomes are members of the same phage clusters. 29 phage clusters containing 141 jumbo phages are persistent, i.e. found in the same GIT site in more than two timepoints. Ten phage clusters with 87 jumbo phages in the oral cavity are found across more than one country, specifically, four from China, the Philippines and the USA, five from China and the USA, and one from the Philippines and the USA.

One phage cluster contains circular and linear jumbo phage genomes that are both persistent and found in all three countries.

Most of the predicted protein-coding genes of these jumbo phages have hypothetical functions, while others mainly encode proteins for replication and nucleic acid metabolism **(Fig. 3.6)**. Although ARGs tend to be found in larger genomes of phages, no ARGs were found in these jumbo phage genomes. However, auxiliary metabolic genes may provide an alternative pathway to modulating resistance to antimicrobial chemicals. Glutaredoxin, being the third most prevalent auxiliary metabolic gene, is a redox enzyme that catalyses the reduction of disulphide bonds of substrates with cofactor glutathione[382].

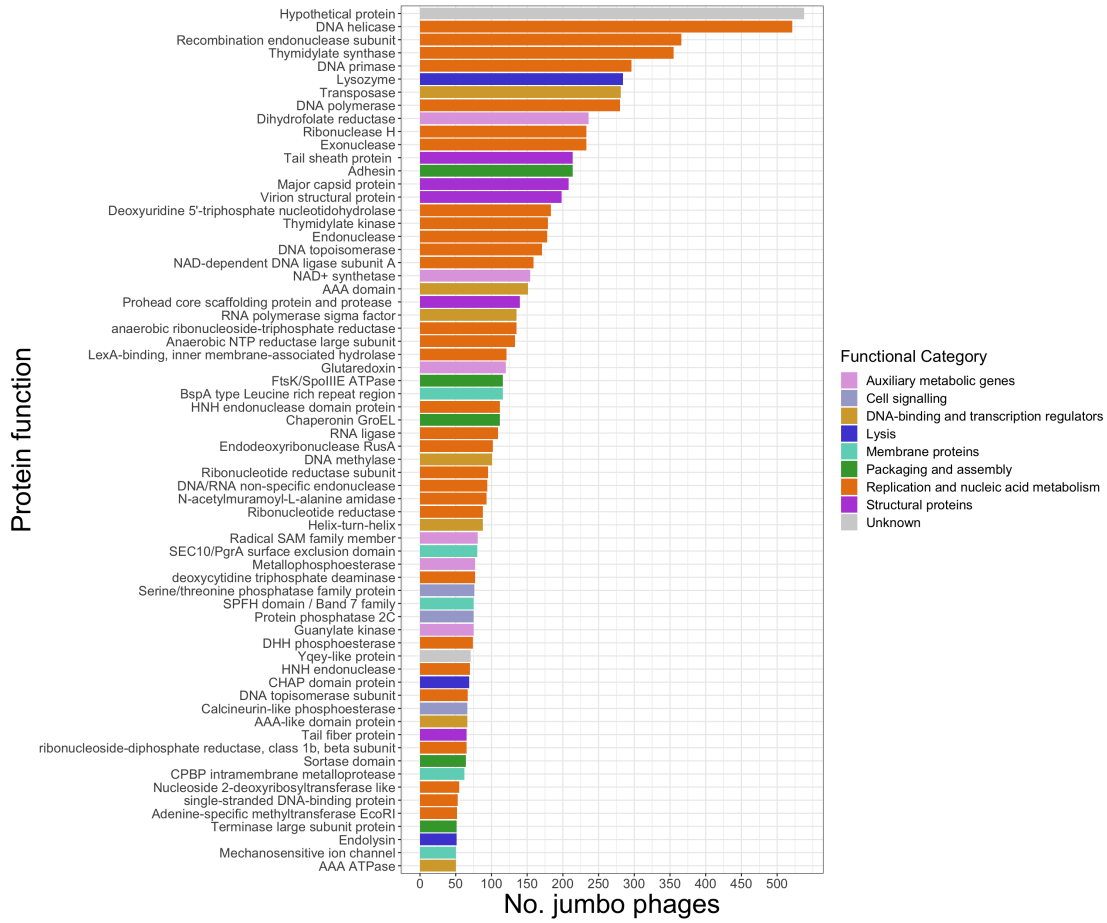

Figure 3.6. Functions and associations of protein-coding gene in jumbo phages.

Number of jumbo phages (50 or greater) containing proteins coloured by functional category. Complete list of functional annotations for each jumbo phage is available here: https://tinyurl.com/yxmej8bu.

3.4 Discussion

Bacteriophages are a major but largely neglected component of the oral microbiome. With their ability to introduce and transport genes between strains, as well as modulating composition through lytic activity, they have a significant impact on the population and function of the microbiome as a whole. Recent advances in bioinformatics and high-throughput sequencing now allow us to explore this community. In this study, we defined the composition of the oral phageome and

compared it with the gut phageome. Firstly, we showed distinct phageome profiles between sites of the GI tract, particularly between the gut and the oral cavity, with sharing of phage clusters between proximal oral sites. Variations in phageome profiles are likely to be associated with differences in bacterial compositions of these sites that are characteristic of the gut and the oral cavity[15]. It has been observed previously that phages, particularly crAss-like, are both stable and individual-specific in the gut[16,29]. Although we did not find the same levels of crAss-like phages in the oral cavity, other phage clusters can persist for months in different oral sites and are more individual-specific than transient phage clusters. However, we cannot rule out that other less persistent phage clusters in some individuals may be attributed to lower abundance phage contigs not being picked up in our metagenomic pipeline.

Very few phage genomes contain putative ARGs, consistent with previous investigations, suggesting phages may not be a major source of MGEs carrying ARGs[96]. Given that all ARG-carrying phage genomes were linear and most of them included integrases, they predominantly have lysogenic behaviour and are likely to be integrated into host DNA as a prophage. Most ARGs associated with these potential prophages mainly encode intrinsic resistance[197]. One of these genes encodes ACI-1, a β-lactamase, found in one stool sample from China. Recently, it has been shown ACI-1 is prevalent worldwide in human gut microbiomes and atypically carried by transposons within propohages[383]. A few phages were associated with ARGs thought to be acquired within the last two decades including *ANT(6)-Ib*[384], *lnu(C)*[385], *lsaE*[386], *mel*, *mefA*, *oqxB*[387] and *vgaC*[388].

Jumbo phages with circularised genomes were identified in most oral sites, particularly on the dorsum of the tongue from USA samples, but were not found in paired stool samples. The lack of circularised jumbo phages in stool samples is interesting given that stool contains large metagenomic assemblies and a higher diversity of phage clusters than saliva and buccal mucosa. Circularised jumbo phages have already been discovered in adult faecal samples from the Bangladesh, Tanzania, Peru and the USA[92,389]. However, in a recent study, 19 circularised jumbo phages were identified in saliva compared to only one in stool samples from pregnant women in the USA, suggesting the oral cavity may have a naturally higher prevalence of jumbo phages[389].

A jumbo phage's large genome size makes it possible for it to carry a broad range of genes[373]. A rich set of both viral and bacterial proteins in jumbo phage genomes are found. No ARGs were found but instead there was a high prevalence of the auxiliary metabolic gene coding for glutaredoxin. It has been shown that *Pseudomonas aeruginosa* with mutated monothiol glutaredoxin was more susceptible to polymyxins, a last-line antibiotic for multidrug-resistant bacteria, although the exact mechanism is not known[390]. Phages may contribute to AMR by more than just HGT of ARGs through encoding metabolic processes that can bypass antimicrobial targets. Bacteriophage genes that are homologous with bacterial genes are likely to have been acquired from host sequences during a previous infection event[391]. Lysogenic phages are also required to adapt their genetic machinery to integrate and cooperate with the host genome. This may explain why we found a relationship between the incidence of specific jumbo phage proteins and their predicted hosts. Since we could only predict bacterial hosts for 11.8% of phages (including jumbo phages), functional protein profiles may be able to

aid host predictions, especially where CRISPR spacer reference information is lacking. Phage hosts can also be refined more accurately at a metagenomic level using techniques such as SMRT metagenomic sequencing[277], single-cell viral tagging[284] or using *de novo* computational methods[391].

The dorsum of the tongue and dental plaque contain the highest diversity of phage clusters compared to other sites, as well as the greatest prevalence of jumbo phages. This could be because bacteria in these sites are largely located in biofilms, which serve to protect microbial communities as well as phages in hostile environments, and bacteriophages themselves have been shown to promote biofilm formation[392]. Dense, protective layers of microbes and their EPS matrix in biofilms could provide an opportunistic environment for the evolution of disparate phages, including jumbo phages with extended genetic machinery and metabolic capacity[373].

Although we only find circularised jumbo phages in oral sites and a higher diversity of phage clusters on the dorsum of the tongue, we cannot rule out that the gut also may contain comparable levels of diverse phages. It is possible that the total phage composition in the gut could be misrepresented in stool metagenomes. Unlike sampling oral sites of particular physical locations, such as dental plaque, surface of the tongue and mucosa from the inner cheek, sampling faecal matter is unable to capture the spatial microbial community structure of the gut, both radially and longitudinally[393]. The microbial community structure of the human gut is of considerable importance to the dynamics of bacteriophage populations. Studies applying transmission electron microscopy of human gut biopsies have shown higher levels of bacteriophage

colonisation in the colonic mucosa layer[394] than in faeces and caecum[395]. Several existing models (piggyback-the-winner[396] and kill-the-winner[397]) have been proposed to explain this biogeographical variation in bacteriophage density across the radial direction in the gut[4]. Lysogeny dominates in low virus to microbe ratios in the gut lumen by the piggyback-the-winner model, but switches to the lytic cycle in higher microbial densities of the mucin layer by the kill-the-winner model.

Already, significant contributions have been made in the discovery of multitudes of novel bacteriophage genomes in humans, animals, outdoor and contained environments[92,372,374,389]. However, further investigations into how genetic and functional variations are manifested at more specific biogeographical sites are required. With better structural resolution, it will be possible to capture the extent of genetic diversity and profile the dynamics of distinct phages in different human microbiomes, such as biofilms on the surface of the tongue. Clinically, there has been renewed interest in using phage therapy with other antimicrobial therapies to control biofilms[392]. Phage therapy exploits the action of lytic phages that infect and lyse specific bacteria. In the future, combinations of genotypic attributes that influence phage persistence and interactions with their hosts could be used to select or design phages for eradicating chronic infections, like periodontitis and dental caries caused by consortia of organisms in biofilms[398]. Given there seems to be very little risk of acquiring ARGs from phages, phage therapy should be seriously considered for eradicating AMR infections where last resort antimicrobial treatments fail.

Chapter 4: Plasmids and Resistance Plasmids

4 Plasmids and Resistance Plasmids

4.1 Introduction

Plasmids are DNA molecules that range from less than a kilobase to a megabase in size and replicate independently within a host cell (replicons). They are separate from chromosomal DNA and most commonly found in a circular, double-stranded form in bacteria. Some plasmids include genes that encode traits for resistance to antimicrobials and heavy metals, also known as resistance plasmids. Many studies focus on plasmids from single or several strains isolated from microbial colonies. This has been useful for determining whether particular opportunistic pathogens have a role in the propagation of known resistance plasmids in GIT sites, such as *Salmonella* in the gut[220] and *Fusobacterium nucleatum* in the oral cavity[399]. However, targeted non-metagenomic approaches like this cannot be applied to discovering new plasmids and resistance plasmids. Alternatively, high-throughput TRACA can capture plasmids by inserting a transposon with an origin of replication that are then cloned into a host bacterium[226] or inverse-PCR[230] can be applied to discover small plasmids in metagenomic samples. However, both these approaches are highly selective for smaller plasmids.

In contrast to isolation and cultivation studies, there has previously been very little research into resistance plasmids from short-read whole metagenomics, including in the GIT. Due to their large size, repetitive sequences and modular structure, it is especially challenging to construct plasmid genomes from short-read sequences. Thus it becomes harder to detect larger plasmids (> 50 kbp) than smaller plasmids (< 10 kbp) from short-

read metagenomic data using currently available tools[400]. Multiple studies have applied other biological approaches to profile plasmids (including plasmids carrying ARGs) from short-read metagenomics, such as proximity ligation[279–281]. However, even when plasmid genomes are isolated, it is especially challenging to identify the bacterial host of origin using computational approaches alone. Proximity ligation and non-computational binning approaches, such as DNA methylation motifs, can be used to overcome some of these problems, thereby aiding host predictions[277].

Computational tools are being developed and adapted to support plasmid detection from short-read metagenomic data, such as plasmidSPAdes, Recycler, cBar and PlasFlow. The most recent and well-established tools, plasmidSPAdes and Recycler, use assembly graphs that represent the final assemblies of metagenomes. An assembly graph consists of nodes of sequences that are connected by edges representing overlaps. The assembly tools resolve paths across the assembly graph to generate continuous contigs. However, sometimes large assembly graphs contain a mixture of both chromosomal and plasmidic edges where sequences are common to both types of DNA. Moreover, plasmids with repeated sequences form loops that traverse the edges of these repeats more than once. To overcome these challenges, both tools rely on finding a complete circular plasmid by taking the path along the graph with the most uniform coverage of edges. Although this approach can work well with single isolates where uniform covered cycles can be easily distinguished, this can be problematic for complex metagenomic datasets where coverage gets confounded by the presence of multiple genomes. Both plasmidSPAdes and Recycler were benchmarked against datasets containing short reads simulated from individual bacterial genomes with 148 plasmids[400]. PlasmidSPAdes and Recycler were

found to have low precisions (number of correct predictions divided by the total number of predictions) of 0.78 and 0.30, respectively, meaning these tools are inapplicable to metagenomic datasets with varying coverages. Recently, a plasmid assembly tool called metaplasmidSPAdes, was adapted from plasmidSPAdes, which was able to discover novel plasmids from short-read metagenomic data, including ones carrying ARGs[269]. The tool iteratively traverses subgraphs while increasing coverage depth until cyclic contigs of candidate plasmids are found. The plasmids are verified by searching their protein-coding regions against HMMs of putative plasmid proteins. The frequency of matches is used to train a naïve Bayesian classifier to remove noise of false matches and classify the contig as either plasmidic or chromosomal.

Here, metaplasmidSPAdes was applied to create a catalogue of plasmids from short-read metagenomes across the human GIT. Furthermore, the composition of plasmids and ARG-carrying plasmids were examined across human oral sites and the gut from China and the USA.

4.2 Methods

4.2.1 Creating a plasmid catalogue

A catalogue of plasmids was created using 624 metagenomic samples from two cohorts:

1) Human Microbiome Project (referred to as USA)[336] containing buccal mucosa (n = 61: 17 with one, 27 with two, 16 with three and 1 with six timepoints), dorsum of tongue (n = 61: 11 with one, 26 with two, 22 with three and 2 with four

timepoints), dental plaque (n = 61: 11 with one, 30 with two, 17 with three, 1 with four and 2 with six timepoints), stool (n = 61: 10 with one, 31 with two, 17 with three, 2 with four and 1 with six timepoints).

2) Healthy control samples from a Chinese rheumatoid arthritis study[331] containing dental plaque (n = 29), saliva (n = 29) and stool (n = 29).

The sequences were trimmed and quality controlled (processed from Chapter 2: 2.2.3.1 and 2.2.3.2). They were then assembled to plasmid contigs using the metaplasmidSPAdes v3.14.0 software with parameters *--plasmid --meta*[269]. The *plasmidverify.py* script from the metaplasmidSPAdes software was then applied to the output circular scaffolds to distinguish putative plasmids (https://github.com/ablab/plasmidVerify). It does this by identifying protein coding genes using Prodigal v2.6.3[210], identifying plasmid proteins from an HMM search of the Pfam database[251] (downloaded 2nd September 2019) using hmmsearch 3.2.1[249], and applying a Naïve Bayes Classifier to classify a genome as either plasmid or chromosomal. The custom helper script, *extract_plasmid_contigs.py* in the Github repository (https://github.com/blue-moon22/plasmidome), was applied to filter putative plasmids from the *plasmidverify.py* output. The plasmid contigs were combined and clustered into a non-redundant plasmid catalogue with the software suite CD-HIT v4.8.1[401] using psi-cd-hit.pl with parameters *-circle 1 -c 0.9 -prog blastn -exec local*, whereby circular genomes are aligned with a clustering threshold of 0.9 using BLASTn v2.9.0. The plasmid taxonomy was determined by aligning the non-redundant plasmid catalogue against the PlasmidFinder database[103] (downloaded on 3rd June 2020) using BLASTn v.2.9.0 with parameter *-evalue 1e-5*.

4.2.2 Plasmid annotation in metagenomes

Metagenomic reads were mapped against the plasmid catalogue with Bowtie2 v2.3.5.1[195] with parameter --*very-sensitive-local*. The output file was indexed, sorted and converted into a BAM file using Samtools v1.9.0[343]. The depth coverage (number of reads mapped) and the breadth coverage (proportion of plasmid contig that had been mapped) was then evaluated using BEDTools v2.29.0[344].

4.2.3 ARG annotation of plasmids in metagenomes

The plasmid catalogue was queried against CARD v3.0.0[197] for ARGs using BLASTn v2.10.0 with an e-value threshold of 1e-5. Using the *get_reads_mapped_to_ARGs.py* helper script, ARGs of plasmids were identified for each sample if at least one read that mapped to the plasmid also mapped to the ARG region within the plasmid genome.

4.2.4 Plasmid analysis

The code and data that this analysis refers to is available here: https://github.com/blue-moon22/plasmidome. The analysis was run using the script *thesis_analysis.R* in R v3.6.1. Plasmid genomes with a breadth coverage of less than 75% were filtered for each sample. This left 615 metagenomes from China (n = 29 dental plaque, n = 29 saliva, n = 29 stool) and the USA (n = 61 buccal mucosa: 17 with one, 27 with two, 16 with three and 1 with six timepoints, n = 61 dental plaque: 11 with one, 30 with two, 17 with three, 1 with four and 2 with six timepoints, n = 61 dorsum of tongue: 11 with one, 26 with two, 22 with three and 2 with four timepoints, and n = 61 stool samples: 12 with

one, 33 with two, 14 with three and 2 with four timepoints). As some of the plasmid genomes may represent the same plasmid and may be redundant, plasmid genomes were organised into plasmid clusters using a similar strategy to vConTACT2 with viral sequences[378]. Pairs of plasmid genomes were connected in a network if the smallest genome length, L_1, was greater than 99% of the other genome length, L_2. The weight, w, between each pair was calculated as:

$$w = \frac{L_1}{L_2} S$$

where S is the proportion of unique plasmid protein-coding genes that are shared between the two plasmid genomes. Pairs of plasmids connected by w greater than 0.8 were assigned the same plasmid cluster number # with prefix name *PC_#*. Others kept their original contig name and remained as plasmid singletons. Plasmid clusters and singletons that contained ARGs were then identified.

4.2.5 Plasmid diversity

To find differences in β-diversity of plasmid profiles between groups of individuals, the Bray-Curtis dissimilarities **(eq. 3.1)** of plasmid singleton and cluster incidence (presence or absence) profiles were computed between individuals and visualised using NMDS. Silhouette analysis **(eq. 2.1)** of *k*-medoids using the cluster package v2.1.0 was used to select the number of distinct groups with the largest average silhouette width.

The α-diversity was calculated as the plasmid singleton and cluster richness which is the number of unique plasmid singletons and clusters for each sample. Only samples with greater than 20 plasmids were included. The plasmid richness was compared between

paired GIT sites from the paired samples. Since the number of plasmid contigs in each sample is significantly linearly correlated with plasmid richness ($p < 2.2x10^{-16}$), the number of plasmids for each sample were subsampled. Subsampling entailed removing a plasmid from a sample based on the probability of finding a particular plasmid singleton/cluster. This was repeated until the number plasmids reached the smallest number of plasmids of an original sample within a paired comparison. Individuals with samples containing less than or equal to three plasmids were excluded from the group comparison to leave the following samples: China dental plaque vs. saliva (n = 28); China stool vs. saliva (n = 28); China stool vs. dental plaque (n = 29); USA buccal mucosa vs. dental plaque (n = 35); USA buccal mucosa vs. dorsum of tongue (n = 35); USA buccal mucosa vs. stool (n = 35); USA dental plaque vs. dorsum of tongue (n = 59); USA dental plaque vs. stool (n = 59); and USA dorsum of tongue vs. stool (n = 61). The plasmid richness between samples in each group was tested for statistical significance with a Two-sided Wilcoxon Rank Sum Test.

4.2.6 Microbial composition

MetaPhlAn2 v2.6.0[349] was used to identify the composition of bacteria and archaea from all 326 metagenomic samples with plasmids (as described in Chapter 2, Section 2.2.3.9). Procrustes analysis was applied to visualise the superposition of NMDS dimensions between plasmid singleton/cluster incidence profiles and microbial genera incidence profiles using the protest function in the vegan package v2.5.6 in R. PROTEST was performed with 999 permutations to a significance of p = 0.001.

4.2.7 Longitudinal analysis of plasmids

The stability of plasmids was investigated by computing the number of timepoints within two years for every plasmid cluster and singleton found in each GIT site of the same individuals from longitudinal USA data. The proportion of plasmid clusters and singletons was calculated for each number of sampled timepoints available for each individual and GIT site: buccal mucosa (n = 17), dental plaque (n = 20), dorsum of the tongue (n = 24) and stool (n = 16). Persistent plasmid clusters/singletons were defined as being found in three or more timepoints, whereas transient phage clusters are defined as being found in less than three timepoints. The Bray-Curtis dissimilarities were computed between persistent and transient plasmid cluster/singleton incidence profiles from GIT sites of individuals containing both persistent and transient plasmid cluster/singletons. NMDS was applied to scale the dissimilarity into a two-dimensional ordination using the metaMDS function in the vegan package v2.5.6 in R. PERMANOVA analysis was performed using the adonis function in the vegan package.

4.3 Results

4.3.1 Plasmid composition across GIT sites

3,929 unique plasmid sequences were identified using metaplasmidSPAdes[269] from 458 oral sites and 157 paired gut metagenomes. 2,536 plasmids with similar sizes and plasmid gene profiles were clustered into 852 plasmid clusters, while the remaining 1,393 plasmids were left as singletons. Most plasmids are shared in fewer than 50% of samples (excluding longitudinal USA samples) for each GIT site and country **(Fig.**

4.1a). GIT sites and country cluster by the incidence of plasmid singletons/clusters that are shared in more than half of samples **(Fig. 4.1b)**. In particular, samples from stool and oral cavity have distinct profiles, while there are also differences in dental plaque and stool samples between USA and China.

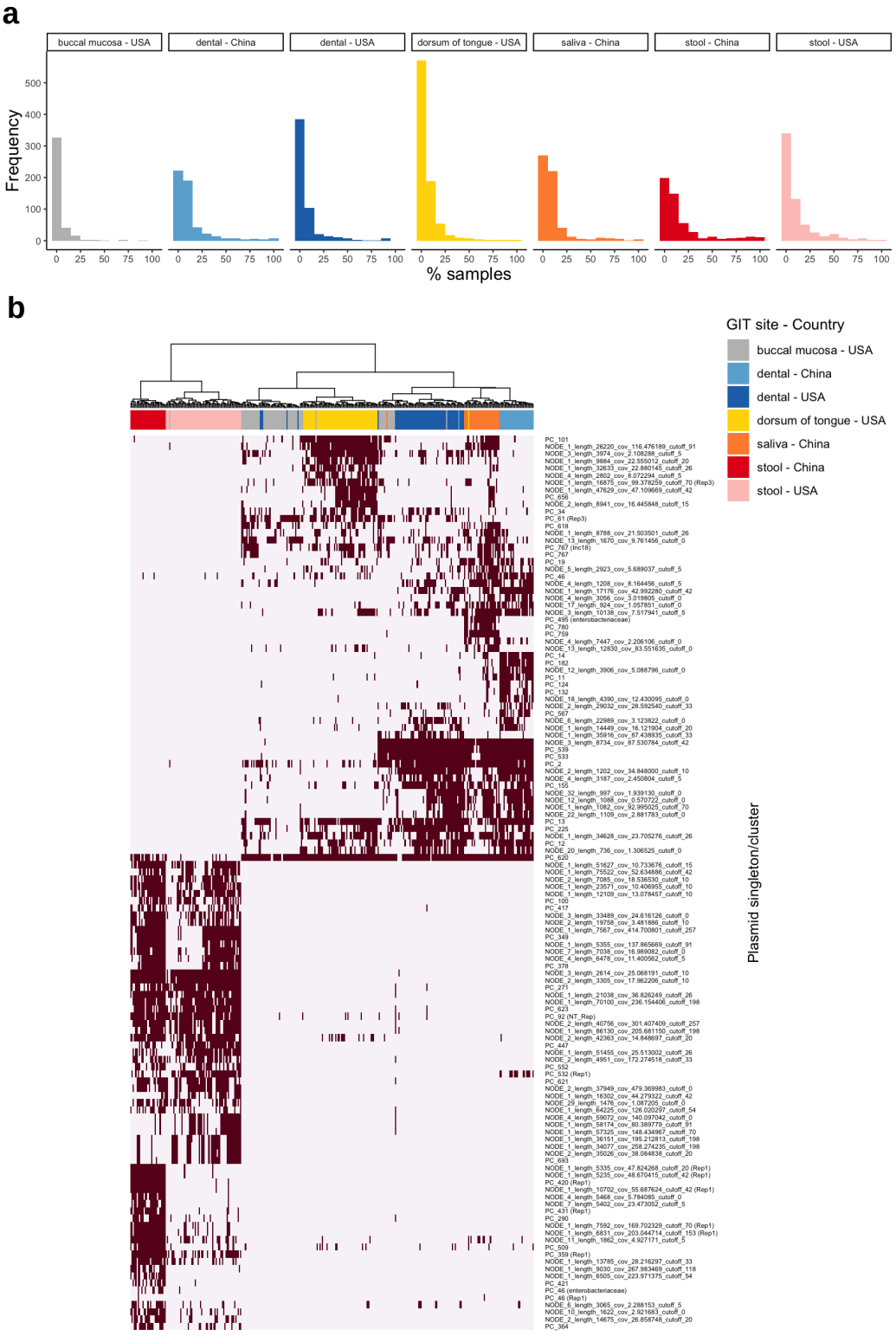

Figure 4.1. Plasmid incidence.

a) Frequency of plasmid singletons and clusters found in a percentage of samples for each GIT site and country. **b)** Incidence of plasmid singletons and clusters found in greater than 50% of samples for each GIT site. USA buccal mucosa (n = 61), dorsum of tongue (n = 61), dental plaque (n = 61) and stool (n = 61); China dental plaque (n = 29), saliva (n = 29) and stool (n = 29).

Variation in plasmid profiles across GIT sites of individuals was evaluated using β-diversity. Specifically, the Bray-Curtis dissimilarity metric was applied to plasmid singleton/cluster abundance profiles. Sample plasmid profiles were clustered by NMDS and the number of groups was found using Silhouette analysis on *k*-medoids on two dimensions. Four groups of plasmid profiles were found between GIT sites, with the largest separation between Groups 1/2/4 consisting of oral samples and Group 3 containing mostly stool samples **(Fig. 4.2a)**. Most dental plaque from China and the USA and saliva from the USA are located in Group 1, most buccal mucosa from the USA reside in Group 2, and most dorsum of the tongue samples from the USA are found in Group 4 **(Fig. 4.2b)**. Shared plasmid singletons/clusters are predominantly across dental plaque, buccal mucosa and dorsum of the tongue samples **(Fig. 4.2c)**.

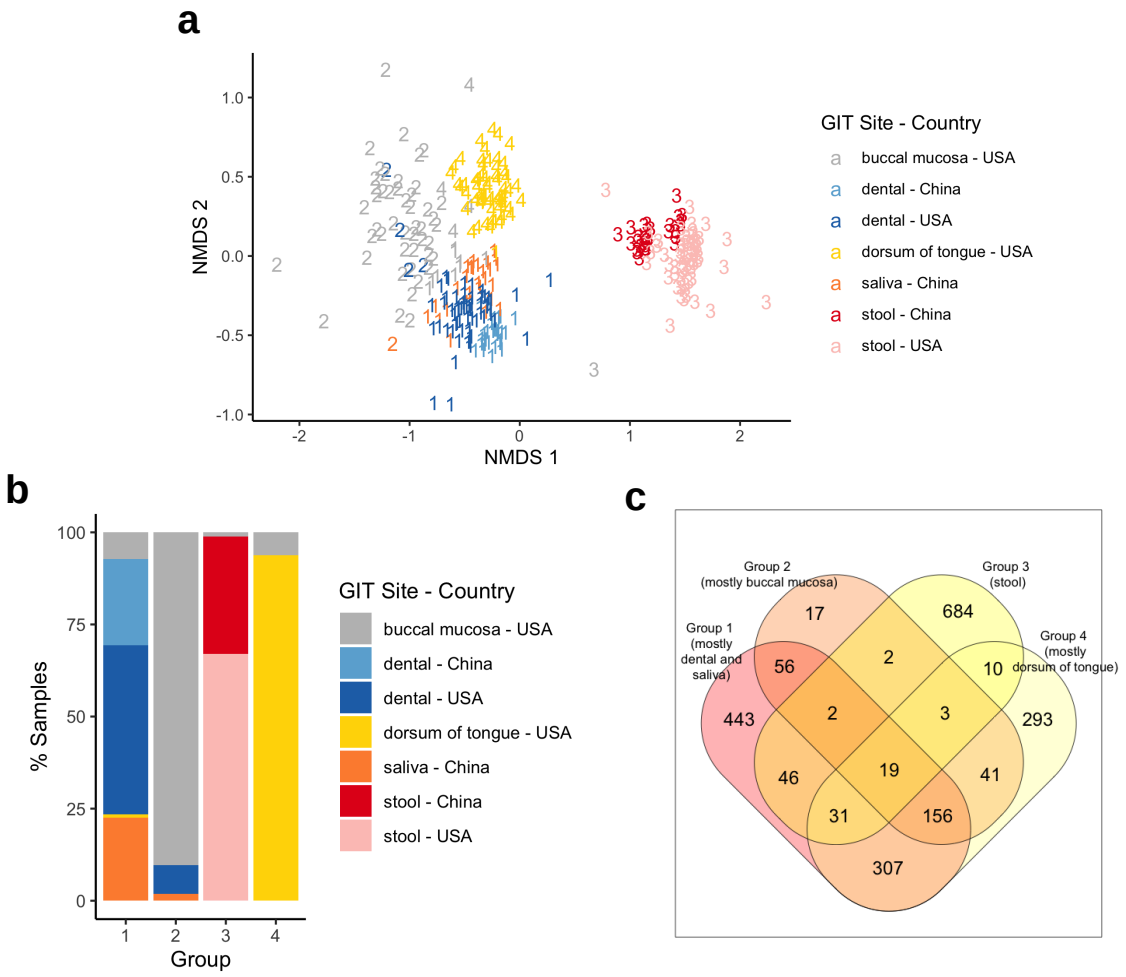

Figure 4.2. Plasmid composition across GIT sites.

a) NMDS of Bray-Curtis dissimilarities between plasmid incidence profiles of samples. The points are shown as numbers representing the groups from *k*-medoids clustering, where number of groups, *k*, has the largest average silhouette width. USA buccal mucosa (n = 61), dorsum of tongue (n = 61), dental plaque (n = 61) and stool (n = 61); China dental plaque (n = 29), saliva (n = 29) and stool (n = 29). **b)** Percentage of samples in each group from 1a, labelled by GIT site and country. **c)** Number of plasmid singletons/clusters in each group.

The plasmid richness of samples was compared between GIT sites. The plasmid richness was evaluated from subsampled plasmid profiles (Methods: Section 4.2.5) as the number of unique plasmid clusters and singletons per sample. In the USA, the dorsum of the tongue has a higher richness than buccal mucosa ($p = 3.12 \times 10^{-6}$; Two-sided Wilcoxon Rank Sum Test), dental plaque ($p = 6.38 \times 10^{-4}$) and stool samples (p =

0.0223) **(Fig. 4.3)**. Buccal mucosa also had the lowest richness, being significantly lower than dental plaque ($p = 1.54x10^{-4}$) and stool samples ($p = 2.85x10^{-4}$). In China, stool samples are significantly richer in plasmid singletons/clusters than dental plaque ($p = 0.00128$) and saliva ($p = 0.0109$), while dental plaque has a greater richness than saliva ($p = 0.0226$).

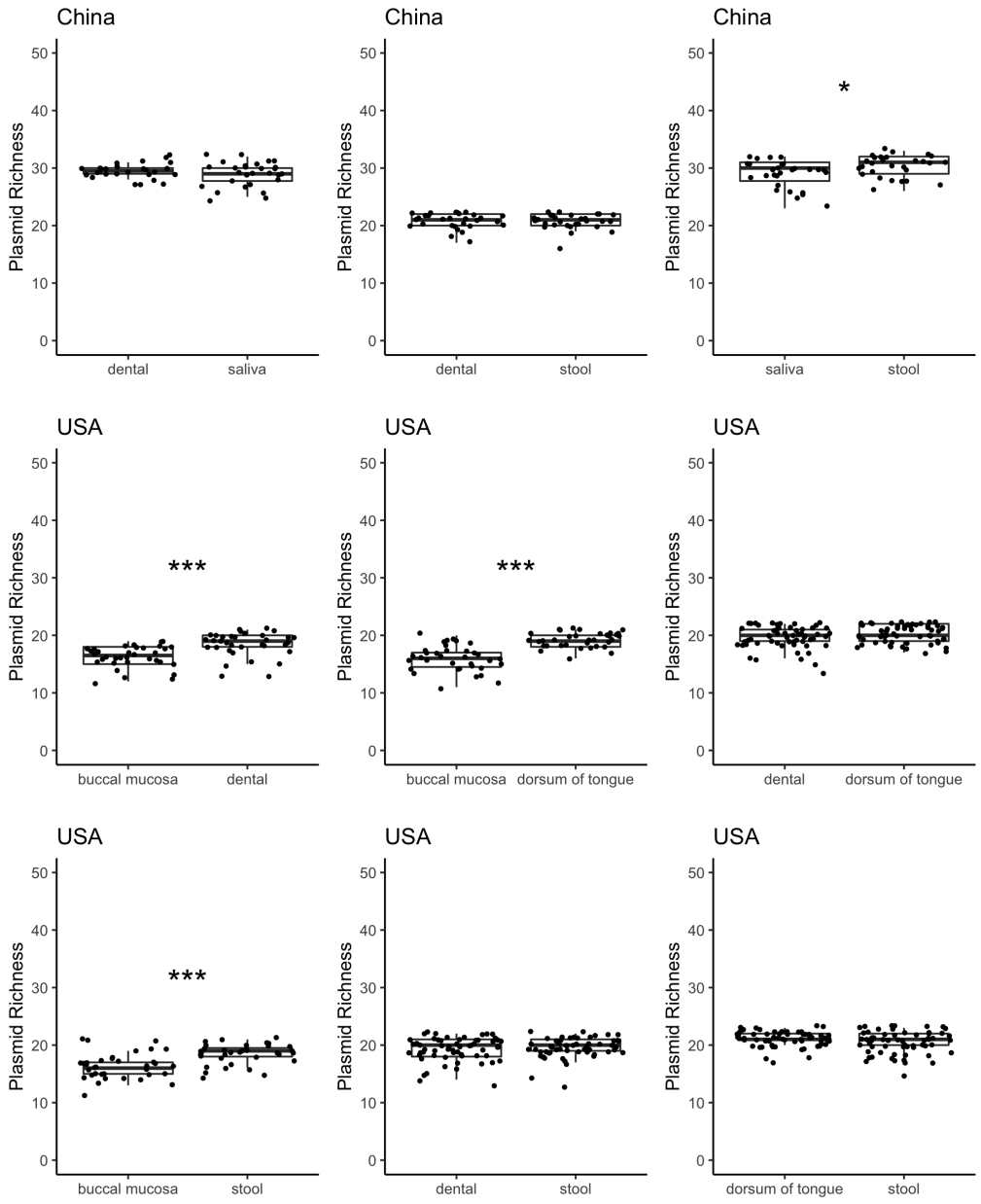

Figure 4.3. Plasmid Cluster Richness between paired GIT sites.

Plasmid Cluster Richness is defined as the number of unique plasmid singletons/clusters in a sample that is subsampled to the smallest number of non-unique clusters. Plasmid Cluster Richness is calculated for paired samples of individuals from China (dental plaque and saliva: n = 28, stool and saliva: n = 28, stool and dental plaque: n = 29) and the USA (buccal mucosa and dental plaque: n = 35, buccal mucosa and dorsum of tongue: n = 35, buccal mucosa and stool: n = 35, dental plaque and dorsum of tongue: n = 59, dental plaque and stool: n = 59, dorsum of tongue and stool: n = 61) with Two-sided Wilcoxon Rank Sum Test ($p < 0.05$ as *, < 0.01 as **, < 0.005 as ***). Centre line is median, box limits are upper and lower quartiles, whiskers are 1.5x interquartile ranges and points beyond whiskers are outliers.

Procrustes analysis was applied to match NMDS dimensions of plasmid and microbial profiles. Microbial profiles were processed from samples using Metaphlan2[349] and NMDS was applied to reduce the profile to two dimensions. PROTEST was then used to determine whether two profiles showed significant association. Microbial composition and plasmid profiles correlate and co-locate by GIT site, especially between stool and dental samples **(Fig. 4.4)** (0.783 to a significance of p = 0.001 in PROTEST).

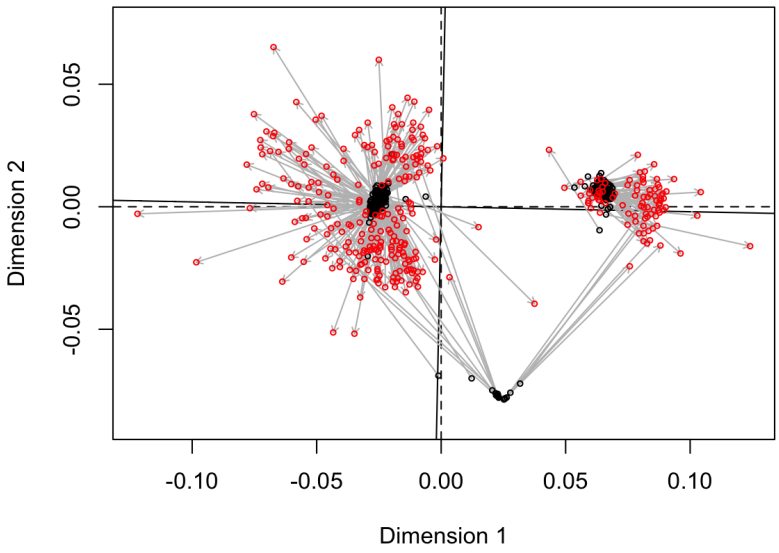

Figure 4.4. Overlay of microbiome composition and plasmid composition.
Procrustes rotation of NMDS coordinates between microbial genera profile (black points) and plasmid incidence profile (red points). Correlation in symmetric Procrustes rotation = 0.783 (p = 0.001; 999 permutations; PROTEST).

4.3.2 Stability of plasmids across longitudinal metagenomes

To show whether plasmids are stable over time, plasmid singletons and clusters were profiled in longitudinal samples from the USA taken over a two-year period with a minimum of two and maximum of six sampling timepoints. The proportions of total

plasmid singletons/clusters drop over time until the third timepoint across all GIT sites **(Fig. 4.5a)**. Due to the small number of samples available with four or more timepoints, it is unclear whether the decrease continues.

To compare the stability profiles between individuals, the plasmid singletons/clusters were separated empirically into persistent and transient plasmid singleton/cluster profiles. Persistent plasmid profiles were defined as being present in three or more timepoints in a given GIT site, whereas transient plasmid profiles are found in less than three timepoints. There are 61 (buccal mucosa), 188 (dental plaque), 381 (dorsum of the tongue) and 315 (stool) persistent plasmid singletons/clusters, and 345 (buccal mucosa), 488 (dental plaque), 719 (dorsum of the tongue) and 292 (stool) transient plasmid singletons/clusters. NMDS visualisations of persistent and transient plasmid profiles from longitudinal samples show greater clustering from the same individuals of persistent **(Fig. 4.5b)** compared to transient profiles **(Fig. 4.5c)**. PERMANOVA was applied to persistent and transient plasmid profiles for each GIT site to find which profile has the highest individuality. A greater percentage of variance in persistent than transient plasmid singletons/clusters can be explained by variability between individuals for all GIT sites. 71.4% (buccal mucosa), 67.4% (dental plaque), 74.5% (dorsum of the tongue) and 85.3% (stool) variance of persistent plasmid profiles and 43.8% (buccal mucosa), 41.4% (dental plaque), 44.3% (dorsum of the tongue) and 44.6% (stool) variance of transient plasmid profiles can be explained by the individual ($p < 0.001$, PERMANOVA). Prevalences of persistent and transient plasmid singletons/clusters vary across these individuals **(Fig. 4.5d)**. The percentages of plasmid singletons/clusters from one individual also seen in another are significantly higher for persistent than

transient plasmids in stool (84.1± 7.8 persistent, 75.7 ± 10.7 transient; p = 0.0167). In contrast, there is a significantly higher percentage of transient than persistent plasmids shared in at least one other individual in dorsum of the tongue samples (71.0 ± 11.1 persistent, 83.9 ± 7.2 transient; p = 9.26×10^{-6}). No significant difference was found for buccal mucosa (66.7 ± 12.8 persistent, 69.2 ± 14.0 transient; p = 0.648).

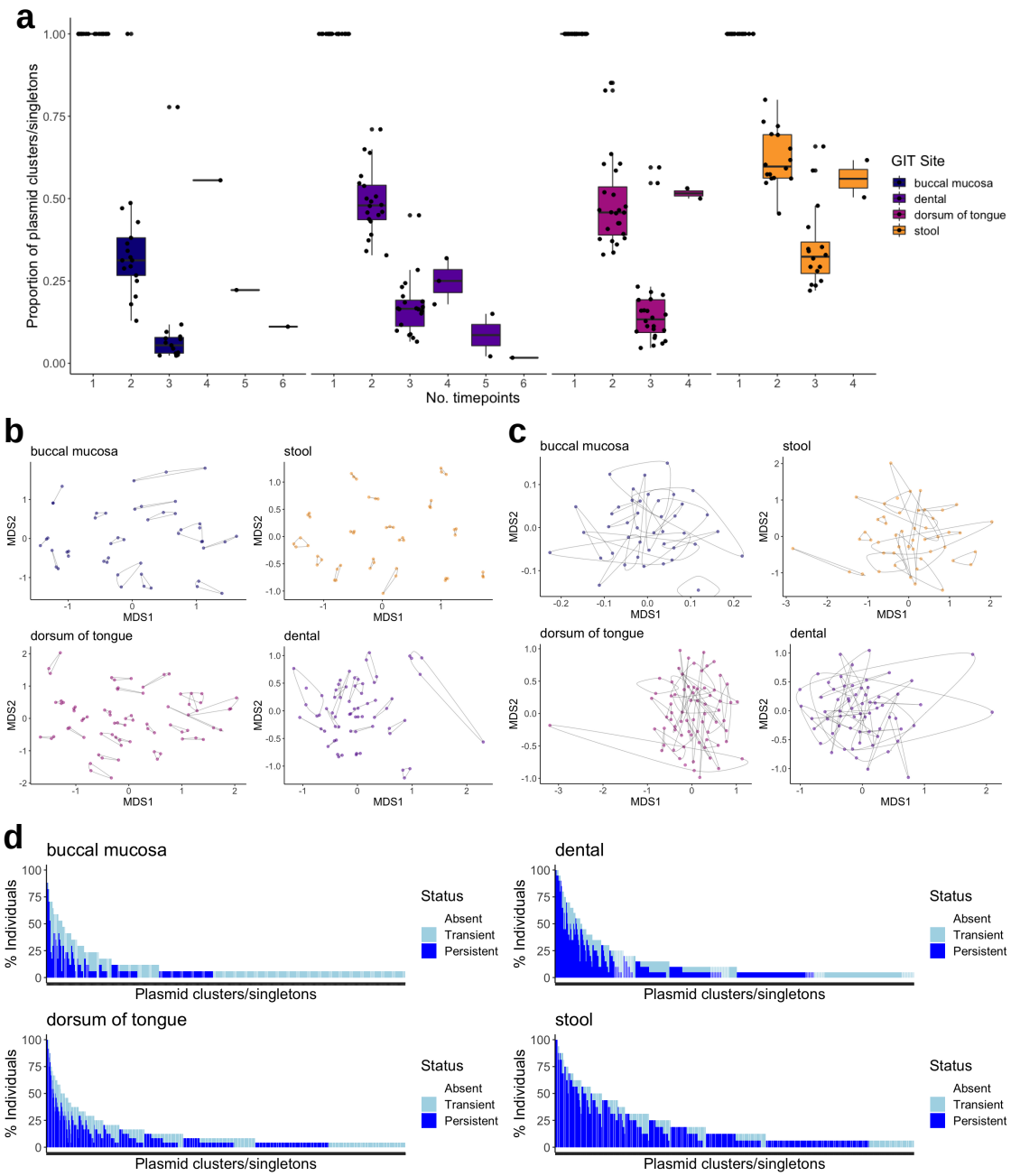

Figure 4.5. Plasmid stability in longitudinal USA GIT sites.

a) Proportion of plasmid clusters/singletons in one to six timepoints for USA individuals with at least three sampling timepoints (buccal mucosa: n = 17, dental plaque: n = 20, dorsum of the tongue: n = 24, stool: n = 16). Non-metric multidimensional scaling of the Bray-Curtis dissimilarities between plasmid cluster/singleton incidence profiles of samples with **b)** persistent and **c)** transient plasmid clusters/singletons. Points represent samples and lines joining points represent grouping samples from the same individual and GIT site. **d)** Prevalence of transient and persistent plasmid clusters/singletons in the same individuals ordered by decreasing prevalence of total and persistent plasmid clusters/singletons.

4.3.3 Resistance plasmids in both oral and gut metagenomes

Plasmid sequences were queried against CARD[197] to identify ARG-carrying plasmids. The ARG classes as well as the specific ARG types were also characterised. Out of the 2,245 (852 plasmid clusters and 1,393 singletons), only 37 were detected to carry ARGs. Eight of those were known plasmids matching to the PlasmidFinder database[103]. A total of 21 ARG homologues are found in plasmids of metagenomic assemblies across all 321 human GIT samples **(Fig. 4.6)**. 13 out of the 21 ARG homologues are located in oral samples: aminoglycoside resistant *APH(3')-Ia*, cephamycin resistant *CfxA2*, multidrug resistant *ErmB* and *ErmX,* lincosamide resistant *lnuC* and *lsaC*, sulfonamide and sulfone resistant *sul2*, two ARG homologues from the multidrug resistant TEM β-lactamase family ARGs, tetracycline resistant *tet(C)* and *tet(G)* by the efflux pump mechanism, and tetracycline resistant *tetM* and *tetO*. 11 out of the 21 ARG homologues are found in the gut: aminoglycoside resistant *APH(3')-IIIa*; *Campylobacter coli* chloramphenical acetytransferase ARG resistant to phenicols; cephamycin resistant *CfxA6*; multidrug resistant ARGs *ErmB*, *ErmF* and *ErmQ*; cephalosporin and penam resistant *OXA-85*; and tetracycline resistant *tetM*, *tetO*, *tetQ* and *tetW*.

Most ARG-carrying plasmids and plasmid clusters are either shared between oral sites or are exclusively in stool samples. Most plasmids in stool samples are found in both China and USA cohorts, and carry ARGs conferring resistance to cephalosporin and penams, cephamycin, glycylcycline and tetracycline antimicrobials, lincosamide, multidrug, phenicol and tetracycline antimicrobial classes. Plasmids in oral sites from China carry a diversity of ARGs conferring resistance to aminoglycoside, cephamycin, lincosamide, multidrug (including penems), sulfonamide/sulfone and tetracycline efflux pumps. All ARG-carrying plasmid singletons/clusters carry a single ARG with the exception of PC_495 that includes two ARGs, aminoglycoside resistant *APH(3')-Ia* and *tet(C)* conferring resistance to tetracycline efflux pumps. The *tetM* ARG, being the most prevalent plasmid-associated ARG in oral sites, is only located on one plasmid cluster PC_620, whereas *ErmB* is located on nine plasmid clusters and singletons. The ARGs from the TEM β-lactamase family are also broadly spread over four different plasmid clusters and singletons, and are found only in oral sites from China.

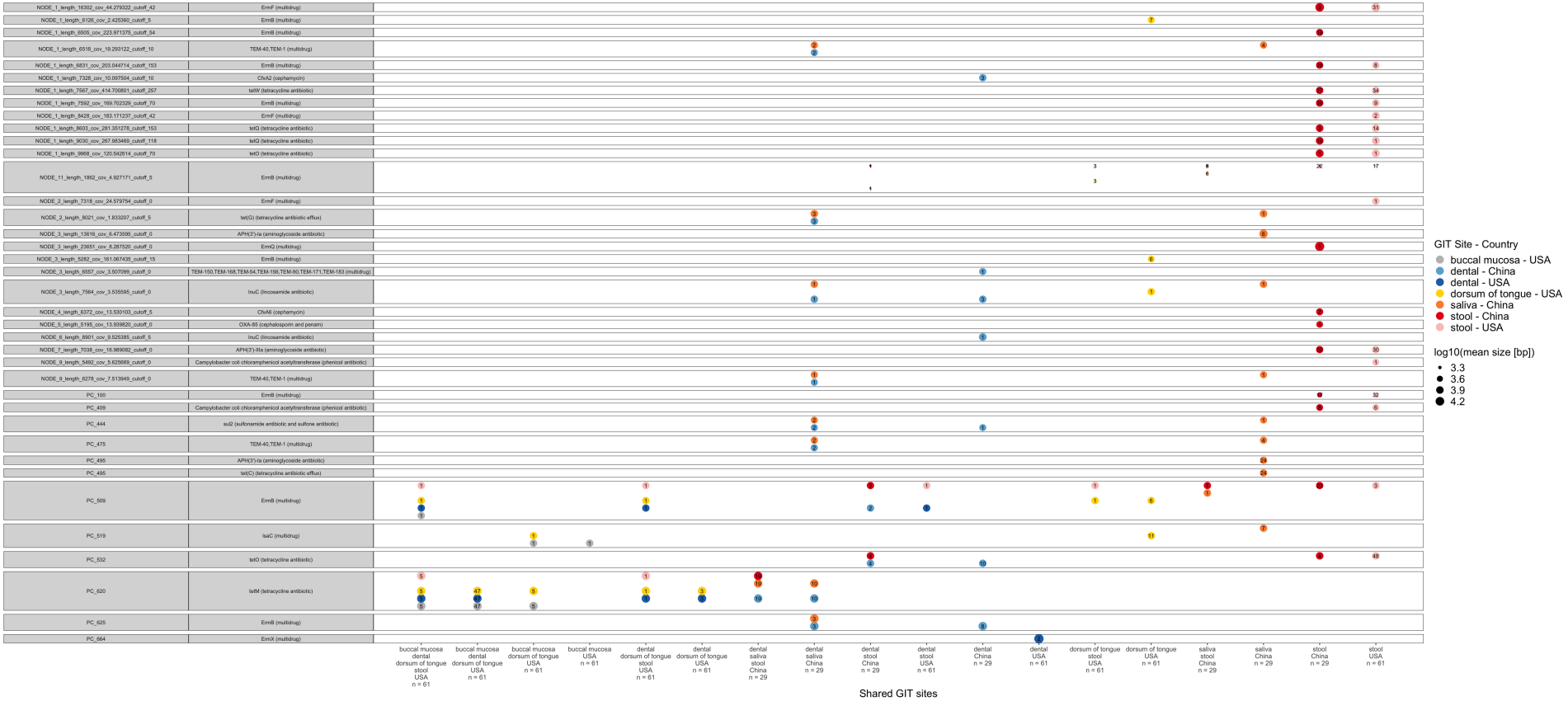

Figure 4.6. ARG-carrying plasmids of GIT sites from China and the USA.

Each row represents an ARG carried by a plasmid cluster or singleton. The first grey column indicating the ARG and ARG class in brackets. Multiple ARG assignments suggest an alternative ARG that is homologous to more than one ARG. The second grey column showing the plasmid cluster or singleton that carries it. Each circle represents a plasmid cluster or singleton with the number of individuals it is found in. The x-axis shows the number of individuals that had GIT sites sampled. Each circle is labelled by body site and country and scaled by $\log_{10}$ average size of plasmid in bp. Some plasmids that were aligned to known plasmid genomes from PlasmidFinder are labelled below the circles. **(Larger figure is available to view here:** https://tinyurl.com/y2kbvyx7**).**

Plasmid clusters *ErmB*-carrying NODE_11_length_1862_cov4.927171_cutoff_5, *ErmB*-carrying PC_509, *tetM*-carrying PC_620 and *tetO*-carrying PC_532 are the only resistance plasmids that are shared between oral sites and the gut. Plasmid cluster *tetM*-carrying PC_620 is found in all oral cavities from China and the USA, and is also shared most between buccal mucosa, dental and dorsum of tongue samples from 47 out of 61 (77.0%) USA individuals. It is also highly prevalent between dental, saliva and stool samples from 19 out of 29 (65.5%) individuals from China. The most prevalent gut ARG-carrying plasmid from the USA is PC_532 with *tetO* in 48 out of 61 (78.7%) USA stool samples, which was also identified as a Gram-positive plasmid from the rep_1 family. From China, however, two rep_1 family plasmids carrying *ErmB* are the most prevalent (28/29, 96.6%) in the gut. Although oral sites tend to share plasmids, 24 out of 29 saliva samples from China exclusively contain plasmid cluster PC_495 carrying *APH(3')-Ia* and *tet(C)*, a small (7076 bp) ColE1 plasmid of the *Enterobacteriaceae* family. Although saliva is in contact with dental plaque, species from the *Enterobacteriaceae* family are only located in saliva (and stool samples) from individuals carrying the ColE1 plasmid **(Fig. 4.7)**. In particular, saliva contains a higher relative abundance of unclassified species of *Escherichia* than stool, making them possible hosts for sequestering ColE1.

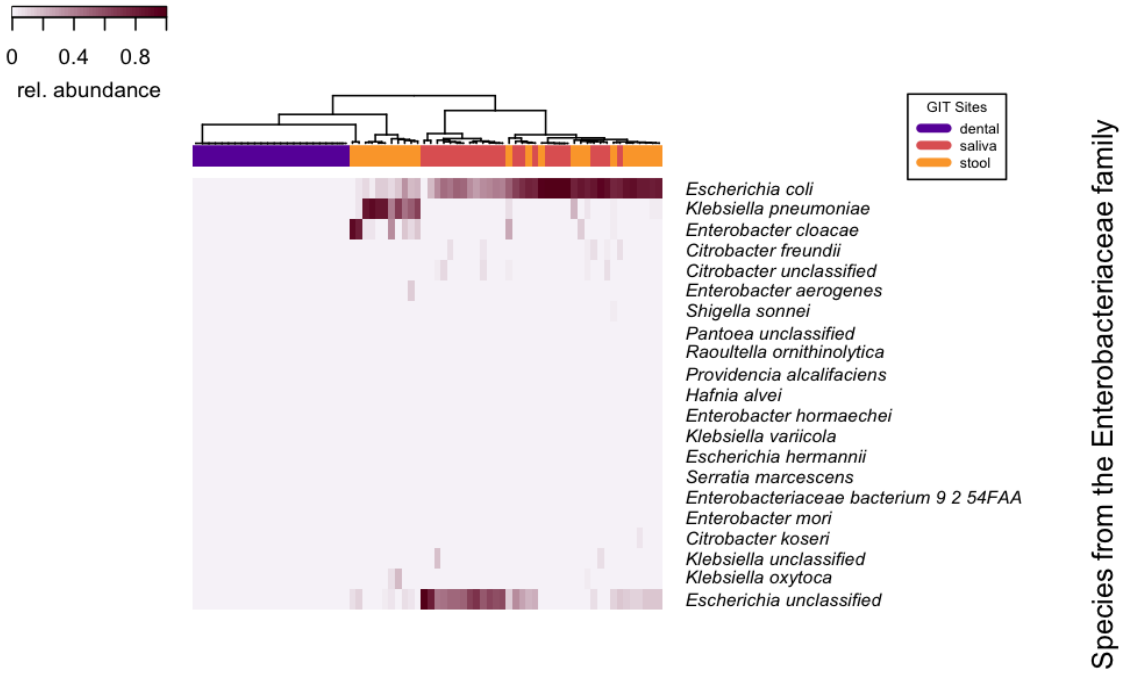

Figure 4.7. Relative abundance of species of the *Enterobacteriaceae* family.
Relative abundance calculated as number of reads mapped to species divided by the total mapped reads to the *Enterobacteriaceae* family derived from Metaphlan2.

To show whether ARG-carrying plasmids are stable, longitudinal USA samples were profiled for plasmid singletons/clusters carrying ARGs. The proportion of plasmid singletons/clusters remaining after one timepoint decreases across GIT sites, particularly in dental plaque, dorsum of the tongue and stool samples **(Fig. 4.8a)**. Fewer resistance plasmids are found in buccal mucosa than in dental plaque or the dorsum of the tongue, but most of these plasmids remain stable for up to three timepoints. They may represent abundant plasmids that are highly detectable in whole metagenomic samples, including mucosa samples with less microbial DNA than other oral samples. For instance, the plasmid cluster PC_620 carrying *tetM* is highly prevalent in three or more timepoint in all oral sites, including buccal mucosa, but not stool samples **(Fig. 4.8b)**. Instead,

plasmids that carry *APH(3')-IIIa*, *ErmB*, *ErmF*, *tetO* or *tetW* are stable in the gut rather than in the oral cavity.

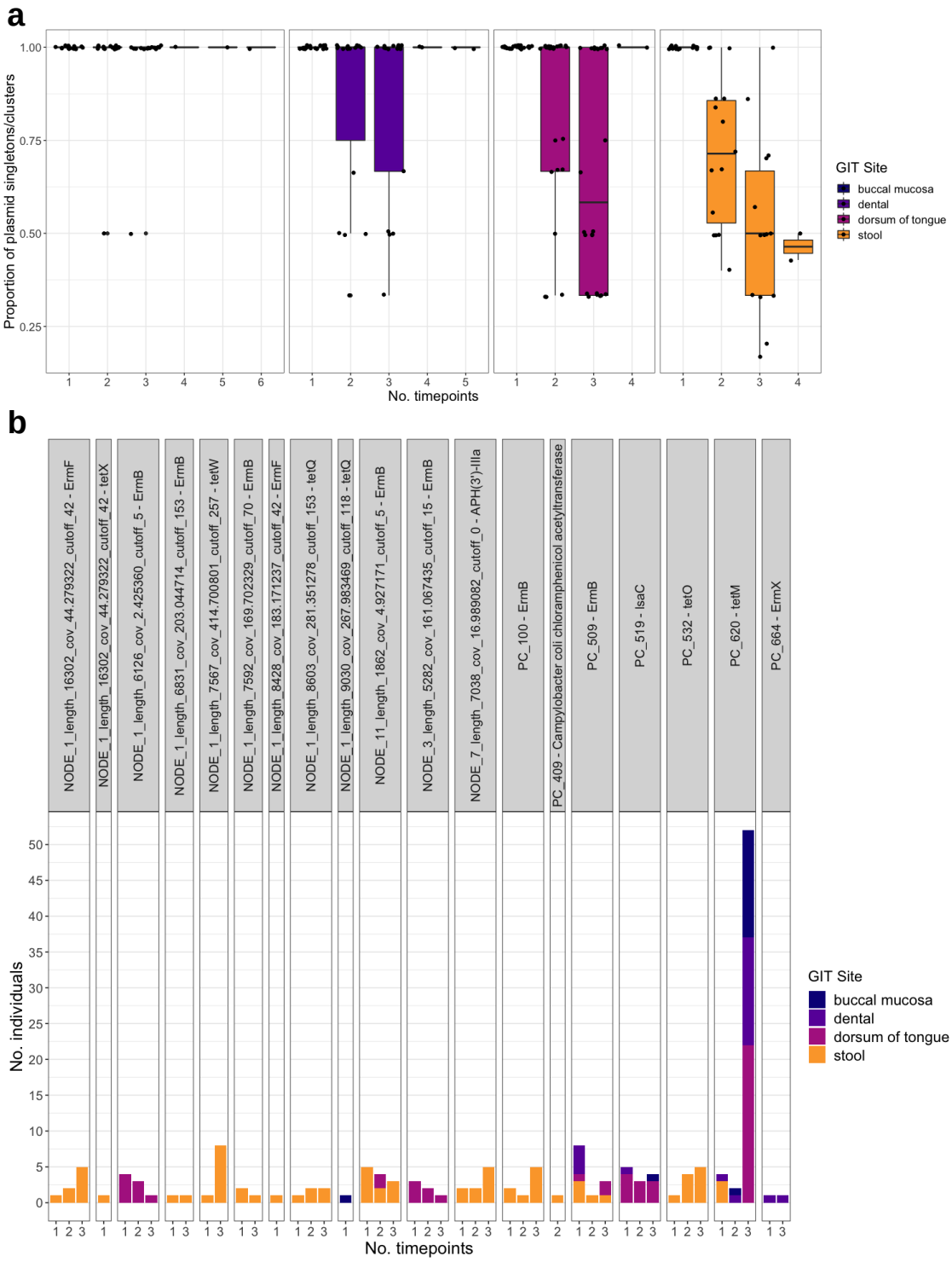

Figure 4.8. Resistance plasmid stability in longitudinal USA GIT sites.

a) Proportion of plasmid singletons/clusters in one to six timepoints and **b)** Number of individuals carrying a resistance plasmid in one to three timepoints for USA individuals with at least three sampling timepoints (buccal mucosa: n = 17, dental plaque: n = 20, dorsum of the tongue: n = 24, stool: n = 16).

4.4 Discussion

Composition of plasmid profiles across the GIT reveals plasmid profiles are distinct between GIT sites, particular between the oral cavity and the gut in both China and the USA. Although most plasmids have low prevalence across GIT sites, there are some highly prevalent plasmids associated with GIT site and country. These plasmid profiles correlate with microbial composition, suggesting that the presence of particular plasmids is dependent on the existence of key species groups of bacteria or archaea that may act as carriers of specific plasmids. The surface of the tongue has the greatest plasmid richness compared to other GIT sites and stool samples from the USA, whereas stool has a greater plasmid richness than dental and saliva in China. Generally, plasmids that persist over several months are more unique to individual than transient plasmids across all GIT sites in individuals in the USA. These differences may be overestimations as it is likely some persistent plasmids in some samples may be labelled as transient in other samples given plasmids may still be undetectable from metagenomic data[217]. Nevertheless, persistent plasmids are more likely than transient plasmids to be shared in at least two individuals for stool samples. In contrast, transient plasmids are more likely than persistent plasmids to be shared in at least two individuals on the dorsum of the tongue. It may be that bacterial communities hosting plasmids on the surface of the tongue fluctuate more than those in the gut.

Oral sites from China have a greater number of plasmids that carry ARGs (resistance plasmids) than from the USA. As the China cohort was taken as controls from a rheumatoid arthritis study and the USA cohort consists of mainly young adults, it is quite possible that participants from China may have acquired a greater diversity of oral

resistance plasmids in their lifetime through being exposed to more anthropogenic antimicrobials that selected for resistance plasmids.

A type of plasmid carrying *tetM* is the most prevalent and highly persistent in USA oral samples. *tetM* has been shown to be more prevalent across oral than stool samples (Chapter 2), but there have been contradictory results[307,313]. In addition, *APH(3')-Ia*, previously shown to be highly prevalent in saliva samples from China in particular (Chapter 2)[307], and *tet(C)* are also linked to a highly prevalent plasmid cluster in matching saliva samples. This plasmid cluster was identified as a ColE1 plasmid of *Enterobacteriaceae* family bacteria: a highly mobile vector for ARGs in animal, human and environmental microbial samples[402]. Aminoglycoside resistant ColE1 plasmids have been previously isolated and characterised from *Salmonella* strains clinically resistant to kanamycin (an aminoglycoside) from veterinary diagnostic labs in the USA[403], and *APH(3')-IIa*-carrying ColE1 plasmids were identified using PCR in human faecal samples[402]. However, this is the first time ColE1 plasmids containing *APH(3')-Ia* have been reported in human saliva. Likewise, *tet(C)* has not been identified from isolated ColE1 plasmids, but has been found previously with computational sequence alignment of *Salmonella enterica* plasmids from the NCBI database[404].

To summarise, most variation in plasmid profiles is found between the oral cavity and the gut in both China and the USA. As well as containing the highest diversity of bacteriophage genotypes (Chapter 3), the surface of the tongue contains a higher plasmid richness than other GIT sites from the USA that may be more mobile than those in other oral sites. Resistance plasmids tend to be specific to either the oral cavity or the

gut in both cohorts, and more unique oral resistance plasmids were from China than the USA. Cohorts that carry highly prevalent ARGs in metagenomes (Chapter 2)[307], are also highly prevalent on resistance plasmids, particularly *tetM* and *APH(3')-Ia*. However, regardless of whether these ARGs are prevalent across a population of individuals, highly abundant ARGs within an individual may be associated with a range of different plasmids that can propagate ARGs more readily in different microbes within a community. For instance, ARGs carried by multiple different plasmids tended to be highly abundant (Chapter 2)[307], especially for *ErmB* and for certain TEM β-lactamases.

Chapter 5: *De novo* Identification of Transposable Elements and their Association with the Resistome

5 *De novo* Identification of Transposable Elements and their Association with the Resistome

5.1 Introduction

Transposable elements are the most abundant type of MGE that transfer ARGs between microbial genomes by HGT. As described in Section 1.4.3.1, an insertion sequence is about 700-2,500 bp long and contains short inverted terminal repeat (ITR) sequences of 10-50 bp at both ends that are reverse complements of each other **(Fig. 1.3)**[117]. A typical composite transposon is made up of two insertion sequences that can flank passenger genes, such as ARGs[405]. Unit transposons are a similar type of transposable element to insertion sequences containing a pair of ITRs but can also carry ARGs. For simplicity, the abbreviation "IS" will be used hereafter to mean insertion sequence or unit transposon that contains ITRs.

As transposable elements are the most ubiquitous and abundant MGE, it is challenging to catalogue all of them. Identifying transposable elements *de novo* from whole metagenomes without reliance on reference databases is a way to circumvent this issue. One study in 2013 applied existing tools to identify ISs from metagenomic assemblies[270]. The authors used the EMBOSS software suite[272] (tools for a variety of sequence analysis) to locate gaps between palindromic motifs in these assemblies to represent insertion sequences flanked by ITRs. They also applied hmmsearch[209] to

amino acid sequences that were translated from these assemblies against a database of HMMs of transposase proteins to identify ISs containing transposase genes. However, current assembly algorithms struggle to resolve repeated elements (e.g. only including one of the repeats or omitted them all together). ITRs are no exception, meaning transposable elements that contain ITRs can be misassembled or incomplete[406]. This study is the only attempt of identifying transposable elements from whole metagenomic data published to date. Therefore, very little is known about the profile of transposable elements in microbiomes and their role in ARG transfer in complex microbial communities.

I developed a software tool, called PaliDIS (**Pali**ndromic **D**etection of **I**nsertion **S**equences) that discovers ISs *de novo* from short-read, paired-end metagenomic data by identifying ITRs from reads and transposase genes from contigs. ITRs were identified from metagenomic reads using another tool I developed called pal-MEM (**pal**indromic **M**aximal **E**xact **M**atching) that was integrated as part of PaliDIS. ITRs are identified from reads using an efficient maximal exact matching algorithm, a fast algorithm optimised for large genomic datasets, such as metagenomes of complex communities[407,408]. A maximal exact match (MEM) between two sequences represents the maximum length of residues (nucleotides or amino acids) that match exactly and cannot be extended in either direction without allowing for a mismatch. After implementing PaliDIS, a catalogue of non-redundant inverted repeat clusters was created to investigate the prevalence of ITRs across metagenomic samples. ARGs were also queried to evaluate their association with ITRs.

5.2 Developing the pal-MEM software

pal-MEM was developed in C++ to identify inverted repeats in metagenomic reads (**https://github.com/blue-moon22/pal-mem**). It uses a similar MEM searching algorithm to an existing tool, called E-MEM[409], but was modified for specifically finding MEMs between reverse complement sequences of inverted repeats. pal-MEM was optimised for better computational efficiency, allowing it to be applied to large metagenomic data files.

5.2.1 The E-MEM algorithm

The E-MEM (efficient computation of MEMs) tool aims to identify MEMs from large genomes with higher computational efficiency than other MEM algorithms[409]. The E-MEM algorithm relies on making smaller exact matches of *k*-mers between two sequences. Then, these matching *k*-mers are extended in both directions to make larger sequence matches until mismatches between two residues on both sequences on both ends disrupt the extension, making a MEM. If the MEM has a length equal to or greater than a length *L*, the start and end coordinates of the MEM in both sequences are recorded.

Firstly, the algorithm must search for *k*-mer matches between two files, specified by the user, containing sequences. One file acts as a reference, while the other is a query. *k*-mers, with defined length *k*, and their positions from the reference are recorded in a hash table. A hash table is a data structure that is indexed in a way that entries can be retrieved quickly, and is ideal for storing and retrieving large amounts of information,

such as a dictionary of all *k*-mers in a large sequence file. Other MEM algorithms developed before E-MEM relied on storing sequence information in suffix arrays or compressed text indexes, which require excessive amounts of memory if applied to large genomes like whole metagenomic data[409–411]. The nucleotides of *k*-mers within the hash table are encoded as unique combinations of two bits (*0* and *1*), where A is *00*, C is *01*, G is *10* and T is *11*, which reduces memory requirements. In addition, not all *k*-mers and their positions have to be stored in the hash table for all MEMs of length *L* or greater to be identified, reducing the demand on memory further. The *k*-mer only has to have a position in the reference that is a multiple of $(L - k) + 1$ (where *k* is the length of the *k*-mer), i.e.

(eq. 5.1) $$b_r \leq j((L - k) + 1) \leq e_r - k + 1$$

where b_r and e_r are the start and end positions of a MEM and $j \geq 1$ **(Fig. 5.1)**.

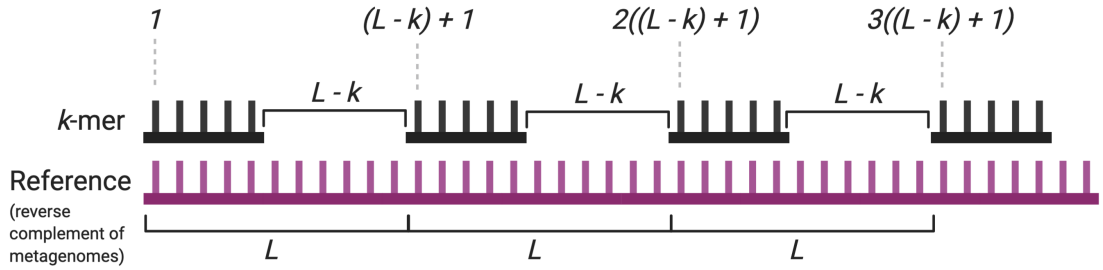

Figure 5.1. *k*-mer positions that are saved into the hash table at the beginning of a reference sequence.

The positions are multiples of $(L - k) + 1$, where *L* is the minimum length of the MEM and *k* is the length of the *k*-mer.

As with the *k*-mers, each nucleotide of the query sequences is encoded as two bits. For query files containing disjointed sequences (such as reads from short-read sequencing), the reads are stored as a continuous sequence in an array of unsigned[xix] 64-bit integers

xix Unsigned integers are integers that can only hold non-negative whole numbers in memory.

representing blocks of 32 nucleotides (as each nucleotide is represented by two bits). A mock sequence of random 20 bits is generated between reads to represent a continuous sequence. The start and end positions of these mock bits are stored in another data structure: *blockofNs*. The start and end positions for each sequence, and their IDs, are also stored in another data structure: *seqData*. Each *k*-mer from the query is looked up against the reference hash table to retrieve a matching *k*-mer. The first *k*-mer window starts from the beginning of the query and continues to shift every two bits, apart from skipping the positions within the *blockofNs*.

Once a *k*-mer match is found between the query and reference sequences, the program attempts to extend the *k*-mers to make larger matches. The length of residues is extended by the same amount for both sequences at the left end of the matching *k*-mers. If an exact match is found between the extensions, it is extended further. If there is no match, the extension has to be updated and the process repeated until a match is found. Once an exact match is found and cannot be extended further, these steps repeat for the right end of the matching *k*-mers. The fastest way for the algorithm to perform this is to use an interval halving approach. The sequence is extended to the left end position of the shortest of the two sequences. If there is no match, the extension is halved until a match is made. Once there is a match, the extension is elongated by one residue at a time until no more exact matches can be made. This is repeated on the right side. Then if the MEM is of length *L* or greater, the coordinates of the MEM between both sequences are recorded. This continues for every *k*-mer query and match between two sequences that are made.

As the MEMs are identified, the positions of the MEMs are written in temporary files. Once all MEMs are identified, the data from these temporary files are then reloaded into the program to be sorted numerically by position. Finally, a file is written which provides the MEM positions between pairs of sequences and their IDs taken from *seqData*. The positions are relative to the individual sequences, rather than to the continuous sequence encoded in E-MEM, and are calculated from the start and end positions of the sequences retrieved from *seqData*.

5.2.2 The pal-MEM algorithm

pal-MEM is a tool based on E-MEM with significant modifications to optimise for processing large files of short-read metagenomic sequences with the aim of identifying inverted repeats. The following sections describe these modifications in more detail.

5.2.2.1 Finding reverse complements only

E-MEM defines MEMs as including reverse complements as well as direct matches. By default, E-MEM searches for direct matches between query and reference files, but it also includes an option for including reverse complements. To identify reverse complement sequences as MEMs, E-MEM transforms the sequences from the reference file into reverse complements of each other, meaning any reverse complements in the query file become direct matches of the transformed reference. As pal-MEM's aim is to identify inverted repeats that are reverse complement sequences, this transformation is set as default.

5.2.2.2 Identifying MEMs from one metagenomic library

E-MEM takes two sequence files (a reference and a query) as an input. These can be different files, but they can also be the same, meaning it is possible to find MEMs within the same library of sequences. Since pal-MEM is only required to find MEMs within one library, it was modified to accept a single library as an input. As short-read sequence libraries can be sequenced by single-read sequencing to generate one file or by paired-end sequencing to generate two files, pal-MEM can either take one single-read file or two paired-end files as inputs.

5.2.2.3 Faster search for inverted repeats

As described, E-MEM queries a *k*-mer match for every two-bit window. Even for an efficient algorithm, this is very time consuming for large files, such as whole short-read metagenomic files with potentially billions of nucleotides. As pal-MEM is designed for identifying inverted repeats from short-read sequences, it is expected a read (~100 bp long) that covers a potential ITR of an IS would only contain one MEM. To speed up the search, instead of querying *k*-mer matches at every two bits, pal-MEM stops querying in the same read if a MEM was already found. So if a MEM was found in a read, the next *k*-mer query would continue at the beginning of the next read.

5.2.2.4 Ignoring technical reverse complements

Read libraries consist of technical reverse complements as well as biological reverse complements, like ITRs. Technical reverse complement reads occur when two

overlapping fragments from both strands (a strand and its reverse complement strand) of a double stranded DNA molecule are sequenced. As a fragment is likely to have originated from a double strand of DNA, technical reverse complements pervade sequence files. In fact, most reads in a sequence file are technical reverse complements. In E-MEM, MEMs of technical reverse complement reads are found at the prefix of one read and the suffix of the other **(Fig. 5.2)**.

Figure 5.2. A MEM between two technical reverse complement reads.
A MEM of a reverse complement, labelled in red, located on the prefix of the purple read and suffix of the green read that are technical reverse complements

To ensure technical reverse complements are not falsely identified as ITRs, an additional algorithm was included in pal-MEM to ignore technical reverse complements. Firstly, MEMs are excluded if their start or end positions are within two nucleotides of either end of one of the reads. As sequencing is not completely accurate, sometimes MEMs are prevented from extending further by mismatches caused by sequencing errors. Therefore, sometimes MEMs of technical reverse complements with errors may not extend fully to the end of the read, meaning the exclusion criteria would miss them. As a result, the MEM would be falsely considered as an inverted repeat. To mitigate this, a buffer of length B is applied to both ends of the reads in order to capture k-mer matches near the end of the reads. Since reference k-mers have positions $j((L - k) + 1)$ **(eq. 5.1)**, it would mean $B \geq L - 1$ to allow at least one k-mer match within the

buffer **(Fig. 5.3)**. Otherwise if *B* is not long enough, there is no guarantee a *k*-mer match will be completely within the buffer region. Sometimes there can be sequence errors within the *k*-mers themselves, meaning a *k*-mer match may not be found within the buffer region. Therefore, this method cannot completely avoid MEMs between technical reverse complements. An additional sequence clustering step before pal-MEM is included in PalidIS to mitigate this further, which is described below in Section 5.3.2.

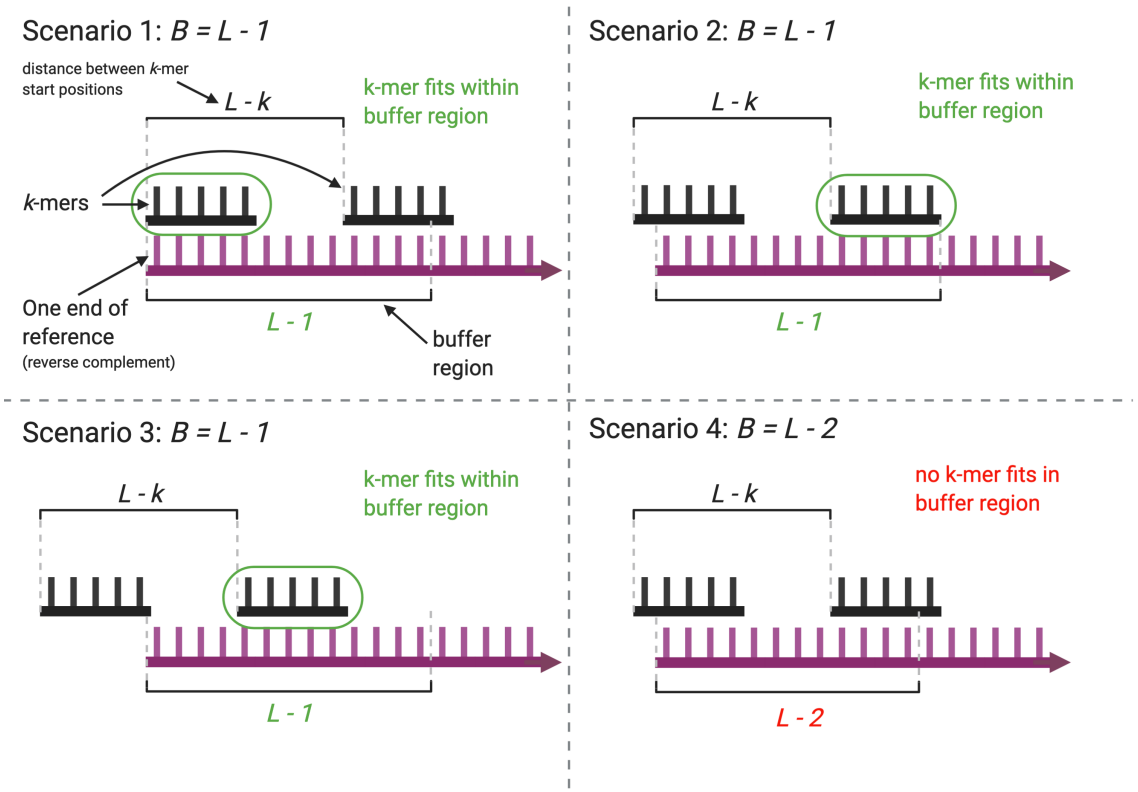

Figure 5.3. Buffer to capture *k*-mer matches at the end of the reads that indicate technical reverse complements.

Scenarios 1 to 3 show potential k-mer matches being captured by the buffer when B is $L - 1$, whereas Scenario 4 shows a *k*-mer match being missed if B is less than $L - 1$.

5.2.2.5 Changing how the MEMs are recorded

In E-MEM, the positions of the MEMs are written in temporary files. Once all MEMs are identified, the data from these temporary files are then reloaded into the program to be sorted numerically by position. As E-MEM queries *k*-mers every two bits, the same MEM is often discovered multiple times. Therefore, the positions of MEMs have to be deduplicated before being written to a final output. For a short-read metagenomic file, MEM positions in temporary files generated by E-MEM can be hundreds of Gigabytes in size, which could be computationally intensive to sort and deduplicate with limited memory resources. As only one copy of MEM positions is recorded in pal-MEM (because only up to one MEM can be found in a read), pal-MEM does not need to rely on sorting and deduplicating MEM positions, nor does it require any temporary files. When a MEM is identified in pal-MEM, the read is flagged as containing a MEM and the MEM positions are stored in *seqData*.

5.2.2.6 Changing the output

E-MEM outputs one file containing the MEM positions between pairs of sequences and their sequence IDs. Similarly, pal-MEM writes the IDs of the sequence pairs with the MEM separated by tabs, where the coordinates of the MEM are added to the ID. For example, the ID *Seq1_ERR589353.7724658/1_Lcoord_23_Rcoord_56* represents the sequence with the ID *ERR589353.7724658/1*, and the numbers after *Lcoord* and *Rcoord* specify the start and end positions of the MEM. Each new line is a different pair of sequences. In addition, pal-MEM also writes the decoded sequences of the reads

containing MEMs in one text file, and the decoded sequences of reads not containing MEMs in another.

5.3 Developing PaliDIS

I created a software tool, called PaliDIS (**Pali**ndromic **D**etection of **I**nsertion **S**equences), to identify ISs from short-read, paired-end metagenomic reads. It consists of a pipeline of five steps:

1) Assembling metagenomic reads to contigs.
2) Clustering metagenomic reads to representative reads.
3) Using pal-MEM to identify inverted repeats.
4) Identifying contigs associated with potential ITRs using reads with inverted repeats.
5) Identifying ITRs from transposases.

A schematic of the pipeline is shown in **Fig. 5.4**. The pipeline is implemented in Nextflow[412] for other researchers to use as a standalone program. The software can be downloaded here: **https://github.com/blue-moon22/PaliDIS**.

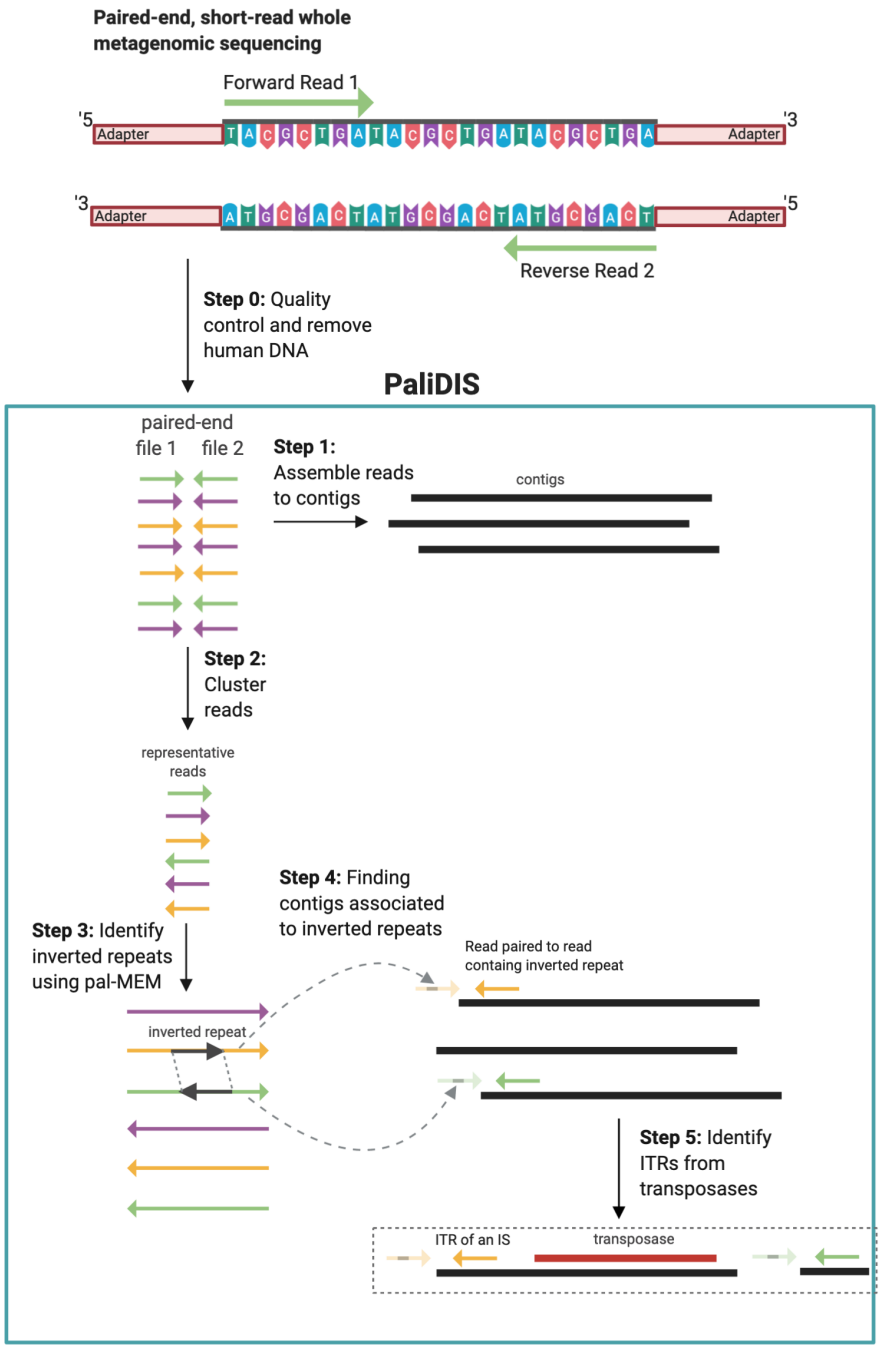

Figure 5.4. Schematic of PaliDIS pipeline for paired-end, short-read whole metagenomes. Before PaliDIS, reads must be quality controlled and contaminant DNA must be removed (Step 0). **Step 1:** Assemble reads to make contigs using metaSPAdes[205]. **Step 2:** Cluster reads by sequence identity to create representative reads using MMSeqs2[413]. **Step 3:** Identify reads carrying inverted repeats using pal-MEM. **Step 4:** Find contigs associated with inverted repeats by mapping reads, that are paired to reads carrying inverted repeats, to contigs using Bowtie2[195]. **Step 5:** Annotate contigs with transposase genes to identify ITRs using HMMER3[209].

5.3.1 Assembling metagenomic reads to contigs

After reads have been quality controlled and filtered (as described in Chapter 2 Sections 2.2.3.1 and 2.2.3.2), they are assembled into contigs using metaSPAdes v3.14.0[205], with parameter *-meta* for metagenomic reads.

5.3.2 Clustering metagenomic reads to representative reads

Reads are clustered by sequence similarity into reference sequences with MMSeqs2 v21d798f09003d0375f0007462e7c4faa1d5eaff7[413]. This is done to cluster some technical reverse complement reads that could be falsely identified as inverted repeats in pal-MEM. The command and parameters *mmseqs linclust* using parameters *--cov-mode 2 -c 0.5* are used to run the clustering. This means the reads are only clustered if their overlap is greater than 50% for at least one of the pairs and the overlap has a sequence identity greater than 90% (default value in *mmseqs linclust*). This coverage threshold is large enough to cluster technical reverse complements, but small enough to avoid removing inverted repeats of up to 50% of the read length. Most short-read lengths are at least 100 bp long, meaning this avoids removing larger ITRs of up to 50 bp long. Reference sequences for each cluster are then generated using *mmseqs createsubdb* and *mmseqs convert2fasta* programmes.

5.3.3 Identifying inverted repeats using pal-MEM

pal-MEM can be run on the command line. It includes mandatory parameters *-f1* and *-f2* that specify the first and second paired-end files or *-fu* that specifies a single-read file.

These read files are required to be in a FASTA sequence file format. Another mandatory parameter is *-o* that specifies the prefix name of the output files. Optional parameters are *-l*, *-k*, *-d* and *-t*. *-l* specify the minimum length of the inverted repeat (default is 24). *-k* is the length of the *k*-mer (default is 15). *-d* is the number of chunks from the sequences file(s) to encode in the reference sequentially, an optional parameter that was repurposed from E-MEM (default is 1). Firstly, the first chunk is encoded in the reference. Once querying has finished on the reference, the second chunk replaces the first chunk that is then encoded as the reference, and so on. It allows hash tables to be built with smaller memory requirements, which is especially useful when analysing large files in limited memory environments. *-t* specifies the number of processing threads for computing in parallel (default is 1).

To test PaliDIS, the pal-MEM parameters *-l 24 -k 15* were used. The minimum length of an ITR has been recorded to be 10 bp[117], but a trade-off minimum length of 24 is specified. The smaller *l* is, the more *k*-mers have to be stored in the reference **(eq. 5.1)**. Instead of decreasing *l*, *k* could be decreased to reduce the number of *k*-mers stored. However, the smaller the *k*-mer, the more matches are made. Smaller *k*-mers are less unique and there is a greater chance of matches and extensions, increasing the computational time further. Therefore, the priority is to have a *k* value that is large enough to generate more unique *k*-mers, but small enough to reduce the number of *k*-mers stored in the reference, with *l* being small enough to find as many potential ITRs as possible. pal-MEM was run on a cluster, therefore *-t 8* was used.

5.3.4 Identifying contigs associated with inverted repeats

Assembly tools struggle to resolve repeated regions, like ITRs, meaning that both ITRs in an IS/transposon, or even one of them, can be missing or incomplete in a contig. Thus it may be difficult to locate inverted repeats within contigs themselves. This was a problem for the method employed by Kamoun et al., where two palindromic motifs representing both ITRs had to be detected on a contig for an IS to be found[270]. In PaliDIS, inverted repeats are detected from reads without relying on contigs (as described in the previous section). However, contigs are still needed to find ITRs of ISs. To detect whether a contig may contain a potential IS without locating ITRs within a contig explicitly, the following method was implemented. Reads containing inverted repeats are representative sequences of clusters that also contain the same inverted repeats (Section 5.3.2). All reads from these clusters are retrieved using a custom script called *get_all_seq_from_clusters.py* in PaliDIS. These reads have paired reads from paired-end sequence files that do not contain inverted repeats. These paired reads are retrieved using *get_discordant_reads.py*. These non-inverted repeat reads are then mapped using Bowtie2[195] against the contigs with parameters *--very-sensitve-local -f*. The parameter *--very-sensitive-local* sets Bowtie2 to map reads using local alignment with high accuracy and *-f* specifies the reads are in a FASTA format. Reads that are mapped to contigs indicate these contigs are associated with potential ITRs.

5.3.5 Identifying ITRs from transposases

As well as being ITRs of transposable elements, inverted repeat sequences are found in promoters and operators[xx] on DNA in the control of gene expression[414]. They also influence genomic instability as part of the microbial evolution and diversity[415], causing the formation of DNA hairpins or cruciform structures that disrupt DNA replication[416]. They have even been shown to lead to the formation of large inverted dimers of plasmids[416]. In order to identify contigs containing potential ISs, these ITRs must be distinguished from other inverted repeats. In many ISs, the transposase gene is usually next to one of the two ITRs[117]. Contigs associated with inverted repeats are searched for transposase-coding genes. Prodigal v2.6.3[210] is applied to identify and translate protein-coding genes from the contigs. As transposases have varied sequence diversity, instead of using sequence alignment the protein sequences are searched using hmmsearch[209] against transposase HMMs from the Pfam database[251] (downloaded on 7th January 2020). Multiple hits of different transposases on the same protein-coding region are filtered to select only the transposase with the lowest protein and domain e-values. Non-inverted repeat reads paired to the inverted repeat reads (from Section 5.3.4) are up to ~250 bp apart depending on the insert size[xxi] of a short-read Illumina sequencer. Given the largest IS is ~2,500 bp long[117] and one of the ITRs may be furthest away from a transposase, if either the start or end positions of the transposase gene is within 2,750 nucleotides from the non-inverted repeat mapped read, then the contig is associated with an ITR and contains an IS. This is run using the custom script *get_contigs_with_sites.py* in PaliDIS.

xx An operator is a genetic sequence that transcriptional regulator proteins bind to, regulating an operon's transcription.

xxi The insert size is the length of the DNA region between the ends of the paired reads in paired-end sequencing (excluding adaptor sequences) (See **Fig. 5.4**)

5.4 Methods to test PaliDIS

To test whether PaliDIS can identifying ISs from short-read, paired-end whole metagenomic data, the following methods were conducted. A catalogue of non-redundant candidate inverted repeats was generated to categorise ITRs that are linked to a transposase belonging to the same ITR family. Clustering ITRs into non-redundant ITR sequences allows distinct ISs shared across samples to be identified. The prevalence of these ISs and the differences between IS profiles (β-diversity) across metagenomic samples can then be evaluated. ARGs were also searched against contigs containing ISs to investigate the association of ISs with ARGs that may indicate the presence of ARG-carrying composite and unit transposons. The code of the analysis was run in R v3.6.1 and can be found here: https://github.com/blue-moon22/IS_analysis.

5.4.1 Creating a catalogue of non-redundant inverted repeats

To investigate the prevalence of specific ISs across samples, a catalogue of non-redundant inverted repeat sequences was created from 1,154 metagenomic samples from five cohorts:

1) Human Microbiome Project (referred to as US)[336] containing buccal mucosa (n = 87: 33 with one, 34 with two, 19 with three and 1 with six timepoints); dorsum of tongue (n = 90: 22 with one, 42 with two, 24 with three and 2 with four timepoints); dental plaque (n = 89: 26 with one, 38 with two, 21 with three,

1 with four and 3 with six timepoints); stool (n = 70: 14 with one, 32 with two, 21 with three, 2 with four and 1 with six timepoints).

2) Healthy control samples from a Chinese rheumatoid arthritis study[331] containing dental plaque (n = 26); saliva (n = 27); stool (n = 72).

3) Saliva (n = 136) and stool (n = 136) samples from Fiji[332].

4) Saliva samples (n = 23) from healthy hunter-gatherers and traditional farmers from the Philippines[333].

5) Saliva (n = 21) and stool (n = 21) samples from Western Europe (5 saliva and 5 stool samples from Germany[301,334], and 16 saliva and 16 stool samples from France[301,335]).

All reads from these samples containing inverted repeats were first trimmed to include only the inverted repeat sequence and clustered using CD-HIT-EST[401] with parameters *-G 0 -aL 0.5 -aS 1.0.* This means pairwise sequences were clustered together when the smallest inverted repeat sequence had complete sequence identity against at least 50% of the largest inverted repeat sequence using local alignment. An identity threshold of 50% was used for the largest sequence instead of a complete sequence identity because inverted repeats in pal-MEM may have been truncated due to mismatches. CD-HIT-EST assigned a cluster ID to each cluster of inverted repeats. Contigs associated with ITRs were labelled with the corresponding ITR cluster IDs.

5.4.2 Searching for ARGs

To identify ARGs, the contigs with ISs were aligned against CARD v3.0.0[197] using BLASTn v2.10.0+[207] with an e-value cut-off of 1e-5 and an identity cut-off above 90%.

Multiple hits of different ARGs that overlapped each other by greater than 20% on a contig were filtered to leave the hit with the lowest e-value and highest identity values. If the e-values and identity values are the same, the hit annotations were combined as they could represent ARG variants from the same ARG family.

5.4.3 IS β-diversity

To find differences in β-diversity of IS profiles between groups of individuals, the Jaccard distance of ITR cluster incidence (presence or absence) profiles was computed between individuals and visualised using NMDS.

The Jaccard distance between incidence profiles is defined as:

$$J_d = 1 - \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

where M_{11} is the number of IS clusters in both samples (A and B), M_{01} is the number of IS clusters in sample B but not sample A, and M_{10} is the number of IS clusters in sample A but not sample B.

Silhouette analysis **(eq. 2.1)** of *k*-medoids using the cluster package v2.1.0 was used to select the number of distinct groups with the largest average silhouette width.

5.5 Results

5.5.1 Detecting ITRs and ISs from metagenomic data using PaliDIS

1,154 metagenomic samples were profiled for ISs using PaliDIS based on finding ITRs in paired-end, short metagenomic reads. A total of 49,104,515 inverted repeats of length 24 bp and greater were detected. Inverted repeats were grouped by sequence similarity to generate a catalogue of 6,833,994 non-redundant inverted repeat clusters. 130,409 (1.91%) inverted repeat clusters were associated with transposases representing ITR clusters. 179,672 contigs associated with ITRs also contained transposases, indicating these contigs carry ISs.

5.5.2 Profiles of ISs across GIT sites

Next, the prevalence of unique ISs was investigated across GIT sites. 160,796 unique pairs of ITR clusters and transposases representing ISs were catalogued in all samples. 17.4% (22,172/127,301) pairs are shared between samples (excluding USA longitudinal samples) across cohorts, where most are found in two samples, although several are found in up to 219 samples **(Fig. 5.5a)**. PDDEXK_2 and DDE_Tnp_ISL3 are the most prevalent types of transposase families of these ISs **(Fig. 5.5b)**.

To find differences in β-diversity of plasmid profiles between groups of individuals, the Jaccard distance between IS incidence profiles was computed between individuals and visualised using NMDS. The profiles were clustered into distinct groups, where the

number of groups was selected as having the largest average silhouette width from Silhouette analysis of *k*-medoids. The IS profiles were separated into two groups, where one contained mostly oral samples and other mostly stool samples **(Fig. 5.5c)**. The Kruskal-Wallis Rank Sum test with Bonferroni multiple test correction identified 298 ISs with incidence profiles significantly conforming to this grouping ($p < 0.05$) **(Supplementary Data 5.1: https://tinyurl.com/y2rpl8rq)**. Out of these ISs, 234 are only found in the gut, 34 are exclusively in the oral cavity, and 30 are located in both sites **(Fig. 5.5d)**.

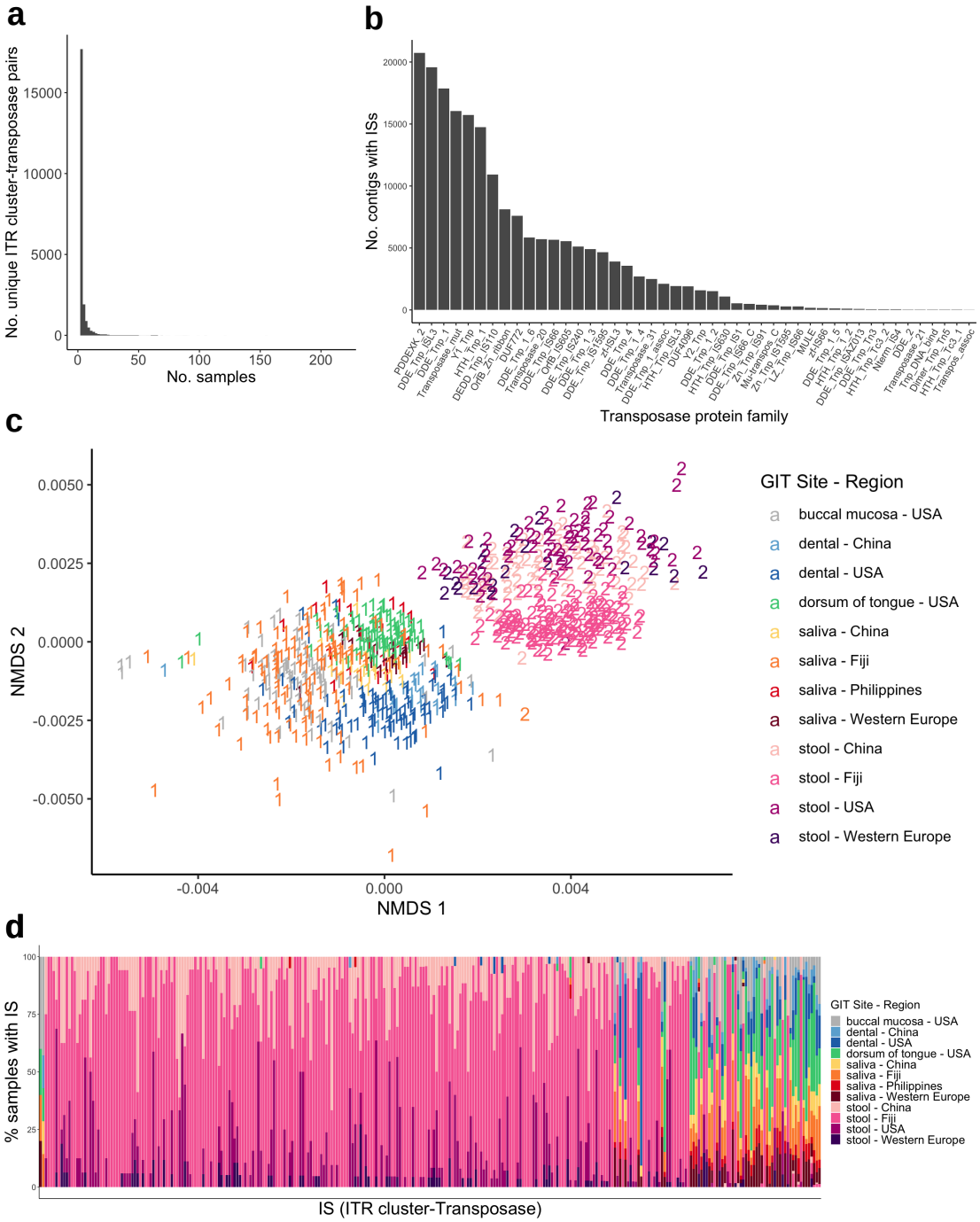

Figure 5.5. IS profiles in metagenomic samples.

a) Number of unique ITR cluster/transposase pairs representing ISs across number of samples. **b)** Number of contigs with ISs carrying different transposase families. **c)** NMDS of Jaccard distance between IS incidence profiles of samples. Ordination coordinates are grouped by *k*-medoids clustering, where number of groups, *k*, has the largest average silhouette width. Missing two outliers. **d)** Percentage of samples with unique ISs that influence grouping in c) (Kruskal-Wallis Rank Test, $p < 0.05$). USA (not longitudinal) buccal mucosa (n = 87), dorsum of tongue (n = 90), dental plaque (n = 89) and stool (n = 70); China dental plaque (n = 26), saliva (n = 27) and stool (n = 72); Fiji saliva (n = 136) and stool (n = 136); Philippines saliva (n = 23); and Western Europe saliva (n = 21) and stool (n = 21).

To verify whether the same ISs can be found across multiple timepoints in the same individual, IS incidences were profiled from longitudinal USA samples. Less than 5% of ISs could be found in two or more sampling timepoints across all GIT sites **(Fig. 5.6)**.

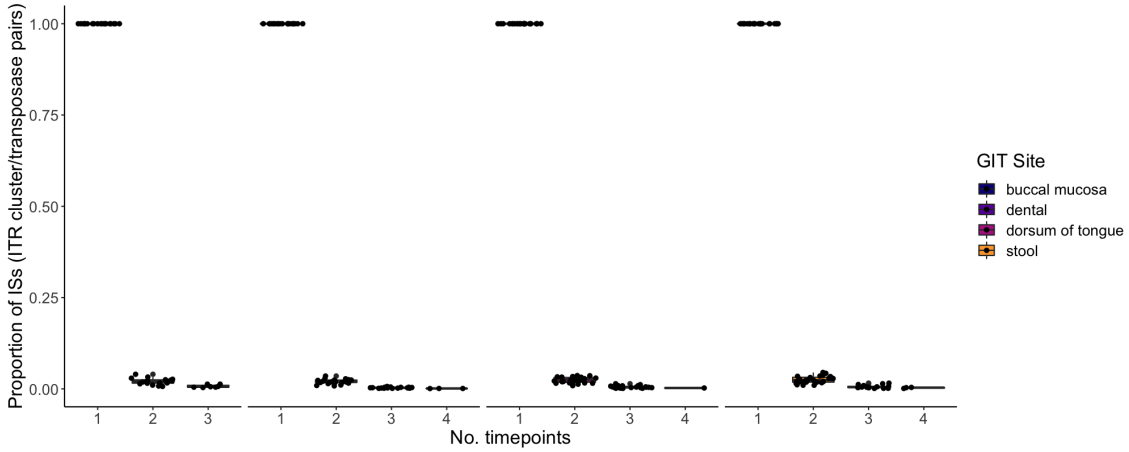

Figure 5.6. IS profiles across longitudinal USA samples.

Proportion of ISs in one to four timepoints for USA individuals with at least three sampling timepoints (buccal mucosa: n = 17, dental plaque: n = 20, dorsum of the tongue: n = 24, stool: n = 16).

5.5.3 ARGs are associated with ITRs

ARGs can be carried within a unit transposon or composite transposon by HGT. As it is challenging for assemblers to resolve more repeat regions, it becomes more difficult to identify complete composite transposons. Therefore, to investigate whether ARGs could be mobilisable by transposition, ARGs are profiled from contigs that are associated with ITRs. These contigs do not have to contain transposases but their associated ITR clusters have been linked to a transposase, perhaps in another contig. 3,995 contigs were found to contain ARGs with ITR clusters across 925 (out of 1,154) metagenomic samples. There were a total of 281 unique pairs of ITR clusters and ARGs. Contigs

associated with ITRs are significantly more enriched for ARGs than those not associated with ITRs (p=0.0332; Wilcoxon Paired Signed-Rank Test) **(Fig. 5.7)**.

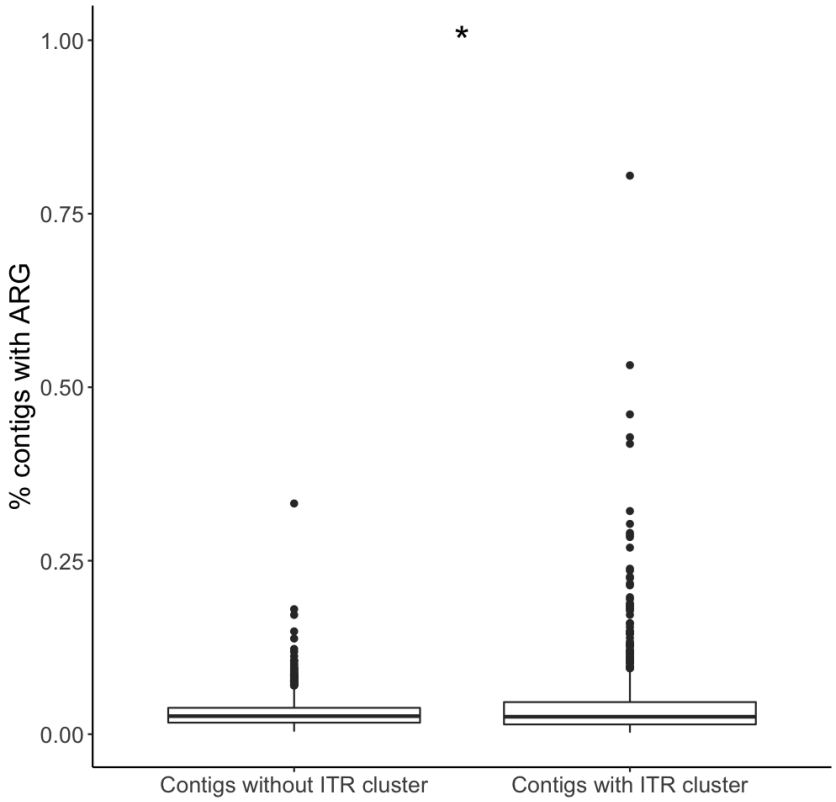

Figure 5.7. Percentage of contigs with or without ITRs carrying ARGs.
From 925 metagenomic samples (p-value < 0.05 as *)

A total of 215 ARG types across 52 ARG classes were associated with insertion sequences **(Fig. 5.8)**. The gut has a higher diversity of resistance linked to ITRs, having 129 ARGs across 19 classes, than the oral cavity with 35 ARGs across 7 classes. 51 ARGs in 26 classes are found in both the gut and the oral cavity. Interestingly, there is variation between these sites within ARG classes. For example, tetracycline resistance (including by efflux pump mechanism) is commonly linked with ISs in both the gut and oral cavity; however, ISs with *tet37, tetA(46), tetA(60), tetB(46), tetB(60)* and *tet(D)* are only located in the oral cavity, whereas ISs with *emrK, emrY* and *tet(40)* reside only in the gut.

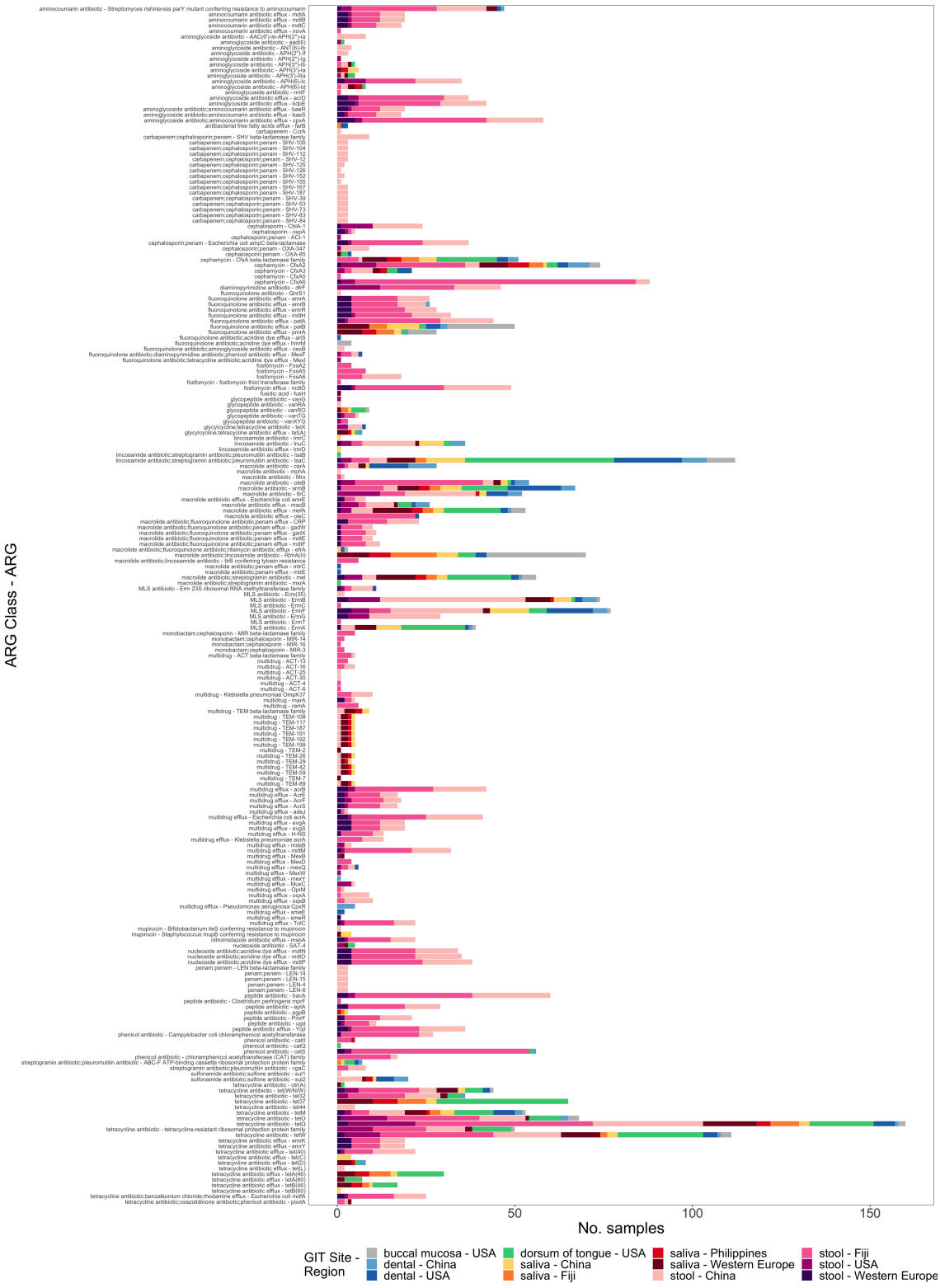

Figure 5.8. Number of samples with ARGs associated with ITRs.

ARGs are shown on the y-axis and are grouped by ARG class. The number of samples are shown in the x-axis with bars coloured by GIT site and geographical region. (**Larger image is available from: https://tinyurl.com/y4o8d9pl).**

To verify whether ARGs and their ITR clusters are persistent in the microbiome, longitudinal samples from the USA were profiled for ITR cluster-ARG pairs across sampling timepoints. 6.72% (73/1,087) of IS-ARG pairs were found in at least two timepoints for all GIT sites samples from the USA cohort, with *tet37* in dorsum of tongue and *tetW* in stool samples in three timepoints **(Fig. 5.9)**.

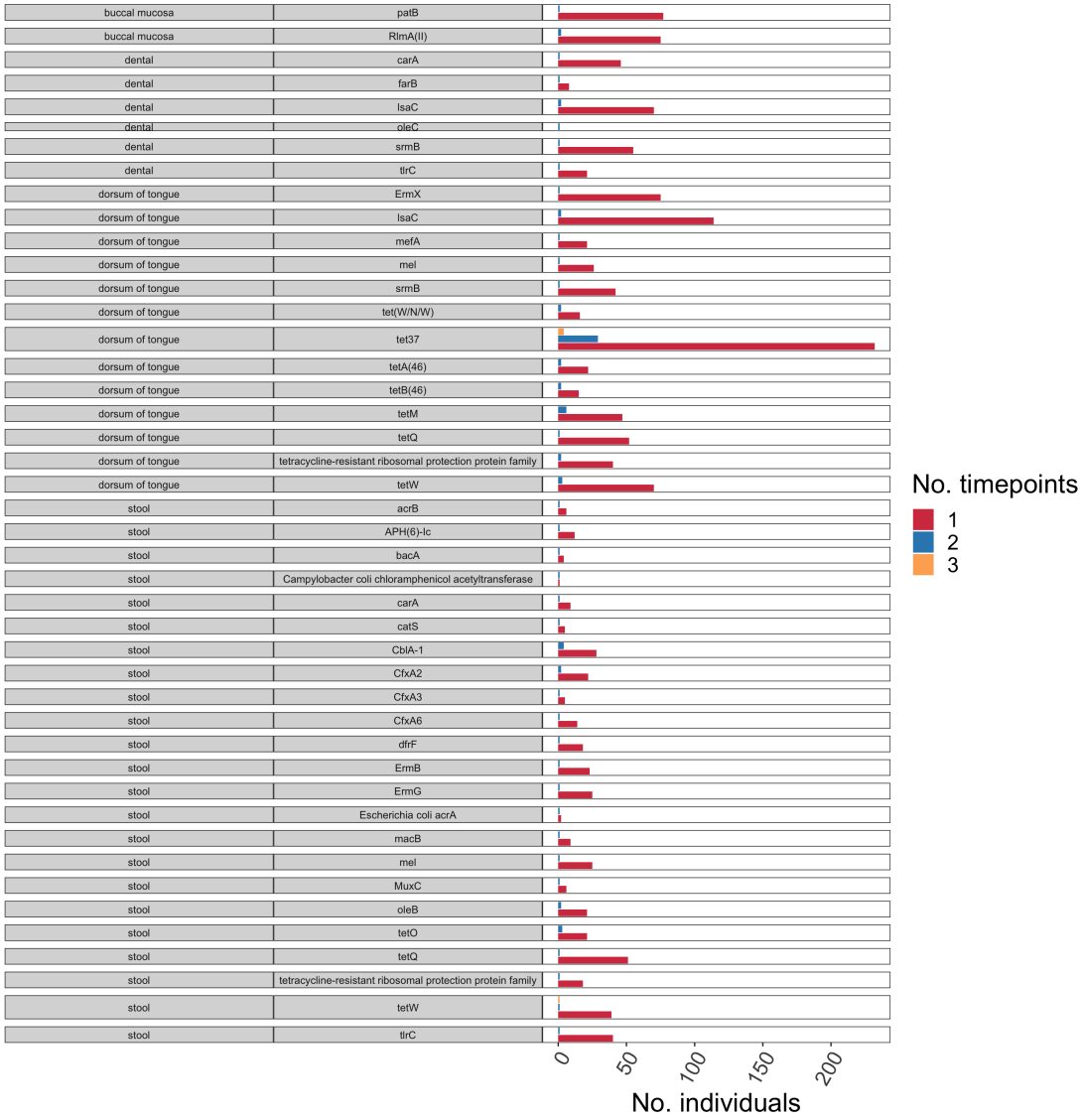

Figure 5.9. Number of individuals with ARGs from the same ITR cluster-ARG pairs found in at least 1,2 or 3 timepoints.

ITR cluster-ARGs are those found in more than one timepoint in longitudinal USA samples.

5.6 Discussion

Here, I built a software package called PaliDIS that identifies ITRs and profiles ISs from short-read, paired-end whole metagenomic data, based on identifying inverted repeats from reads using pal-MEM, another tool I developed. 130,409 unique ITRs were identified from 1,154 metagenomes using PaliDIS. A much earlier study detected a median of 499 ITRs from 63 chromosomes of single isolate genomes from GenBank representing 58 bacterial genomes, ranging from none in *Chlamydia trachomatis* to 66,860 in *Neisseria meningitidis*[415]. Given that 4,644 different culturable and unculturable species have been catalogued from the human gut[417] and at least 700 species reside in the oral cavity[418], with multiple sub-species and strains, there could be millions of ITRs in metagenomes. It is likely many ITRs may have been missed. Only inverted repeats with lengths of 24 bp or greater were identified in pal-MEM, even though ITRs can be as short as 10 bp[117]. This is because it would have been computationally time consuming to search for lower length (as described in 5.3.3). It may be possible to parallelise pal-MEM processing further with a larger number of computer threads to allow for smaller lengths to be found within a reasonable time. Additionally, given transposase genes are genetically diverse, not all ITRs may have been identified from a search of transposase HMMs.

Nevertheless, PaliDIS was able to identify 179,672 contigs containing ISs and find distinct IS profiles between human oral sites and gut samples, driven by a particular set of 298 ISs. Less than 5% of ISs were also shown to persist across multiple timepoints in longitudinal USA samples. This verifies that PaliDIS is able to detect stable ISs that are integrated the microbiome. However, when compared to plasmids and bacteriophages,

this proportion is lower than expected. Since some inverted repeats may have been truncated by disruptions to MEM extensions due to mismatches, it is possible the same inverted repeats may not have had sequences similar enough to be clustered. This would lead to a greater number of ITR clusters and a smaller overlap of ITR clusters between longitudinal samples. In addition, a small minority of ISs lack ITRs, such as IS*91*-like and IS*CR* elements[122]. As all transposable elements contain transposases, an alternative method of identifying unique ISs could be to cluster contigs containing transposases by sequence similarity into a non-redundant catalogue. ITR clusters associated with contigs can be assigned to their contig clusters, whereas those clusters that are not assigned an ITR could be considered potential ISs lacking ITRs. Clusters of contigs could be merged together, along with ITRs, to create more complete ISs sequences. However, this relies on identifying all possible transposases, which is difficult given transposase genes are genetically diverse and can only be identified in genomic data by reference-based approaches. Nevertheless, discoveries of novel ISs could be catalogued alongside the IS reference database, ISFinder[258], to monitor the prevalence of ISs across different ecological samples.

Contigs associated with ITRs are significantly more enriched for ARGs than contigs not associated with ITRs. This suggests that ARGs are more likely to be linked to ISs than without, perhaps as part of composite or unit transposons. ARGs *patB, RlmA(II)* and *tetA(60)* that are highly abundant in the oral cavity compared to the gut (Chapter 2)[307] are mostly associated with ITRs in oral sites. HGT involving ISs as part of composite or unit transposons may be driving the increase in ARG abundances. However, the higher

read depths of these ARGs could have led to better assemblies of contigs and highly detection of ISs associated with these ARGs.

In the future, pal-MEM may be implemented to include shorter ITR sequences and PaliDIS may be developed further to construct and discover new ISs from metagenomic data. Applying PaliDIS to whole metagenomic data is a promising avenue for discovering multiple ISs rapidly, especially from uncultivable strains and those without an expressible phenotype on which functional metagenomics relies.

Once PaliDIS has been developed further, its specificity and sensitivity should be tested on various datasets to inform other users what results they may expect from their data. The sensitivity is the proportion of contigs containing ISs that are correctly identified, and the specificity is the proportion of contigs not containing ISs that are correctly identified. In other words, a high sensitivity would suggest PaliDIS performs well at identifying ISs that exist, and a high specificity would indicate PaliDIS is good at avoiding sequences that are not ISs. In order to make these measurements, PaliDIS must be tested on data where ISs are known. Unfortunately, there are no whole metagenomic datasets of real samples with known ISs. Although it is difficult to recapitulate the complexity of microbial communities in the human GIT, simulated metagenomic datasets representing simple microbial communities can be created by combining genomes from public datasets[419].

Chapter 6: Discussion

6 Discussion

The perennial problem of AMR has made it imperative to understand more about its causes, especially how ARGs are acquired in microorganisms. The work in this study has highlighted associations between ARGs and MGEs, and how they may spread through the microbiome by HGT. In particular, ARGs are linked to plasmids and transposable elements in the human GIT that are shared across populations, unlike phages that seldom encode ARGs. However, to begin exploring the differential influence these types of MGEs have on HGT of ARGs, data of these multiple MGE-ARG associations from this study must be combined. In this final chapter, associations between antibiotic use data and the prevalence of ARG-carrying MGEs will be explored to indicate whether antibiotic use could be driving acquired resistance involving particular types of MGEs. Associations between the MGE incidence and the ARG abundance are then investigated to show which MGEs are most influential in propagating ARGs throughout a microbial community. Additionally, ARG and MGE prevalence data for each ARG are ranked for each antibiotic class, country and GIT site to highlight ARGs that are commonly acquired by HGT and may be important to monitor for AMR surveillance. The code for this analysis is available here: https://github.com/blue-moon22/thesis_summary. Analysis was run using R v3.6.1. I discuss what further data are required to make predictions of clinical outcomes of infections in patients with specific AMR based on HGT-acquired ARGs. Finally, I propose how these predictions can support interventions in preventing the spread of AMR and facilitating alternative therapies to treat patients with AMR.

Previously in Chapter 2, the abundance and diversity of ARGs were profiled from openly available whole metagenomic data of human GIT microbiomes from China, Fiji, the Philippines, the USA and Western Europe (France and Germany). In Chapter 3, a computational pipeline based on existing viral identification software was applied to profile bacteriophages in samples from China, the Philippines and the USA. (The Fiji and Western Europe samples were not available at the time of this study.) Chapter 4 used a similar computational pipeline based on an existing, but relatively new, metagenomic plasmid assembly tool for identifying circular plasmids from China and USA samples. Finally, a new tool called PaliDIS was developed to identify ISs from short-read, paired-end metagenomic data, which was applied to the metagenomic data from all cohorts and linked to ARGs.

6.1 Antibiotic use and prevalence of ARG-associated MGEs

Anthropogenic use of antimicrobials is a major driver of acquired resistance. The use of antimicrobials can lead to a higher incidence of ARGs acquired by HGT through selective pressures. To investigate this, the relationship between the prevalence of ARG-carrying MGEs and antibiotic use in 2015 was evaluated. Prevalence data were used from previous chapters of ARG-carrying phages from China and the USA (Chapter 3), plasmids from China and the USA (Chapter 4) and ISs from China, the Philippines, the USA and Western Europe (Chapter 5). Antibiotic class use data were taken from Chapter 2 (Fig. 2.2c). No antibiotic use data were available for Fiji. The linear relationship between the prevalence of ARG-carrying MGEs and the DDD Per 1,000

individuals in 2015 for the different antibiotic classes was evaluated for each GIT site and MGE type. There is no significant linear correlation between prevalence of ARG-carrying MGEs and DDD for any case (Student's t-test) **(Fig. 6.1)**. Confounders that could not be considered here (due to lack of data) is likely to influence this relationship. For one, it is unclear how the antibiotics were administrated and what body sites were exposed, e.g. orally, topically or intravenously. However, it could be predicted from the antibiotic class used to some extent, e.g. aminoglycosides are administered intravenously or intramuscularly as they cannot be absorbed from the gut. Another reason that no correlation is present could be that these data were taken in just the year 2015 alone, which do not accurately reflect the antibiotic exposures preceding this. Therefore, it could be that ARGs were acquired earlier. However ongoing antibiotic exposure may be required for ARGs to persist in the microbiota given the possible fitness costs of retaining an ARG in a microbial genome. Even once an ARG is retained with or without antibiotic selection pressures, the presence of an ARG does not necessarily equate to it being expressed. Since expression is usually triggered by antibiotic pressures, the expression of acquired ARGs is likely to be driven by antibiotic use levels.

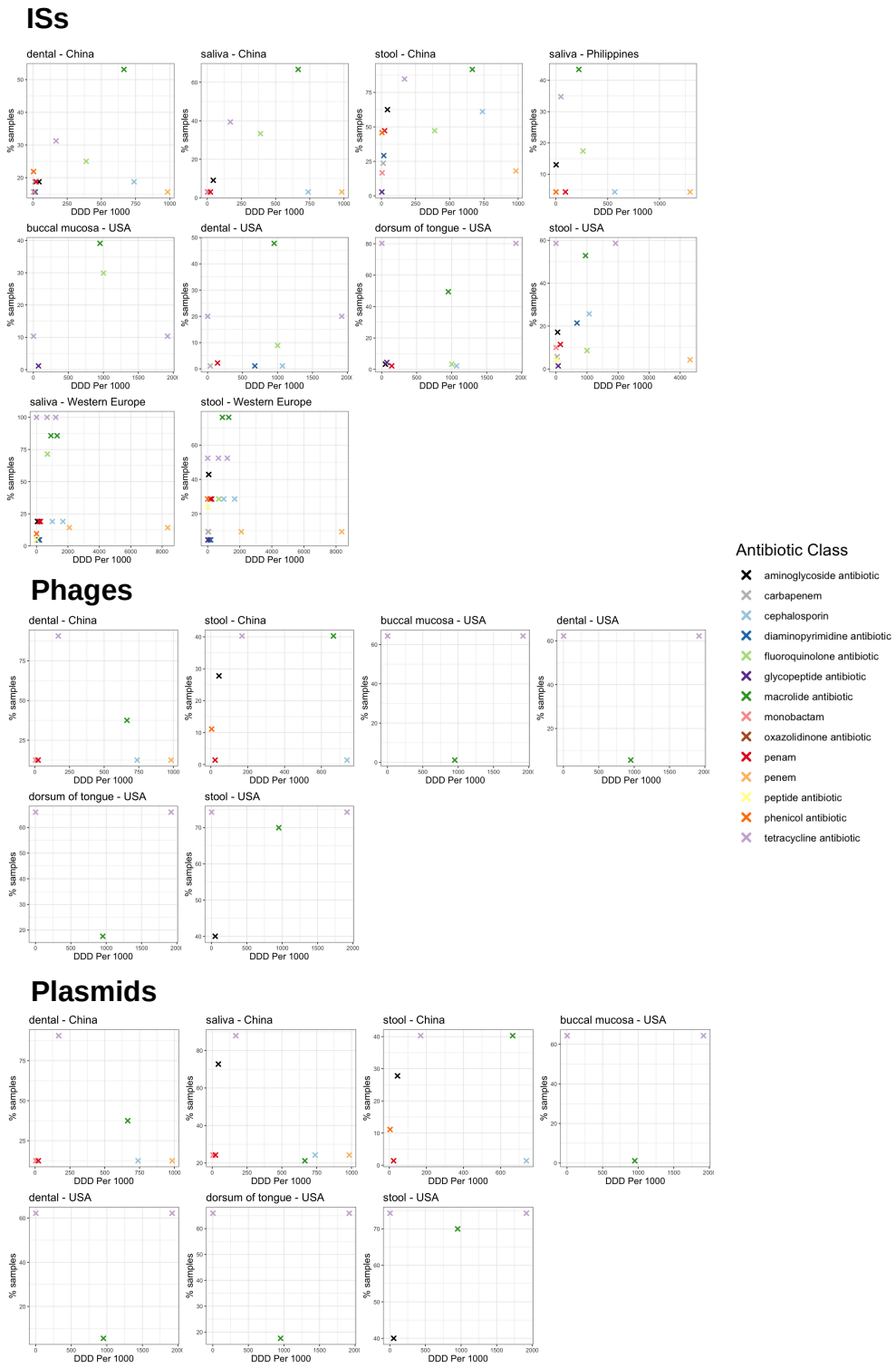

Figure 6.1. Prevalence of ARG-carrying ISs, phages and plasmids against DDD Per 1,000 in 2015 for each antibiotic class.

Samples from China (dental: n = 32, saliva: n = 33, stool: n = 72), Philippines (saliva: n = 23), the USA (buccal mucosa: n = 87, dental: n = 90, dorsum of tongue: n = 91, stool: n = 70), and Western Europe (saliva: n = 21, stool: n = 21). Excluding longitudinal USA samples.

6.2 Comparing prevalence of ARG-associated MGEs

To investigate the relative contributions of different MGE types on the HGT of ARGs, prevalence data for ISs, plasmids and phages were compared in China and the USA samples (as plasmids were not profiled from other cohorts) at each GIT site. There is a higher prevalence of ARG-carrying ISs and plasmids than ARG-carrying phages **(Fig. 6.2)**. Out of these types, ARG-carrying ISs carry the highest number of ARGs targeting the greatest variety of antimicrobial classes. However, for certain ARGs linked to MGEs, ISs are less prevalent than plasmids. For instance, plasmid-associated *tetM* ARGs are common across oral sites in China and the USA, but in fewer samples when IS-associated. It is possible that an ARG, like *tetM*, found in both plasmids and ISs may be within a composite/unit transposon that is part of a plasmid. This could mean *tetM* may be acquired by multiple mechanisms of HGT, allowing it to propagate extensively in microbial communities. For instance, *tetM* could transfer between plasmids and chromosomes by transposable elements, making it possible for *tetM* to spread across a variety of species without being limited to a particular plasmid. This prediction can be tested by searching for ISs within circular plasmid contigs. Similarly, ARGs that are linked with phages and ISs (such as *mdtC* and *mdtA*) may be located in ISs within viral DNA. To further explore the relative influence of MGE types on the ARG incidence across cohorts, the relationship between the prevalence of all ARGs and their associated phages, plasmids and ISs was evaluated for each GIT site from China and the USA using multiple linear regression. ARG-associated ISs are likely to impact the spread of ARGs across GIT sites in China (dental plaque: $p = 0.00798$, saliva: $p = 0.0391$, stool: $p = 7.49 \times 10^{-5}$; Student's t-test on multiple linear regression) and the USA (dorsum of the

tongue: p = 0.0158 and stool: p = 2.17×10^{-5}). Likewise, ARG-carrying plasmids influence ARG incidence in stool samples from the USA (p = 0.0123).

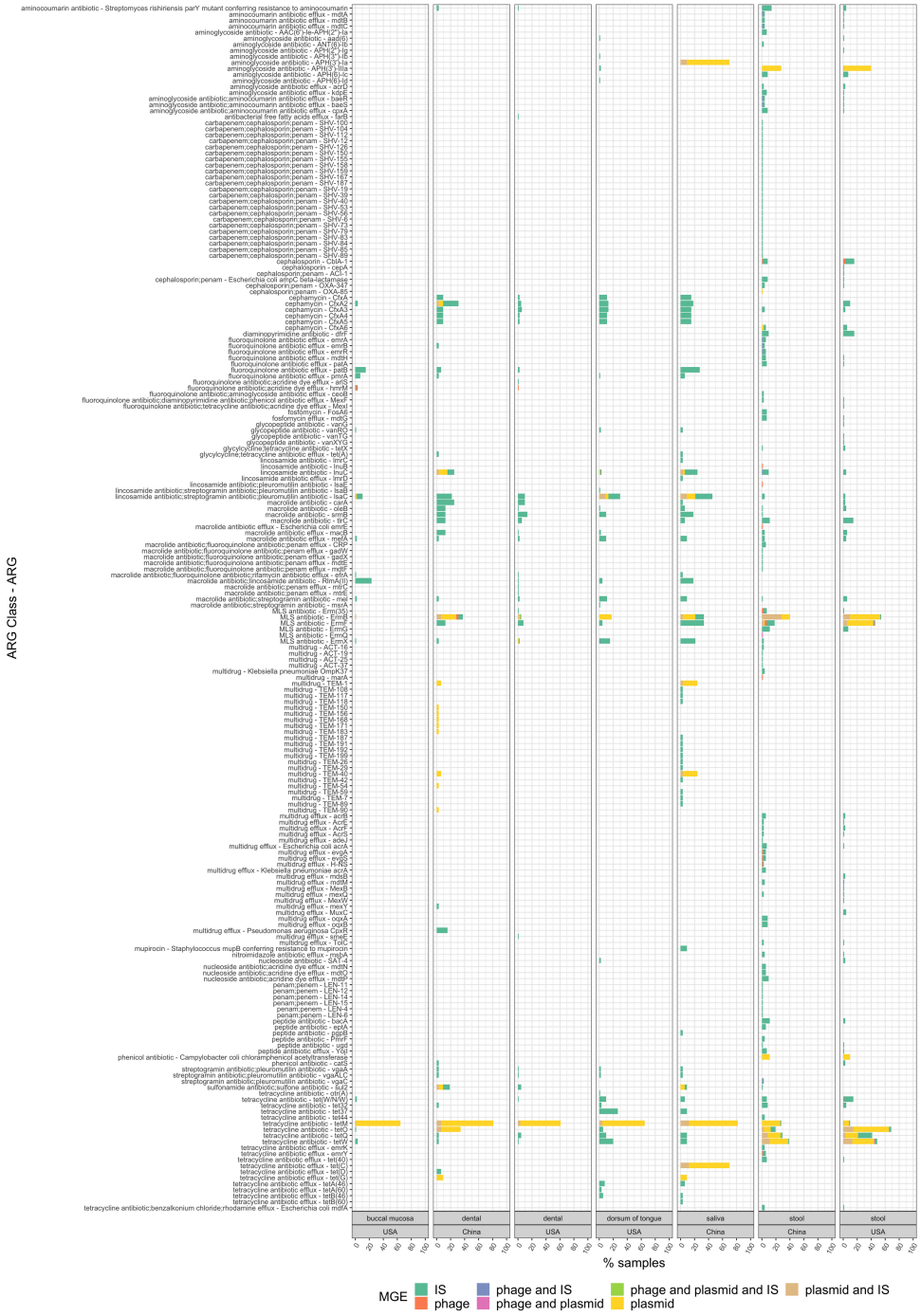

Figure 6.2. Percentage of samples containing ARGs associated with MGE types.

ARGs are grouped by ARG classes associated with a phage only, plasmid only, IS only, phage and plasmid, phage and IS, plasmid and IS, or phage, plasmid and IS. Samples from the USA (not longitudinal) (buccal mucosa: n = 61, dental: n = 60, dorsum of the tongue: n = 61, stool: n = 61) and China (saliva: n = 27, dental: n = 26, stool: n = 29). **(Larger figure can be downloaded here:** https://tinyurl.com/y4aqkhxr**).**

6.3 The influence of MGE incidence on ARG abundance

Just because an ARG is associated with an MGE does not guarantee it is mobilisable and able to transfer into other genomes. For example, an ARG that integrates into a prophage may inactivate it, preventing the phage genome being excised from the host genome during prophage induction. When an ARG is transferred by HGT across the microbiome, its abundance should increase. To identify which MGEs may be most influential in this increase, the incidence of ARG-associated MGEs was modelled to predict the ARG abundance using logistic regression. The presence of four ISs and four plasmids are found to be significantly related to an increase in ARG abundance across different cohorts **(Fig. 6.3)** ($p < 0.05$ from Student's t-test of estimated MGE coefficient, and $p < 0.05$ chi-square test between model and null model). No phages are associated with ARG abundance. The presence of IS 126965 (ITR cluster ID from Chapter 5) is associated with an increase in the abundance of ARG *bacA* (conferring resistance to peptide antibiotics in stool samples from China [$p = 5.02 \times 10^{-4}$]) and Fiji [$p = 4.08 \times 10^{-9}$]), *acrD* (encoding resistance to aminoglycosides [$p = 9.71 \times 10^{-11}$]) and *mdtM* (conferring multidrug resistance [$p = 1.83 \times 10^{-8}$] in stool samples from Fiji only). The abundance of MLS resistant ARG *ermB* is also significantly related to the incidence of three different plasmids: one in stool samples from China ($p = 2.67 \times 10^{-4}$) and two in stool samples from the USA (both $p = 5.00 \times 10^{-6}$). In addition, the abundances of fosfomycin resistant ARGs *mdtG*, *lsaC* and *tetW* are significantly associated with the incidence of an IS in Fiji stool samples ($p = 0.00103$), a plasmid in USA dorsum of the tongue ($p = 6.02 \times 10^{-4}$), and a plasmid in USA stool samples ($p = 0.0179$), respectively.

These eight MGEs are likely to be involved in the spread and propagation of their associated ARGs within GIT microbiomes leading to their increased abundances. As logistic regression was only performed when there were ten or more samples carrying MGE-associated ARGs, other MGEs with a lower prevalence that were omitted in this model are also likely to be involved in propagating ARGs. More samples would be required to detect whether less prevalent MGEs also propagate ARGs.

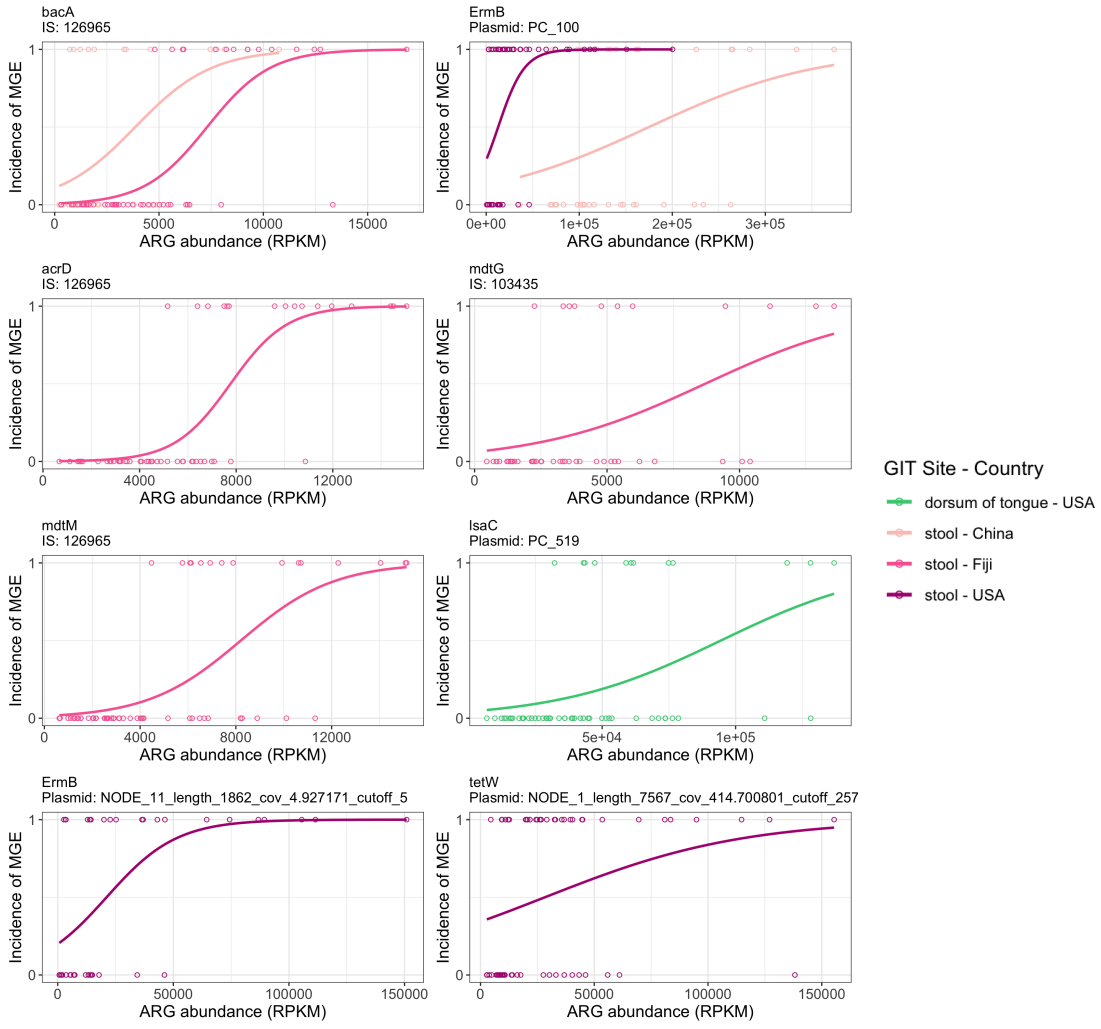

Figure 6.3. MGE incidence versus ARG abundance for MGEs with statistical significance.
ARG abundance is measured by RPKM, and incidence of MGEs describes presence (1) or absence (0). Logistic regression is applied when 10 or more samples have ARGs with or without an associated MGE. Statistical significance is found when $p < 0.05$ from Student's t-test of estimated MGE coefficient and $p < 0.05$ from chi-square test between model and null model.

6.4 Which ARGs that associate with MGEs may be important in AMR?

The prevalence of ARGs and ARG-associated MGEs can be used to predict which ARGs and MGEs may be implicated in causing AMR for particular antibiotic classes within populations and body sites. Highly prevalent ARG-associated MGEs are likely to be more successful in spreading ARGs in human populations compared to less common ARG-associated MGEs. The following methodology was applied to highlight these MGEs. For every antibiotic class, cohort and GIT site, each ARG was ranked in descending order by their prevalence of being associated with ISs, phages and plasmids in samples. For example, in dental plaque samples from China, the tetracycline resistant ARGs *tetM* and *tetO* are located in a plasmid in all samples (100%). The ARG *tetM* is associated with ISs in 60% of samples, whereas *tetO* is associated with ISs in 50% of individuals. Neither ARGs are associated with phages (0% samples). The sum of the proportions of individuals with ARGs associated with MGEs are 160% for *tetM* and 150% for *tetO*. *tetM* has the highest sum and is assigned a rank of "1", then *tetO* is given a rank of "2". If no other tetracycline resistant ARG is linked to an MGE in that cohort, then the rank is continued in descending order by the proportion of individuals containing the ARG without being associated with an MGE. For example, *tetQ* is found in 96% samples but are not associated with any MGEs, so is given a rank of "3". The next most prevalent ARG, *tetB(60)* in 60% of samples, is assigned the rank "4". ARGs with the same percentages are given a tied rank, e.g. if *tet32* also has a prevalence of 60%, it is also given the rank "4", and the next ARG has the rank "6". **Table 6.1** summarises the top ranking (i.e. rank "1") of ARGs for each antibiotic class, country

and GIT site. Highest ranked ARGs that are highly prevalent and linked to MGEs usually confer resistance to multiple antibiotics. Rankings of all ARGs for all antibiotic classes can be found in Supplementary Data 6.1 available here: https://tinyurl.com/y6amgn4n. (Each antibiotic is separated by sheet and follows the same layout as Table 6.1.)

Table 6.1. ARGs that have the highest prevalence of being associated with an MGE.
ARGs are ranked by the sum of their prevalence of being associated with an IS, phage and plasmid for each antibiotic class, country and GIT site.

Antibiotic Class	Country	GIT Site	ARG	Prevalence of ARGs (%)	Prevalence of ARG-associated ISs (%)	Prevalence of ARG-carrying plasmids (%)	Prevalence of ARG-carrying phages (%)
acridine dye	China	stool	*mdtM*	55.6	4.17	0	0
	USA	buccal mucosa	*hmrM*	37.9	0	0	2.3
	USA	dental	*hmrM*	11.1	0	0	1.11
	USA	stool	*mdtM*	7.14	1.43	0	1.43
aminoglycoside antibiotic	China	saliva	*APH(3')-Ia*	100	0	69.7	0
	China	stool	*APH(3')-IIIa*	41.7	0	27.8	0
	USA	dorsum of tongue	*APH(3')-IIIa*	6.59	2.2	0	0
	USA	stool	*APH(3')-IIIa*	67.1	0	40	0
benzalkonium chloride	China	stool	*Escherichia coli mdfA*	62.5	4.17	0	0
	USA	stool	*Escherichia coli mdfA*	7.14	1.43	0	0
carbapenem	China	stool	*Klebsiella pneumoniae OmpK37*	37.5	4.17	0	0
cephalosporin	China	dental	*TEM-1*	6.25	0	6.25	0
	China	dental	*TEM-40*	3.12	0	6.25	0
	China	saliva	*TEM-1*	36.4	0	24.2	0
	China	saliva	*TEM-40*	3.03	0	24.2	0
	China	stool	*Escherichia coli ampC beta-lactamase*	68.1	8.33	0	0
	USA	stool	*CblA-1*	87.1	12.9	0	0
cephamycin	China	dental	*CfxA2*	87.5	25	0	0
	China	saliva	*CfxA2*	69.7	18.2	0	0
	China	stool	*CfxA6*	29.2	2.78	2.78	0
	USA	buccal mucosa	*CfxA2*	39.1	3.45	0	0
	USA	dental	*CfxA3*	14.4	5.56	0	0

Continues next page

	USA	dorsum of tongue	*CfxA2*	67	13.2	0	0
	USA	dorsum of tongue	*CfxA3*	33	13.2	0	0
	USA	stool	*CfxA2*	60	10	0	0
diaminopyrimidine antibiotic	China	stool	*dfrF*	100	9.72	0	0
	USA	stool	*dfrF*	82.9	15.7	0	0
fluoroquinolone antibiotic	China	dental	*pmrA*	46.9	3.12	0	0
	China	saliva	*pmrA*	100	6.06	0	0
	China	stool	*Escherichia coli acrA*	69.4	6.94	0	0
	USA	buccal mucosa	*pmrA*	100	6.9	0	0
	USA	dental	*hmrM*	11.1	0	0	1.11
	USA	dorsum of tongue	*pmrA*	79.1	1.1	0	0
	USA	stool	*mdtM*	7.14	1.43	0	1.43
fosfomycin	China	stool	*FosA6*	40.3	6.94	0	0
	China	stool	*mdtG*	59.7	6.94	0	0
	USA	stool	*mdtG*	8.57	1.43	0	0
glycopeptide antibiotic	USA	stool	*vanTG*	2.86	1.43	0	0
	USA	stool	*vanXYG*	7.14	1.43	0	0
glycylcycline	China	dental	*tet(A)*	71.9	3.12	0	0
	China	saliva	*tet(A)*	84.8	3.03	0	0
	China	stool	*Escherichia coli acrA*	69.4	6.94	0	0
	USA	stool	*tetX*	68.6	2.86	0	0
lincosamide antibiotic	China	dental	*ErmB*	100	0	28.1	0
	China	saliva	*ErmF*	100	33.3	0	0
	China	saliva	*lsaC*	97	33.3	0	0
	China	stool	*ErmB*	100	0	40.3	0
	USA	buccal mucosa	*RlmA(II)*	98.9	23	0	0
	USA	dental	*lsaC*	90	10	0	0
	USA	dorsum of tongue	*lsaC*	100	25.3	0	0
	USA	stool	*ErmB*	82.9	0	52.9	0
macrolide antibiotic	China	dental	*ErmB*	100	0	28.1	0
	China	saliva	*ErmF*	100	33.3	0	0
	China	stool	*ErmB*	100	0	40.3	0
	USA	buccal mucosa	*RlmA(II)*	98.9	23	0	0
	USA	dental	*ErmF*	81.1	7.78	0	0
	USA	dorsum of tongue	*ErmB*	87.9	0	17.6	0
	USA	stool	*ErmB*	82.9	0	52.9	0
monobactam	China	dental	*TEM-1*	6.25	0	6.25	0
	China	dental	*TEM-40*	3.12	0	6.25	0
	China	saliva	*TEM-1*	36.4	0	24.2	0
	China	saliva	*TEM-40*	3.03	0	24.2	0
	China	stool	*Klebsiella pneumoniae OmpK37*	37.5	4.17	0	0

Continues next page

nitroimidazole antibiotic	China	stool	*msbA*	72.2	4.17	0	0
	USA	stool	*msbA*	14.3	1.43	0	0
nucleoside antibiotic	China	stool	*mdtM*	55.6	4.17	0	0
	USA	dorsum of tongue	*SAT-4*	7.69	2.2	0	0
	USA	stool	*mdtM*	7.14	1.43	0	1.43
	USA	stool	*SAT-4*	51.4	2.86	0	0
penam	China	dental	*TEM-1*	6.25	0	6.25	0
	China	dental	*TEM-40*	3.12	0	6.25	0
	China	saliva	*TEM-1*	36.4	0	24.2	0
	China	saliva	*TEM-40*	3.03	0	24.2	0
	China	stool	*Escherichia coli ampC beta-lactamase*	68.1	8.33	0	0
	USA	stool	*ACI-1*	2.86	1.43	0	0
	USA	stool	*Escherichia coli acrA*	8.57	1.43	0	0
	USA	stool	*Escherichia coli ampC beta-lactamase*	12.9	1.43	0	0
	USA	stool	*OXA-347*	17.1	1.43	0	0
penem	China	dental	*TEM-1*	6.25	0	6.25	0
	China	dental	*TEM-40*	3.12	0	6.25	0
	China	saliva	*TEM-1*	36.4	0	24.2	0
	China	saliva	*TEM-40*	3.03	0	24.2	0
	China	stool	*Klebsiella pneumoniae OmpK37*	37.5	4.17	0	0
peptide antibiotic	China	saliva	*pgpB*	51.5	3.03	0	0
	China	stool	*bacA*	63.9	11.1	0	0
	USA	stool	*bacA*	12.9	2.86	0	0
phenicol antibiotic	China	dental	*catS*	3.12	3.12	0	0
	China	stool	*Campylobacter coli chloramphenicol acetyltransferase*	47.2	0	11.1	0
	USA	stool	*Campylobacter coli chloramphenicol acetyltransferase*	20	0	10	0
pleuromutilin antibiotic	China	dental	*lsaC*	100	21.9	0	0
	China	saliva	*lsaC*	97	33.3	0	0
	China	stool	*lsaC*	1.39	4.17	0	0
	USA	buccal mucosa	*lsaC*	60.9	9.2	0	0
	USA	dental	*lsaC*	90	10	0	0
	USA	dorsum of tongue	*lsaC*	100	25.3	0	0
rhodamine	China	stool	*Escherichia coli mdfA*	62.5	4.17	0	0
	USA	stool	*Escherichia coli mdfA*	7.14	1.43	0	0
rifamycin antibiotic	China	stool	*Escherichia coli acrA*	69.4	6.94	0	0
	USA	stool	*Escherichia coli acrA*	8.57	1.43	0	0
streptogramin antibiotic	China	dental	*ErmB*	100	0	28.1	0
	China	saliva	*ErmF*	100	33.3	0	0
	China	saliva	*lsaC*	97	33.3	0	0
	China	stool	*ErmB*	100	0	40.3	0

Continues next page

	USA	buccal mucosa	*lsaC*	60.9	9.2	0	0
	USA	dental	*lsaC*	90	10	0	0
	USA	dorsum of tongue	*lsaC*	100	25.3	0	0
	USA	stool	*ErmB*	82.9	0	52.9	0
sulfonamide antibiotic	China	dental	*sul2*	50	12.5	0	0
	China	saliva	*sul2*	60.6	0	6.06	0
	China	stool	*sul2*	73.6	1.39	0	0
	USA	dental	*sul2*	16.7	4.44	0	0
sulfone antibiotic	China	dental	*sul2*	50	12.5	0	0
	China	saliva	*sul2*	60.6	0	6.06	0
	China	stool	*sul2*	73.6	1.39	0	0
	USA	dental	*sul2*	16.7	4.44	0	0
tetracycline antibiotic	China	dental	*tetM*	100	0	81.2	0
	China	saliva	*tetM*	100	0	81.8	0
	China	stool	*tetW*	100	0	37.5	0
	USA	buccal mucosa	*tetM*	86.2	0	64.4	0
	USA	dental	*tetM*	91.1	0	61.1	0
	USA	dorsum of tongue	*tetM*	100	0	64.8	0
	USA	stool	*tetO*	94.3	0	65.7	0
triclosan	China	stool	*Escherichia coli acrA*	69.4	6.94	0	0
	USA	stool	*Escherichia coli acrA*	8.57	1.43	0	0

6.5 To what extent can short-read, whole metagenomic data be useful in predicting AMR?

Surveillance of the spread of antimicrobial resistant pathogens and the clinical consequences of infections caused by resistant pathogens is important in understanding both the spread and health impacts of AMR in human populations. The genetic determinants of how a pathogen becomes resistant to antimicrobials are also crucial to our understanding of how AMR emerges. An ARG acquired by an MGE or by mutation that is retained in the genome is the first step towards a microbe being able to survive against antimicrobial treatments. Even so, the likelihood of an acquired ARG to pose a threat to human and animal health relies on other conditions, including whether the acquired ARG is in a pathogen and whether the ARG can be expressed sufficiently

enough for pathogen to survive antimicrobial treatment. Although profiling the prevalence of the mobile resistome gives an indication of which genetic determinants may pose a threat, a combination of alternative modelling and experimental techniques may better inform us as to whether an acquired ARG can pass these necessary conditions. Modelling allows us to represent a biological system based on empirical evidence from experiments to make predictions. The purpose of the model would be to predict a pathogen's AMR phenotype to an antimicrobial drug, such as its MIC of ECOFF value, based on acquired ARGs from short-read metagenomic data. Hereafter, I discuss to what extent information about acquired ARGs, microbial hosts and ARG expression can be drawn from short-read metagenomics and other technologies to predict the AMR phenotype of a pathogen in a microbial community.

6.5.1 Profiling acquired ARGs

In this study, only a subset of MGEs, i.e. ISs, phages and plasmids, were considered as vectors for ARG transfer. In reality, ARGs are also transferred by other MGEs, including gene cassettes/integrons, ICEs and IMEs[217], and also can be part of multiple MGEs, such as within transposable elements on plasmids[106]. Integron_Finder[133] and ICEberg[259] are reference-based tools that identify integrons and ICEs/IMEs from assembled metagenomic data, respectively. However, as previously discussed, *de novo* discovery tools are more desirable for identifying novel MGEs that may be missed by reference-based tools. An approach to cataloguing mobile genes in species that do not rely on profiling MGEs, e.g. ISs, phages or plasmids, is to identify identical or similar genes in reference genomes and in distantly related sequences of assembled metagenomes, indicating these genes must have moved across different species via

HGT[142]. ARGs can also be acquired by mutation, including SNPs and indels. A non-reference-based *de novo* method using De Bruijn graph structures of co-assembled multiple genomes can identify SNPs and indels[364,365]. This has been previously applied to whole metagenomic data of simple microbial communities[366], and could be developed further for more complex communities, such as within the GIT.

6.5.2 Identifying pathogens carrying ARGs

The hosts of the acquired ARGs need to be found to determine whether they could cause clinically relevant AMR. For example, an *ermB*-carrying plasmid that is in *Streptococcus oralis* (a frequent commensal of the oral cavity) but not in the pathogen *Staphylococcus aureus*, would not cause clinically relevant AMR unless this plasmid is able to transfer later to *S. aureus*. In the case of phages, assemblies of short-read metagenomics contain CRISPR spacers in a small proportion of phage genomes that are recognisable in bacterial reference sequences, which have enabled bacterial host predictions to be made (Chapter 3). Similarly, in this study, a handful of plasmid sequences showed similarities to plasmid DNA isolated from known strains (Chapter 4). However, the hosts of ISs were not predicted in this study. PaliDIS, the tool created to identify ISs from short-read, paired-end metagenomes (Chapter 5), could also incorporate a host prediction function in the future. One method would be to create a catalogue of metagenomic species from computational binning of metagenomic reads by their differential abundance across samples that are then assembled per bin[235]. These metagenomic species can be mapped against reference genomes to identify its species or strain. This has already been done for both the human gut and oral cavity[420,421], so the reads associated with ITRs of ISs can be mapped to these metagenomic species to

predict the ISs' hosts. As well as using short-read metagenomics, long-read sequencing technologies, such as Nanopore or PacBio, can generate longer lengths of unique chromosomal DNA containing prophages and ISs, which can be resolved at the strain-level. Since the majority of plasmids are separate from chromosomal DNA, proximity ligation technologies or clustering by methylation motifs can be applied to metagenomes to locate their hosts[217]. These techniques are currently the only ways to retrieve reliable host predictions given plasmids are highly promiscuous, such that one plasmid can be found in multiple species[105].

It is also important to taxonomically profile the microbial community to identify other pathogens or pathobionts that may acquire ARGs from a commensal species in the future. For example, *ermB*-carrying plasmids could transfer from commensal *Streptococcus oralis* to pathogen *Staphylococcus aureus* within the microbial community. If a pathobiont was part of a microbial community that colonised a particular niche some time ago and did not acquire a particular ARG, it could be argued that, given the microbial community continues to remain stable with unchanging exposures to metabolites and exogenous compounds, it is unlikely to acquire an ARG in the future. However, the composition of external environments can often change, especially in the oral cavity. HGT within microbial communities can be promoted by exposure to sublethal levels of antimicrobials[422] and other exogenous compounds[423]. For example, cefotaxime exposure significantly raised the conjugation frequency of *blaCTX−M−1*-carrying IncI1 resistance plasmid in an *Escherichia coli* strain[424]. The rate of MGE transfer of ARGs between different microorganisms may need to be incorporated to model the likelihood of a pathogen or pathobiont acquiring an ARG.

Published mathematical models have attempted to recreate the dynamics of HGT of ARGs, but have often fallen short of representing a realistic scenario. Most of them are simplified deterministic mathematical models[xxii] of one ARG transferred only by conjugation between *E. coli* strains[425]. They are also constrained by parameters from *in vitro* experiments without antimicrobial exposure, but including some fitness costs to the bacteria acquiring new ARGs. In reality, a representative mathematical model of HGT should include stochasticity as HGT can be a rare event[426,427]. It should also include multiple ARGs, transferred by different HGT mechanisms between multiple species simulating under a variety of *in vivo* conditions, including antimicrobial exposure. These transfer rates of different MGEs and HGT mechanisms can vary between bacterial species. For example, transformation is more common in *Neisseria gonorrhoea* than in *Staphylococcus aureus*[428,429]. This is a very complex system. Without accurate measurements of these conditions to confine all possible solutions, the outcomes of complex mathematical models become less precise and thus meaningless. Instead, predictions of whether a pathogen or pathobiont could acquire an ARG in a microbial community could be aided by finding if they exist in metagenomic samples and reference genomes from other environmental niches, and thus have the potential to host ARGs in the human body[87]. If a pathogen or pathobiont outside the human body is found with an acquired ARG, it has the potential to spread into the human body at sites of infection. For example, the MLS resistant ARG *ermG* was first found in the soil bacterium *Bacillus sphaericus*[430] and much later in several human intestinal *Bacteroides species*[431] and in a conjugative transposon[432]. This indicates that there may have been HGT of *ermG* from *Bacillus sphaericus* and proximal enteric *Bacteroides* species, which then colonised the human gut. Therefore, metagenomic data needs to be collected

xxii No randomness is involved and will always produce the same result given initial conditions.

from all other environments of human contact, such as households, transportation and hospitals, to find ARG-carrying pathogens or pathobionts that could colonise the human body.

6.5.3 ARG expression of a resistance phenotype

Finally, once ARGs are newly modelled into the genomes, it has to be established whether they are expressed and what resistance phenotype is presented during antimicrobial treatment of infections. It is impossible to determine directly whether an ARG from metagenomic data is expressible or not. Metatranscriptomic sequencing of mRNA can determine whether acquired ARGs are expressed during antimicrobial treatment. However, this would only detect protein-coding ARGs that are expressed and promoter regions that are increasing their expression at that time, and usually only while there is antimicrobial exposure. Moreover, expression levels can only be profiled from known ARGs using metatranscriptomics. Instead, functional metagenomics can be applied to discover novel genetic resistance determinants associated with MGEs[359], including promoters as well as ARGs. Only genetic determinants presenting a resistant phenotype in the surrogate host (usually *E. coli*) that survive antimicrobial exposure are sequenced, meaning that genetic determinants that are missed may be expressed in the original host or genetic determinants that are discovered in the surrogate may not be expressed in their original host. Even when an ARG is found to be expressible in one genome under a particular condition of antimicrobial exposure, it may be expressed differently in other genetic backgrounds. Multiple ARGs may interact to potentiate each other's expression more than if expressed separated. For example, the efflux pump NorA in *Staphylococcus aureus* that exports fluoroquinolone antibiotics can lead to a

more rapid recruitment of other mechanisms of ciprofloxacin (a fluoroquinolone) resistance during exposure, including intrinsic resistance mediated by a DNA topoisomerase[146].

6.5.4 Future work with short-read whole metagenomes

It is possible to identify the hosts of acquired ARGs with short-read, whole metagenomic data, yet it is impossible to directly ascertain whether acquired ARGs are expressible via this approach. However, there is potential to predict AMR phenotype without knowledge of an ARG's expression from short-read metagenomic data. Long-read metagenomic sequencing of sputum samples using Nanopore MinION of metagenomes and pre-existing susceptibility and genomic data for genomic neighbour typing[xxiii] were enough to accurately predict antibiotic resistance and susceptibility phenotypes of *Streptococcus pneumoniae* strains in these samples[215]. It would be challenging to apply this particular genomic neighbouring typing method to short-read metagenomes, mainly due to limitations in resolving near-complete genomes of strains. However, predicting the host species of the mobile resistome (without resolving genomes) from short-read metagenomes can provide information on the incidence of acquired ARGs for different species within a metagenomic sample. Comparing these incidences between metagenomic data and reference genomes with known susceptibility data could aid predictions of whether a metagenomic species with an acquired ARG may be more resistant than its susceptible reference counterpart. Future work might include calculating the distances (such as Euclidean distance) of acquired ARG profiles between metagenomic species and reference strains of the same species. The

xxiii Predicting the phenotype of the pathogen by their closest relatives using *k*-mer content of reads.

susceptibility of the metagenomic species to the antimicrobial would be most similar to the susceptibility of reference strain with the most similar acquired ARG profiles. Currently, susceptibility to an antimicrobial can only be derived from phenotypic testing of single isolates in culture. Nevertheless, continuing to characterise the susceptibility of reference microbial genomes could improve AMR phenotype predictions from whole metagenomic data.

6.6 Interventions against AMR

Pathogens that are predicted to be resistant to an antimicrobial from metagenomic data are likely to lead to poor clinical outcomes (prolonged infections and fatalities) when they survive and colonise at sites of infection under antimicrobial treatment. Once these pathogens are identified, three strategies can be implemented to prevent these clinical outcomes: 1) surveillance to monitor the spread and colonisation of these pathogens in human populations; 2) diagnostics to choose appropriate antimicrobial treatments; and 3) alternative therapies to eradicate acquired ARGs and MGEs.

6.6.1 Surveillance

Surveillance of AMR has been highlighted as a major action point from public health bodies, including the World Health Organisation. The spread of antimicrobial resistant pathogens and acquired ARGs that pose a threat to human life can be monitored across different environments, such as hospitals, transport, farms and wastewater. Acquired ARGs and MGEs or marker sequences of the pathogen can be targeted and amplified

with specific sequence probes to rapidly test the presence of the pathogen. These tests can be done using PCR or loop-mediated isothermal amplification (LAMP) that does not require alternating temperature cycles like PCR[433]. Antimicrobial resistant pathogens found in environments where there is frequent human contact, such as hospitals, are considered to be a higher risk of causing clinical AMR. This can inform decision-making in implementing procedures to prevent further spreading in hospitals, such as regular hand sanitisation for all staff and visitors. The genomes of resistant pathogens can also be sequenced based on culture-based or long-read metagenomic techniques to track the evolution of their genetic virulence (such as how they colonise niches, interrupt the immune response and disrupt human metabolism) and resistance determinants, and whether these pose a greater danger to health.

6.6.2 Diagnostics

PCR, LAMP and metagenomic sequencing to identify antimicrobial resistant pathogens used in surveillance can also be applied to diagnose whether a patient has a pathogen that could colonise a site of infection under antimicrobial treatment. If an individual tests positive, an alternative antimicrobial treatment can be administered to eradicate an infection, reducing the risk of poor clinical outcomes.

6.6.3 Therapies to prevent resistant infections

Other prevention methods, like vaccines, can be developed and administered to prevent infection of resistant pathogens and other related species[434]. One possibility is using phage therapy that infect and lyse specific bacteria, which has been used especially to

eradicate bacterial infections in Eastern Europe since WW2[95]. Phage therapy can also be applied to modulate the microbial community by removing microbes that transfer ARGs into pathogen genomes by MGEs. Although phage therapy has been shown to work in practice[435], approving its use in countries with strict drug regulations is difficult as phage therapy is administered as cocktail of several different types of phages[436], meaning clinical trials on their safety and efficacy have to be conducted on each phage. In contrast to antimicrobials, phage therapy is also highly personalised and used against very specific bacteria, making it more difficult to test safety and efficacy. Another alternative therapy is faecal microbiota transplantation where faeces from a healthy donor is transplanted into the gut of the recipient, restoring the gut microbiota after antimicrobial treatment to eradicate recurrent infections, like *Clostridium difficile*[437], or to prevent colonisation of resistant pathogens, such as multidrug resistant *Enterobacteriaceae*[438]. In both circumstances there are concerns that off-target effects may remove keystone species that could alter the microbiota function and produce long-term side-effects. CRISPR-Cas based technologies are promising avenues for the eradication of antimicrobial resistant pathogens by the selective knockdown of pathogens with undesirable genetic AMR determinants. This is achieved using RNA-guide nucleases (RGNs) that target specific DNA sequences and have a modified CRISPR-Cas system to affect cell death once these sequences are detected. These RGNs can be delivered directly into a specific bacterium using phages or plasmids in other bacteria that can be transferred via conjugation[439].

6.7 Concluding remarks

This study has provided computational methods and frameworks for profiling both the resistome and the mobilome in whole, short-read metagenomes. These methods have been applied to profile the resistome and mobilome in human GIT sites worldwide from publicly available short-read metagenomic data. In doing so this thesis provides, for the first time, a comparison of the resistome and mobilome between the oral cavity and the gut, and the development of new pipelines and software for identifying bacteriophages, plasmids and transposons from short-read metagenomic data. Additionally, this study integrates the analysis of the resistome and mobilome to provoke ideas on how AMR can be predicted using whole, short-read metagenomic data. These ideas are not just limited to short-read metagenomic data, but can be built alongside advancements in sequencing technology, such as SMRT and Nanopore sequencing that are being developed for effective surveillance and point-of-care diagnostics. Thus, I hope the methods, data, results and discussion of this thesis will prove useful and interesting for researchers developing a better understanding of the spread of AMR.

References

1. Sender, R., Fuchs, S. & Milo, R. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell* **164**, 337–340 (2016).
2. Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLOS Biol.* **14**, e1002533 (2016).
3. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
4. Shkoporov, A. N. & Hill, C. Bacteriophages of the Human Gut: The “Known Unknown” of the Microbiome. *Cell Host Microbe* **25**, 195–209 (2019).
5. Sam, Q. H., Chang, M. W. & Chai, L. Y. A. The Fungal Mycobiome and Its Interaction with Gut Bacteria in the Host. *Int. J. Mol. Sci.* **18**, 330 (2017).
6. Lurie-Weinberger, M. N. & Gophna, U. Archaea in and on the Human Body: Health Implications and Future Directions. *PLOS Pathog.* **11**, e1004833 (2015).
7. Chabé, M., Lokmer, A. & Ségurel, L. Gut Protozoa: Friends or Foes of the Human Gut Microbiota? *Trends Parasitol.* **33**, 925–934 (2017).
8. Gilbert, J. A. *et al.* Current understanding of the human microbiome. *Nat. Med.* **24**, 392–400 (2018).
9. Turnbaugh, P. J. *et al.* The Human Microbiome Project. *Nature* **449**, 804–810 (2007).
10. Maynard, C. L., Elson, C. O., Hatton, R. D. & Weaver, C. T. Reciprocal interactions of the intestinal microbiota and immune system. *Nature* **489**, 231–241 (2012).
11. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).

12. Frank, D. N. *et al.* Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 13780–13785 (2007).
13. Kostic, A. D. *et al.* Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe* **14**, 207–215 (2013).
14. Jiang, H. *et al.* Altered fecal microbiota composition in patients with major depressive disorder. *Brain. Behav. Immun.* **48**, 186–194 (2015).
15. Segata, N. *et al.* Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. *Genome Biol.* **13**, R42 (2012).
16. Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.* **14**, 20–32 (2016).
17. Schnorr, S. L. *et al.* Gut microbiome of the Hadza hunter-gatherers. *Nat. Commun.* **5**, 3654 (2014).
18. Smits, S. A. *et al.* Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. *Science* **357**, 802–806 (2017).
19. de Goffau, M. C. *et al.* Human placenta has no microbiome but can contain potential pathogens. *Nature* **572**, 329–334 (2019).
20. Vaishampayan, P. A. *et al.* Comparative metagenomics and population dynamics of the gut microbiota in mother and infant. *Genome Biol. Evol.* **2**, 53–66 (2010).
21. Gueimonde, M. *et al.* Effect of maternal consumption of lactobacillus GG on transfer and establishment of fecal bifidobacterial microbiota in neonates. *J. Pediatr. Gastroenterol. Nutr.* **42**, 166–170 (2006).
22. Koenig, J. E. *et al.* Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4578–4585 (2011).

23. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity, stability and resilience of the human gut microbiota. *Nature* **489**, 220–230 (2012).
24. Shao, Y. *et al.* Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**, 117–121 (2019).
25. Goodrich, J. K. *et al.* Human Genetics Shape the Gut Microbiome. *Cell* **159**, 789–799 (2014).
26. Schirmer, M. *et al.* Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. *Cell* **167**, 1125-1136.e8 (2016).
27. Zheng, D., Liwinski, T. & Elinav, E. Interaction between microbiota and immunity in health and disease. *Cell Res.* **30**, 492–506 (2020).
28. Ivanov, I. I. *et al.* Specific Microbiota Direct the Differentiation of IL-17-Producing T-Helper Cells in the Mucosa of the Small Intestine. *Cell Host Microbe* **4**, 337–349 (2008).
29. Hamada, H. *et al.* Identification of Multiple Isolated Lymphoid Follicles on the Antimesenteric Wall of the Mouse Small Intestine. *J. Immunol.* **168**, 57–64 (2002).
30. Bouskra, D. *et al.* Lymphoid tissue genesis induced by commensals through NOD1 regulates intestinal homeostasis. *Nature* **456**, 507–510 (2008).
31. Freire, M. *et al.* Longitudinal Study of Oral Microbiome Variation in Twins. *Sci. Rep.* **10**, 7954 (2020).
32. Kovatcheva-Datchary, P. *et al.* Simplified Intestinal Microbiota to Study Microbe-Diet-Host Interactions in a Mouse Model. *Cell Rep.* **26**, 3772-3783.e6 (2019).
33. Song, S. J. *et al.* Cohabiting family members share microbiota with one another and with their dogs. *eLife* **2**, (2013).
34. Cook, M. D. *et al.* Exercise and gut immune function: evidence of alterations in colon immune cell homeostasis and microbiome characteristics with exercise training. *Immunol. Cell Biol.* **94**, 158–163 (2016).

35. Karl, J. P. *et al.* Changes in intestinal microbiota composition and metabolism coincide with increased intestinal permeability in young adults under prolonged physiological stress. *Am. J. Physiol. Gastrointest. Liver Physiol.* **312**, G559–G571 (2017).
36. Ying, S. *et al.* The Influence of Age and Gender on Skin-Associated Microbial Communities in Urban and Rural Human Populations. *PLoS ONE* **10**, (2015).
37. Huang, C. & Shi, G. Smoking and microbiome in oral, airway, gut and some systemic diseases. *J. Transl. Med.* **17**, 225 (2019).
38. Benedict, C. *et al.* Gut microbiota and glucometabolic alterations in response to recurrent partial sleep deprivation in normal-weight young individuals. *Mol. Metab.* **5**, 1175–1186 (2016).
39. Modi, S. R., Collins, J. J. & Relman, D. A. Antibiotics and the gut microbiota. *J. Clin. Invest.* **124**, 4212–4218 (2014).
40. Jernberg, C., Löfmark, S., Edlund, C. & Jansson, J. K. Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. *ISME J.* **1**, 56–66 (2007).
41. Dethlefsen, L. & Relman, D. A. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 4554–4561 (2011).
42. Sturød, K., Dhariwal, A., Dahle, U. R., Vestrheim, D. F. & Petersen, F. C. Impact of narrow-spectrum penicillin V on the oral and faecal resistome in a young child treated for otitis media. *J. Glob. Antimicrob. Resist.* **20**, 290–297 (2020).
43. Palleja, A. *et al.* Recovery of gut microbiota of healthy adults following antibiotic exposure. *Nat. Microbiol.* **3**, 1255 (2018).
44. Forslund, K., Sunagawa, S., Coelho, L. P. & Bork, P. Metagenomic insights into the human gut resistome and the forces that shape it. *BioEssays* **36**, 316–329 (2014).

45. Imhann, F. *et al.* Proton pump inhibitors affect the gut microbiome. *Gut* **65**, 740–748 (2016).

46. Rogers, M. A. M. & Aronoff, D. M. The influence of non-steroidal anti-inflammatory drugs on the gut microbiome. *Clin. Microbiol. Infect.* **22**, 178.e1-178.e9 (2016).

47. Flowers, S. A., Evans, S. J., Ward, K. M., McInnis, M. G. & Ellingrod, V. L. Interaction Between Atypical Antipsychotics and the Gut Microbiome in a Bipolar Disease Cohort. *Pharmacother. J. Hum. Pharmacol. Drug Ther.* **37**, 261–267 (2017).

48. Maier, L. *et al.* Extensive impact of non-antibiotic drugs on human gut bacteria. *Nature* **555**, 623–628 (2018).

49. Bokulich, N. A. *et al.* Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci. Transl. Med.* **8**, 343ra82 (2016).

50. Örtqvist, A. K., Lundholm, C., Halfvarson, J., Ludvigsson, J. F. & Almqvist, C. Fetal and early life antibiotics exposure and very early onset inflammatory bowel disease: a population-based study. *Gut* **68**, 218–225 (2019).

51. Arrieta, M.-C. *et al.* Early infancy microbial and metabolic alterations affect risk of childhood asthma. *Sci. Transl. Med.* **7**, 307ra152-307ra152 (2015).

52. *Antimicrobial resistance: global report on surveillance*. (World Health Organization, 2014).

53. Kraker, M. E. A. de, Stewardson, A. J. & Harbarth, S. Will 10 Million People Die a Year due to Antimicrobial Resistance by 2050? *PLOS Med.* **13**, e1002184 (2016).

54. Klein, E. Y. *et al.* Global increase and geographic convergence in antibiotic consumption between 2000 and 2015. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E3463–E3470 (2018).

55. Llor, C. & Bjerrum, L. Antimicrobial resistance: risk associated with antibiotic overuse and initiatives to reduce the problem: *Ther. Adv. Drug Saf.* (2014) doi:10.1177/2042098614554919.

56. Wright, G. D. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nat. Rev. Microbiol.* **5**, 175–186 (2007).

57. Cheng, J. *et al.* Knowledge and behaviors in relation to antibiotic use among rural residents in Anhui, China. *Pharmacoepidemiol. Drug Saf.* **27**, 652–659 (2018).

58. Wang, X. *et al.* Massive misuse of antibiotics by university students in all regions of China: implications for national policy. *Int. J. Antimicrob. Agents* **50**, 441–446 (2017).

59. Barber, D. A. *et al.* Prevalence and correlates of antibiotic sharing in the Philippines: antibiotic misconceptions and community-level access to non-medical sources of antibiotics. *Trop. Med. Int. Health* **22**, 567–575 (2017).

60. López-Lozano, J.-M. *et al.* A nonlinear time-series analysis approach to identify thresholds in associations between population antibiotic use and rates of resistance. *Nat. Microbiol.* 1 (2019) doi:10.1038/s41564-019-0410-0.

61. Cheng, G. *et al.* Antibiotic alternatives: the substitution of antibiotics in animal husbandry? *Front. Microbiol.* **5**, (2014).

62. Smith, R. Regulation (EC) No 764/2008 of the European Parliament and of the Council. in *Core EU Legislation* 183–186 (Macmillan Education UK, 2015). doi:10.1007/978-1-137-54482-7_19.

63. Liu, Y.-Y. *et al.* Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. *Lancet Infect. Dis.* **16**, 161–168 (2016).

64. Brown, N. M. *et al.* An outbreak of meticillin-resistant *Staphylococcus aureus* colonization in a neonatal intensive care unit: use of a case–control study to investigate and control it and lessons learnt. *J. Hosp. Infect.* **103**, 35–43 (2019).

65. Alvarez-Uria, G., Gandra, S. & Laxminarayan, R. Poverty and prevalence of antimicrobial resistance in invasive isolates. *Int. J. Infect. Dis.* **52**, 59–61 (2016).

66. Ayukekbong, J. A., Ntemgwa, M. & Atabe, A. N. The threat of antimicrobial resistance in developing countries: causes and control strategies. *Antimicrob. Resist. Infect. Control* **6**, 47 (2017).

67. Unemo, M. & Shafer, W. M. Antimicrobial Resistance in *Neisseria gonorrhoeae* in the 21st Century: Past, Evolution, and Future. *Clin. Microbiol. Rev.* **27**, 587–613 (2014).

68. Friedman, N. D., Temkin, E. & Carmeli, Y. The negative impact of antibiotic resistance. *Clin. Microbiol. Infect.* **22**, 416–422 (2016).

69. Årdal, C. *et al.* Antibiotic development — economic, regulatory and societal challenges. *Nat. Rev. Microbiol.* **18**, 267–274 (2020).

70. Belkum, A. van *et al.* Innovative and rapid antimicrobial susceptibility testing systems. *Nat. Rev. Microbiol.* **18**, 299–311 (2020).

71. Nathan, C. Resisting antimicrobial resistance. *Nat. Rev. Microbiol.* **18**, 259–260 (2020).

72. Hernando-Amado, S., Coque, T. M., Baquero, F. & Martínez, J. L. Defining and combating antibiotic resistance from One Health and Global Health perspectives. *Nat. Microbiol.* **4**, 1432–1442 (2019).

73. Wellcome. Reframing resistance. (2019).

74. Reygaert, W. C. An overview of the antimicrobial resistance mechanisms of bacteria. *AIMS Microbiol.* **4**, 482–501 (2018).

75. Bush, K., Jacoby, G. A. & Medeiros, A. A. A functional classification scheme for beta-lactamases and its correlation with molecular structure. *Antimicrob. Agents Chemother.* **39**, 1211–1233 (1995).

76. Giske, C. G. *et al.* Redefining extended-spectrum β-lactamases: balancing science and clinical need. *J. Antimicrob. Chemother.* **63**, 1–4 (2009).

77. Bush, K. & Jacoby, G. A. Updated Functional Classification of β-Lactamases. *Antimicrob. Agents Chemother.* **54**, 969–976 (2010).

78. Iovleva, A. & Doi, Y. Carbapenem-Resistant *Enterobacteriaceae*. *Clin. Lab. Med.* **37**, 303–315 (2017).

79. Marshall, C. G., Broadhead, G., Leskiw, B. K. & Wright, G. D. D-Ala-D-Ala ligases from glycopeptide antibiotic-producing organisms are highly homologous to the enterococcal vancomycin-resistance ligases VanA and VanB. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 6480–6483 (1997).

80. Hegde, S. S. *et al.* A Fluoroquinolone Resistance Protein from *Mycobacterium tuberculosis* That Mimics DNA. *Science* **308**, 1480–1483 (2005).

81. Utsui, Y. & Yokota, T. Role of an altered penicillin-binding protein in methicillin- and cephem-resistant *Staphylococcus aureus*. *Antimicrob. Agents Chemother.* **28**, 397–403 (1985).

82. Randall, L. P. & Woodward, M. J. The multiple antibiotic resistance (mar) locus and its significance. *Res. Vet. Sci.* **72**, 87–93 (2002).

83. Chevalier, S. *et al.* Structure, function and regulation of *Pseudomonas aeruginosa* porins. *FEMS Microbiol. Rev.* **41**, 698–722 (2017).

84. Thanassi, D. G., Cheng, L. W. & Nikaido, H. Active efflux of bile salts by *Escherichia coli*. *J. Bacteriol.* **179**, 2512–2518 (1997).

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

86. Brinkac, L., Voorhies, A., Gomez, A. & Nelson, K. E. The Threat of Antimicrobial Resistance on the Human Microbiome. *Microb. Ecol.* **74**, 1001–1008 (2017).

87. Forslund, K. *et al.* Country-specific antibiotic use practices impact the human gut resistome. *Genome Res.* **23**, 1163–1169 (2013).

88. Wu, S. W., Lencastre, H. de & Tomasz, A. Recruitment of the mecA Gene Homologue of *Staphylococcus sciuri* into a Resistance Determinant and Expression of the Resistant Phenotype in *Staphylococcus aureus*. *J. Bacteriol.* **183**, 2417–2424 (2001).

89. Griffiths, A. J., Miller, J. H., Suzuki, D. T., Lewontin, R. C. & Gelbart, W. M. Transduction. *Introd. Genet. Anal. 7th Ed.* (2000).

90. Chen, J. *et al.* Genome hypermobility by lateral transduction. *Science* **362**, 207–212 (2018).

91. Johnston, C., Martin, B., Fichant, G., Polard, P. & Claverys, J.-P. Bacterial transformation: distribution, shared mechanisms and divergent control. *Nat. Rev. Microbiol.* **12**, 181–196 (2014).

92. Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).

93. Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host Microbe* **26**, 527-541.e5 (2019).

94. Pehrsson, E. C. *et al.* Interconnected microbiomes and resistomes in low-income human habitats. *Nature* **533**, 212–216 (2016).

95. Brives, C. & Pourraz, J. Phage therapy as a potential solution in the fight against AMR: obstacles and possible futures. *Palgrave Commun.* **6**, 1–11 (2020).

96. Enault, F. *et al.* Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J.* **11**, 237–247 (2017).

97. Debroas, D. & Siguret, C. Viruses as key reservoirs of antibiotic resistance genes in the environment. *ISME J.* 1 (2019) doi:10.1038/s41396-019-0478-9.

98. Kleinheinz, K. A., Joensen, K. G. & Larsen, M. V. 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* **4**, (2014).

99. Enav, H., Mandel-Gutfreund, Y. & Béjà, O. Comparative metagenomic analyses reveal viral-induced shifts of host metabolism towards nucleotide biosynthesis. *Microbiome* **2**, 9 (2014).

100. Hayes, F. The Function and Organization of Plasmids. in *E. coli Plasmid Vectors: Methods and Applications* (eds. Casali, N. & Preston, A.) 1–17 (Humana Press, 2003). doi:10.1385/1-59259-409-3:1.

101. Solar, G. del, Giraldo, R., Ruiz-Echevarría, M. J., Espinosa, M. & Díaz-Orejas, R. Replication and Control of Circular Bacterial Plasmids. *Microbiol. Mol. Biol. Rev.* **62**, 434–464 (1998).

102. Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P. C. & Cruz, F. de la. Mobility of Plasmids. *Microbiol Mol Biol Rev* **74**, 434–452 (2010).

103. Carattoli, A. *et al.* In Silico Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. *Antimicrob. Agents Chemother.* **58**, 3895–3903 (2014).

104. Jensen, L. B. *et al.* A classification system for plasmids from enterococci and other Gram-positive bacteria. *J. Microbiol. Methods* **80**, 25–43 (2010).

105. Kent, A. G., Vill, A. C., Shi, Q., Satlin, M. J. & Brito, I. L. Widespread transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial Hi-C. *Nat. Commun.* **11**, 4379 (2020).

106. Partridge, S. R., Kwong, S. M., Firth, N. & Jensen, S. O. Mobile Genetic Elements Associated with Antimicrobial Resistance. *Clin. Microbiol. Rev.* **31**, (2018).

107. Schultsz, C. & Geerlings, S. Plasmid-Mediated Resistance in *Enterobacteriaceae*. *Drugs* **72**, 1–16 (2012).

108. Bush, K. Alarming β-lactamase-mediated resistance in multidrug-resistant *Enterobacteriaceae*. *Curr. Opin. Microbiol.* **13**, 558–564 (2010).

109. Jensen, S. O. & Lyon, B. R. Genetics of antimicrobial resistance in *Staphylococcus aureus*. *Future Microbiol.* **4**, 565–582 (2009).

110. Malachowa, N. & DeLeo, F. R. Mobile genetic elements of *Staphylococcus aureus*. *Cell. Mol. Life Sci.* **67**, 3057–3071 (2010).

111. Clewell, D. B. *et al.* Extrachromosomal and Mobile Elements in *Enterococci*: Transmission, Maintenance, and Epidemiology. in *Enterococci: From Commensals to Leading Causes of Drug Resistant Infection* (eds. Gilmore, M. S., Clewell, D. B., Ike, Y. & Shankar, N.) (Massachusetts Eye and Ear Infirmary, 2014).

112. Laverde Gomez, J. A. *et al.* A multiresistance megaplasmid pLG1 bearing a hylEfm genomic island in hospital *Enterococcus faecium* isolates. *Int. J. Med. Microbiol.* **301**, 165–175 (2011).

113. Francia, M. V. & Clewell, D. B. Amplification of the Tetracycline Resistance Determinant of pAMα1 in *Enterococcus faecalis* Requires a Site-Specific Recombination Event Involving Relaxase. *J. Bacteriol.* **184**, 5187–5193 (2002).

114. Hallet, B. & Sherratt, D. J. Transposition and site-specific recombination: adapting DNA cut-and-paste mechanisms to a variety of genetic rearrangements. *FEMS Microbiol. Rev.* **21**, 157–178 (1997).

115. Chandler, M., Fayet, O., Rousseau, P., Hoang, B. T. & Duval-Valentin, G. Copy-out–Paste-in Transposition of IS911: A Major Transposition Pathway. *Mob. DNA III* 591–607 (2015) doi:10.1128/microbiolspec.MDNA3-0031-2014.

116. Aziz, R. K., Breitbart, M. & Edwards, R. A. Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res.* **38**, 4207–4217 (2010).

117. Mahillon, J. & Chandler, M. Insertion Sequences. *Microbiol Mol Biol Rev* **62**, 725–774 (1998).

118. Harmer, C. J., Moran, R. A. & Hall, R. M. Movement of IS26-Associated Antibiotic Resistance Genes Occurs via a Translocatable Unit That Includes a Single IS26 and Preferentially Inserts Adjacent to Another IS26. *mBio* **5**, (2014).

119. Boyd, D. A. *et al.* Complete Nucleotide Sequence of a 92-Kilobase Plasmid Harboring the CTX-M-15 Extended-Spectrum Beta-Lactamase Involved in an Outbreak in Long-Term-Care Facilities in Toronto, Canada. *Antimicrob. Agents Chemother.* **48**, 3758–3764 (2004).

120. Poirel, L., Carrër, A., Pitout, J. D. & Nordmann, P. Integron Mobilization Unit as a Source of Mobility of Antibiotic Resistance Genes. *Antimicrob. Agents Chemother.* **53**, 2492–2498 (2009).

121. Snesrud, E. *et al.* A Model for Transposition of the Colistin Resistance Gene mcr-1 by ISApl1. *Antimicrob. Agents Chemother.* **60**, 6973–6976 (2016).

122. Chandler, M. *et al.* Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. *Nat. Rev. Microbiol.* **11**, 525–538 (2013).

123. Partridge, S. R. & Hall, R. M. In34, a Complex In5 Family Class 1 Integron Containing orf513 and dfrA10. *Antimicrob. Agents Chemother.* **47**, 342–349 (2003).

124. Toleman, M. A., Bennett, P. M. & Walsh, T. R. ISCR Elements: Novel Gene-Capturing Systems of the 21st Century? *Microbiol. Mol. Biol. Rev.* **70**, 296–316 (2006).

125. Siguier, P., Gagnevin, L. & Chandler, M. The new IS1595 family, its relation to IS1 and the frontier between insertion sequences and transposons. *Res. Microbiol.* **160**, 232–241 (2009).

126. Nicolas, E. *et al.* The Tn3-family of Replicative Transposons. *Mob. DNA III* 693–726 (2015) doi:10.1128/microbiolspec.MDNA3-0060-2014.

127. Siguier, P., Filée, J. & Chandler, M. Insertion sequences in prokaryotic genomes. *Curr. Opin. Microbiol.* **9**, 526–531 (2006).

128. Delihas, N. Impact of Small Repeat Sequences on Bacterial Genome Evolution. *Genome Biol. Evol.* **3**, 959–973 (2011).

129. Bardaji, L. *et al.* Miniature Transposable Sequences Are Frequently Mobilized in the Bacterial Plant Pathogen *Pseudomonas syringae pv. phaseolicola*. *PLoS ONE* **6**, (2011).

130. Salyers, A. A., Shoemaker, N. B., Stevens, A. M. & Li, L. Y. Conjugative transposons: an unusual and diverse set of integrated gene transfer elements. *Microbiol. Rev.* **59**, 579–590 (1995).

131. Botelho, J. & Schulenburg, H. The Role of Integrative and Conjugative Elements in Antibiotic Resistance Evolution. *Trends Microbiol.* **0**, (2020).

132. Gillings, M. R. Integrons: Past, Present, and Future. *Microbiol. Mol. Biol. Rev. MMBR* **78**, 257–277 (2014).

133. Cury, J., Jové, T., Touchon, M., Néron, B. & Rocha, E. P. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res.* **44**, 4539–4550 (2016).

134. Boucher, Y. *et al.* Recovery and evolutionary analysis of complete integron gene cassette arrays from *Vibrio*. *BMC Evol. Biol.* **6**, 3 (2006).

135. Partridge, S. R., Tsafnat, G., Coiera, E. & Iredell, J. R. Gene cassettes and cassette arrays in mobile resistance integrons. *FEMS Microbiol. Rev.* **33**, 757–784 (2009).

136. Guédon, G., Libante, V., Coluzzi, C., Payot, S. & Leblond-Bourget, N. The Obscure World of Integrative and Mobilizable Elements, Highly Widespread Elements that Pirate Bacterial Conjugative Systems. *Genes* **8**, (2017).

137. Mark Osborn, A. & Böltner, D. When phage, plasmids, and transposons collide: genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum. *Plasmid* **48**, 202–212 (2002).

138. O'Brien, F. G. *et al.* Origin-of-transfer sequences facilitate mobilisation of non-conjugative antimicrobial-resistance plasmids in *Staphylococcus aureus*. *Nucleic Acids Res.* **43**, 7971–7983 (2015).

139. Roberts, A. P. & Kreth, J. The impact of horizontal gene transfer on the adaptive ability of the human oral microbiome. *Front. Cell. Infect. Microbiol.* **4**, (2014).

140. Ramsay, J. P. & Firth, N. Diverse mobilization strategies facilitate transfer of non-conjugative mobile genetic elements. *Curr. Opin. Microbiol.* **38**, 1–9 (2017).

141. Liu, L. *et al.* The human microbiome: A hot spot of microbial horizontal gene transfer. *Genomics* **100**, 265–270 (2012).

142. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241–244 (2011).

143. Martínez, J. L., Coque, T. M. & Baquero, F. What is a resistance gene? Ranking risk in resistomes. *Nat. Rev. Microbiol.* **13**, 116–123 (2015).

144. Bengtsson-Palme, J. & Larsson, D. G. J. Antibiotic resistance genes in the environment: prioritizing risks. *Nat. Rev. Microbiol.* **13**, 396–396 (2015).

145. Skippington, E. & Ragan, M. A. Lateral genetic transfer and the construction of genetic exchange communities. *FEMS Microbiol. Rev.* **35**, 707–735 (2011).

146. Papkou, A., Hedge, J., Kapel, N., Young, B. & MacLean, R. C. Efflux pump activity potentiates the evolution of antibiotic resistance across *S. aureus* isolates. *Nat. Commun.* **11**, 3970 (2020).

147. Martínez, J. L. & Rojo, F. Metabolic regulation of antibiotic resistance. *FEMS Microbiol. Rev.* **35**, 768–789 (2011).

148. Andersson, D. I. & Hughes, D. Antibiotic resistance and its cost: is it possible to reverse resistance? *Nat. Rev. Microbiol.* **8**, 260–271 (2010).

149. Baquero, F., Tedim, A. P. & Coque, T. M. Antibiotic resistance shaping multi-level population biology of bacteria. *Front. Microbiol.* **4**, (2013).

150. Ashbolt Nicholas J. *et al.* Human Health Risk Assessment (HHRA) for Environmental Development and Transfer of Antibiotic Resistance. *Environ. Health Perspect.* **121**, 993–1001 (2013).

151. Baquero, F., Alvarez-Ortega, C. & Martinez, J. L. Ecology and evolution of antibiotic resistance. *Environ. Microbiol. Rep.* **1**, 469–476 (2009).

152. Wiegand, I., Hilpert, K. & Hancock, R. E. W. Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nat. Protoc.* **3**, 163–175 (2008).

153. Bauer, A. W., Kirby, W. M. M., Sherris, J. C. & Turck, M. Antibiotic Susceptibility Testing by a Standardized Single Disk Method. *Am. J. Clin. Pathol.* **45**, 493–496 (1966).

154. Espinel-Ingroff, A. & Turnidge, J. The role of epidemiological cutoff values (ECVs/ECOFFs) in antifungal susceptibility testing and interpretation for uncommon yeasts and moulds. *Rev. Iberoam. Micol.* **33**, 63–75 (2016).

155. Morrissey, I. *et al.* Evaluation of Epidemiological Cut-Off Values Indicates that Biocide Resistant Subpopulations Are Uncommon in Natural Isolates of Clinically-Relevant Microorganisms. *PLOS ONE* **9**, e86669 (2014).

156. Sannes, M. R., Kuskowski, M. A. & Johnson, J. R. Geographical distribution of antimicrobial resistance among *Escherichia coli* causing acute uncomplicated pyelonephritis in the United States. *FEMS Immunol. Med. Microbiol.* **42**, 213–218 (2004).

157. Lee, J. Y. H. *et al.* Global spread of three multidrug-resistant lineages of *Staphylococcus epidermidis*. *Nat. Microbiol.* **3**, 1175–1185 (2018).

158. Walker, T. M. *et al.* Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect. Dis.* **15**, 1193–1202 (2015).

159. Kuehbacher, T. *et al.* Intestinal TM7 bacterial phylogenies in active inflammatory bowel disease. *J. Med. Microbiol.* **57**, 1569–1576 (2008).

160. Gijón, D., Curiao, T., Baquero, F., Coque, T. M. & Cantón, R. Fecal Carriage of Carbapenemase-Producing *Enterobacteriaceae*: a Hidden Reservoir in Hospitalized and Nonhospitalized Patients. *J. Clin. Microbiol.* **50**, 1558–1563 (2012).

161. Geser, N., Stephan, R., Korczak, B. M., Beutin, L. & Hächler, H. Molecular identification of extended-spectrum-β-lactamase genes from *Enterobacteriaceae* isolated from healthy human carriers in Switzerland. *Antimicrob. Agents Chemother.* **56**, 1609–1612 (2012).

162. Severin, J. A. *et al.* Faecal carriage of extended-spectrum β-lactamase-producing *Enterobacteriaceae* among humans in Java, Indonesia, in 2001–2002. *Trop. Med. Int. Health* **17**, 455–461 (2012).

163. Vo, A. T. T., Duijkeren, E. van, Gaastra, W. & Fluit, A. C. Antimicrobial Resistance, Class 1 Integrons, and Genomic Island 1 in *Salmonella* Isolates from Vietnam. *PLOS ONE* **5**, e9440 (2010).

164. Cuthbertson, L. *et al.* Time between Collection and Storage Significantly Influences Bacterial Sequence Composition in Sputum Samples from Cystic Fibrosis Respiratory Infections. *J. Clin. Microbiol.* **52**, 3011–3016 (2014).

165. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).

166. Wesolowska-Andersen, A. *et al.* Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. *Microbiome* **2**, 19 (2014).

167. Bag, S. *et al.* An Improved Method for High Quality Metagenomics DNA Extraction from Human and Environmental Samples. *Sci. Rep.* **6**, 26775 (2016).

168. Vesty, A., Biswas, K., Taylor, M. W., Gear, K. & Douglas, R. G. Evaluating the Impact of DNA Extraction Method on the Representation of Human Oral Bacterial and Fungal Communities. *PLOS ONE* **12**, e0169877 (2017).

169. Gray, M. W., Sankoff, D. & Cedergren, R. J. On the evolutionary descent of organisms and organelles: a global phylogeny based on a highly conserved structural core in small subunit ribosomal RNA. *Nucleic Acids Res.* **12**, 5837–5852 (1984).

170. Yang, B., Wang, Y. & Qian, P.-Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics* **17**, (2016).

171. Bartram, A. K., Lynch, M. D. J., Stearns, J. C., Moreno-Hagelsieb, G. & Neufeld, J. D. Generation of Multimillion-Sequence 16S rRNA Gene Libraries from Complex Microbial Communities by Assembling Paired-End Illumina Reads. *Appl. Environ. Microbiol.* **77**, 3846–3852 (2011).

172. Burke, C. M. & Darling, A. E. A method for high precision sequencing of near full-length 16S rRNA genes on an Illumina MiSeq. *PeerJ* **4**, (2016).

173. Schoch, C. L. *et al.* Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 6241–6246 (2012).

174. Nilsson, R. H. *et al.* Mycobiome diversity: high-throughput sequencing and identification of fungi. *Nat. Rev. Microbiol.* **17**, 95–109 (2019).

175. Callahan, B. J. *et al.* High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Res.* **47**, e103–e103 (2019).

176. Langille, M. G. I. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* **31**, 814–821 (2013).

177. Tessler, M. *et al.* Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Sci. Rep.* **7**, 6589 (2017).

178. Sunagawa, S. *et al.* Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods* **10**, 1196–1199 (2013).

179. Pehrsson, E. C., Forsberg, K. J., Gibson, M. K., Ahmadi, S. & Dantas, G. Novel resistance functions uncovered using functional metagenomic investigations of resistance reservoirs. *Front. Microbiol.* **4**, (2013).

180. Gibson, M. K. *et al.* Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. *Nat. Microbiol.* **1**, 16024 (2016).

181. Penders, J., Stobberingh, E. E., Savelkoul, P. H. M. & Wolffs, P. The human microbiome as a reservoir of antimicrobial resistance. *Front. Microbiol.* **4**, (2013).

182. Lam, K. N., Cheng, J., Engel, K., Neufeld, J. D. & Charles, T. C. Current and future resources for functional metagenomics. *Front. Microbiol.* **6**, (2015).

183. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463 (1977).

184. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24**, 133–141 (2008).

185. Check Hayden, E. Genome sequencing: the third generation. *Nature* (2009) doi:10.1038/news.2009.86.

186. Quail, M. A. *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, 341 (2012).

187. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13**, 278–289 (2015).

188. Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323**, 133–138 (2009).

189. Suzuki, Y. *et al.* Long-read metagenomic exploration of extrachromosomal mobile genetic elements in the human gut. *Microbiome* **7**, 119 (2019).

190. Tyler, A. D. *et al.* Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. *Sci. Rep.* **8**, 10931 (2018).

191. Taxt, A. M., Avershina, E., Frye, S. A., Naseer, U. & Ahmad, R. Rapid identification of pathogens, antibiotic resistance genes and plasmids in blood cultures by nanopore sequencing. *Sci. Rep.* **10**, 7622 (2020).

192. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).

193. Charalampous, T. *et al.* Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection. *Nat. Biotechnol.* **37**, 783–792 (2019).

194. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinforma. Oxf. Engl.* **32**, 3047–3048 (2016).

195. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

196. Ruppé, E. *et al.* Prediction of the intestinal resistome by a three-dimensional structure-based method. *Nat. Microbiol.* **4**, 112 (2019).

197. McArthur, A. G. *et al.* The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother.* **57**, 3348–3357 (2013).

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

199. Gupta, S. K. *et al.* ARG-ANNOT, a New Bioinformatic Tool To Discover Antibiotic Resistance Genes in Bacterial Genomes. *Antimicrob. Agents Chemother.* **58**, 212–220 (2014).

200. Lakin, S. M. *et al.* MEGARes: an antimicrobial resistance database for high throughput sequencing. *Nucleic Acids Res.* **45**, D574–D580 (2017).

201. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **46**, D8–D13 (2018).

202. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25**, 1754–1760 (2009).

203. Clausen, P. T. L. C., Aarestrup, F. M. & Lund, O. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics* **19**, 307 (2018).

204. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

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

206. 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).

207. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

208. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59 (2014).

209. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).

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

211. Brown, G. D. *et al.* Hidden Killers: Human Fungal Infections. *Sci. Transl. Med.* **4**, 165rv13-165rv13 (2012).

212. Mukherjee, P. K. *et al.* Oral Mycobiome Analysis of HIV-Infected Patients: Identification of Pichia as an Antagonist of Opportunistic Fungi. *PLOS Pathog.* **10**, e1003996 (2014).

213. Nash, A. *et al.* MARDy: Mycology Antifungal Resistance Database. *Bioinforma. Oxf. Engl.* **34**, 3233–3234 (2018).

214. Bromley, M. J. *et al.* Occurrence of azole-resistant species of *Aspergillus* in the UK environment. *J. Glob. Antimicrob. Resist.* **2**, 276–279 (2014).

215. Břinda, K. *et al.* Rapid inference of antibiotic resistance and susceptibility by genomic neighbour typing. *Nat. Microbiol.* 1–10 (2020) doi:10.1038/s41564-019-0656-6.

216. Wang, Y. *et al.* Integrated metagenomic and metatranscriptomic profiling reveals differentially expressed resistomes in human, chicken, and pig gut microbiomes. *Environ. Int.* **138**, 105649 (2020).

217. Carr, V. R., Shkoporov, A., Hill, C., Mullany, P. & Moyes, D. L. Probing the Mobilome: Discoveries in the Dynamic Microbiome. *Trends Microbiol.* **0**, (2020).

218. Sitaraman, R. Prokaryotic horizontal gene transfer within the human holobiont: ecological-evolutionary inferences, implications and possibilities. *Microbiome* **6**, 163 (2018).

219. Hsu, B. B. *et al.* Dynamic Modulation of the Gut Microbiota and Metabolome by Bacteriophages in a Mouse Model. *Cell Host Microbe* **25**, 803-814.e5 (2019).

220. Bakkeren, E. *et al. Salmonella* persisters promote the spread of antibiotic resistance plasmids in the gut. *Nature* **573**, 276–280 (2019).

221. Kleiner, M., Hooper, L. V. & Duerkop, B. A. Evaluation of methods to purify virus-like particles for metagenomic sequencing of intestinal viromes. *BMC Genomics* **16**, 7 (2015).

222. Conceição-Neto, N. *et al.* Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci. Rep.* **5**, 16532 (2015).

223. Shkoporov, A. N. *et al.* Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome* **6**, 68 (2018).

224. Milani, C. *et al.* Tracing mother-infant transmission of bacteriophages by means of a novel analytical tool for shotgun metagenomic datasets: METAnnotatorX. *Microbiome* **6**, 145 (2018).

225. Aggarwala, V., Liang, G. & Bushman, F. D. Viral communities of the human gut: metagenomic analysis of composition and dynamics. *Mob. DNA* **8**, 12 (2017).

226. Jones, B. V. & Marchesi, J. R. Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome. *Nat. Methods* **4**, 55–61 (2007).

227. Smalla, K., Jechalke, S. & Top, E. M. Plasmid Detection, Characterization, and Ecology. *Microbiol. Spectr.* **3**, (2015).

228. Solden, L. M. *et al.* Interspecies cross-feeding orchestrates carbon degradation in the rumen ecosystem. *Nat. Microbiol.* **3**, 1274 (2018).

229. Dib, J. R., Wagenknecht, M., Farías, M. E. & Meinhardt, F. Strategies and approaches in plasmidome studies—uncovering plasmid diversity disregarding of linear elements? *Front. Microbiol.* **6**, (2015).

230. Jørgensen, T. S., Xu, Z., Hansen, M. A., Sørensen, S. J. & Hansen, L. H. Hundreds of Circular Novel Plasmids and DNA Elements Identified in a Rat Cecum Metamobilome. *PLOS ONE* **9**, e87924 (2014).

231. Tansirichaiya, S., Mullany, P. & Roberts, A. P. PCR-based detection of composite transposons and translocatable units from oral metagenomic DNA. *FEMS Microbiol. Lett.* **363**, (2016).

232. Ghai, R., Mehrshad, M., Mizuno, C. M. & Rodriguez-Valera, F. Metagenomic recovery of phage genomes of uncultured freshwater actinobacteria. *ISME J.* **11**, 304–308 (2017).

233. Waller, A. S. *et al.* Classification and quantification of bacteriophage taxa in human gut metagenomes. *ISME J.* **8**, 1391–1402 (2014).

234. Ogilvie, L. A. *et al.* Genome signature-based dissection of human gut metagenomes to extract subliminal viral sequences. *Nat. Commun.* **4**, 2420 (2013).

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

236. Sutton, T. D. S., Clooney, A. G., Ryan, F. J., Ross, R. P. & Hill, C. Choice of assembly software has a critical impact on virome characterisation. *Microbiome* **7**, 12 (2019).

237. Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A. & Sullivan, M. B. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* **5**, e3817 (2017).

238. Krawczyk, P. S., Lipinski, L. & Dziembowski, A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.* **46**, e35 (2018).

239. Siguier, P., Gourbeyre, E. & Chandler, M. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol. Rev.* **38**, 865–891 (2014).

240. Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D. & Bushman, F. D. Hypervariable loci in the human gut virome. *Proc. Natl. Acad. Sci.* **109**, 3962–3966 (2012).

241. Manrique, P. *et al.* Healthy human gut phageome. *Proc. Natl. Acad. Sci.* **113**, 10400–10405 (2016).

242. Lima-Mendez, G., Toussaint, A. & Leplae, R. A modular view of the bacteriophage genomic space: identification of host and lifestyle marker modules. *Res. Microbiol.* **162**, 737–746 (2011).

243. Martinez-Hernandez, F. *et al.* Single-virus genomics reveals hidden cosmopolitan and abundant viruses. *Nat. Commun.* **8**, 15892 (2017).

244. Warwick-Dugdale, J. *et al.* Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. *PeerJ* **7**, e6800 (2019).

245. Beaulaurier, J. *et al.* Assembly-free single-molecule nanopore sequencing recovers complete virus genomes from natural microbial communities. *bioRxiv* 619684 (2019) doi:10.1101/619684.

246. Somerville, V. *et al.* Long read-based de novo assembly of low complex metagenome samples results in finished genomes and reveals insights into strain diversity and an active phage system. *bioRxiv* 476747 (2018) doi:10.1101/476747.

247. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* 1–8 (2019) doi:10.1038/s41587-019-0217-9.

248. Bertrand, D. *et al.* Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. *Nat. Biotechnol.* 1 (2019) doi:10.1038/s41587-019-0191-2.

249. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).

250. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745 (2016).

251. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).

252. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

253. Li, C., Jiang, Y. & Li, S. LEMON: a method to construct the local strains at horizontal gene transfer sites in gut metagenomics. *BMC Bioinformatics* **20**, 702 (2019).

254. Jiang, X., Hall, A. B., Xavier, R. J. & Alm, E. J. Comprehensive analysis of chromosomal mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. *PLOS ONE* **14**, e0223680 (2019).

255. Grazziotin, A. L., Koonin, E. V. & Kristensen, D. M. Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res.* **45**, D491–D498 (2017).

256. Wattam, A. R. *et al.* PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* **42**, D581–D591 (2014).

257. Paez-Espino, D. *et al.* IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Res.* **47**, D678–D686 (2019).

258. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**, D32-36 (2006).

259. Liu, M. *et al.* ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic Acids Res.* **47**, D660–D665 (2019).

260. Filée, J., Siguier, P. & Chandler, M. Insertion Sequence Diversity in Archaea. *Microbiol. Mol. Biol. Rev.* **71**, 121–157 (2007).

261. Mangul, S. *et al.* Systematic benchmarking of omics computational tools. *Nat. Commun.* **10**, 1–11 (2019).

262. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).

263. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69 (2017).

264. Amgarten, D., Braga, L. P. P., da Silva, A. M. & Setubal, J. C. MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. *Front. Genet.* **9**, (2018).

265. Zheng, T. *et al.* Mining, analyzing, and integrating viral signals from metagenomic data. *Microbiome* **7**, 42 (2019).

266. Tampuu, A., Bzhalava, Z., Dillner, J. & Vicente, R. ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples. *PLOS ONE* **14**, e0222271 (2019).

267. Zhou, F. & Xu, Y. cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics* **26**, 2051–2052 (2010).

268. Rozov, R. *et al.* Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics* **33**, 475–482 (2017).

269. Antipov, D., Raiko, M., Lapidus, A. & Pevzner, P. A. Plasmid detection and assembly in genomic and metagenomic data sets. *Genome Res.* **29**, 961–968 (2019).

270. Kamoun, C., Payen, T., Hua-Van, A. & Filée, J. Improving prokaryotic transposable elements identification using a combination of de novo and profile HMM methods. *BMC Genomics* **14**, 700 (2013).

271. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinforma. Oxf. Engl.* **21 Suppl 1**, i351-358 (2005).

272. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet. TIG* **16**, 276–277 (2000).

273. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

274. Roux, S., Tournayre, J., Mahul, A., Debroas, D. & Enault, F. Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics* **15**, 76 (2014).

275. Noguchi, H., Taniguchi, T. & Itoh, T. MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes. *DNA Res.* **15**, 387–396 (2008).

276. Brown Kav, A., Benhar, I. & Mizrahi, I. A method for purifying high quality and high yield plasmid DNA for metagenomic and deep sequencing approaches. *J. Microbiol. Methods* **95**, 272–279 (2013).

277. Beaulaurier, J. *et al.* Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat. Biotechnol.* **36**, 61–69 (2018).

278. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**, 289–293 (2009).

279. Beitel, C. W. *et al.* Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ* **2**, e415 (2014).

280. Stewart, R. D. *et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat. Commun.* **9**, 870 (2018).

281. Stalder, T., Press, M. O., Sullivan, S., Liachko, I. & Top, E. M. Linking the resistome and plasmidome to the microbiome. *ISME J.* 1 (2019) doi:10.1038/s41396-019-0446-4.

282. Burton, J. N., Liachko, I., Dunham, M. J. & Shendure, J. Species-Level Deconvolution of Metagenome Assemblies with Hi-C–Based Contact Probability Maps. *G3 Genes Genomes Genet.* **4**, 1339–1346 (2014).

283. Bickhart, D. *et al.* Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation. *bioRxiv* 491175 (2018) doi:10.1101/491175.

284. Džunková, M. *et al.* Defining the human gut host–phage network through single-cell viral tagging. *Nat. Microbiol.* 1–12 (2019) doi:10.1038/s41564-019-0526-2.

285. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 (2013).

286. Herath, D., Tang, S.-L., Tandon, K., Ackland, D. & Halgamuge, S. K. CoMet: a workflow using contig coverage and composition for binning a metagenomic sample with high precision. *BMC Bioinformatics* **18**, 571 (2017).

287. Girotto, S., Pizzi, C. & Comin, M. MetaProb: accurate metagenomic reads binning based on probabilistic sequence signatures. *Bioinformatics* **32**, i567–i575 (2016).

288. Plaza Oñate, F. *et al.* MSPminer: abundance-based reconstitution of microbial pan-genomes from shotgun metagenomic data. *Bioinformatics* doi:10.1093/bioinformatics/bty830.

289. Yu, G., Jiang, Y., Wang, J., Zhang, H. & Luo, H. BMC3C: binning metagenomic contigs using codon usage, sequence composition and read coverage. *Bioinformatics* **34**, 4172–4179 (2018).

290. Wang, Z., Wang, Z., Lu, Y. Y., Sun, F. & Zhu, S. SolidBin: improving metagenome binning with semi-supervised normalized cut. *Bioinformatics* doi:10.1093/bioinformatics/btz253.

291. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).

292. Stern, A., Mick, E., Tirosh, I., Sagy, O. & Sorek, R. CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res.* **22**, 1985–1994 (2012).

293. Wang, J., Gao, Y. & Zhao, F. Phage–bacteria interaction network in human oral microbiome. *Environ. Microbiol.* **18**, 2143–2158 (2016).

294. Zhang, Q., Rho, M., Tang, H., Doak, T. G. & Ye, Y. CRISPR-Cas systems target a diverse collection of invasive mobile genetic elements in human microbiomes. *Genome Biol.* **14**, R40 (2013).

295. Gogleva, A. A., Gelfand, M. S. & Artamonova, I. I. Comparative analysis of CRISPR cassettes from the human gut metagenomic contigs. *BMC Genomics* **15**, 202 (2014).

296. Arredondo-Alonso, S., Willems, R. J., van Schaik, W. & Schürch, A. C. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb. Genomics* **3**, (2017).

297. van Schaik Willem. The human gut resistome. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20140087 (2015).

298. Roberts, A. P. & Mullany, P. Oral biofilms: a reservoir of transferable, bacterial, antimicrobial resistance. *Expert Rev. Anti Infect. Ther.* **8**, 1441–1450 (2010).

299. Yang, I., Nell, S. & Suerbaum, S. Survival in hostile territory: the microbiota of the stomach. *FEMS Microbiol. Rev.* **37**, 736–761 (2013).

300. Mark Welch, J. L., Ramírez-Puebla, S. T. & Borisy, G. G. Oral Microbiome Geography: Micron-Scale Habitat and Niche. *Cell Host Microbe* **28**, 160–168 (2020).

301. Schmidt, T. S. B. *et al.* Extensive transmission of microbes along the gastrointestinal tract. *eLife* **8**, e42693 (2019).

302. Kent, A. G., Vill, A. C., Shi, Q., Satlin, M. J. & Brito, I. L. Widespread transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial Hi-C. *Nat. Commun.* **11**, 4379 (2020).

303. Feng, J. *et al.* Antibiotic resistome in a large-scale healthy human gut microbiota deciphered by metagenomic and network analyses. *Environ. Microbiol.* **20**, 355–368 (2018).

304. Diaz-Torres, M. L. *et al.* Determining the antibiotic resistance potential of the indigenous oral microbiota of humans using a metagenomic approach. *FEMS Microbiol. Lett.* **258**, 257–262 (2006).

305. Sommer, M. O. A., Dantas, G. & Church, G. M. Functional Characterization of the Antibiotic Resistance Reservoir in the Human Microflora. *Science* **325**, 1128–1131 (2009).

306. Warburton, P. *et al.* Characterization of tet(32) Genes from the Oral Metagenome. *Antimicrob. Agents Chemother.* **53**, 273–276 (2009).

307. Carr, V. R. *et al.* Abundance and diversity of resistomes differ between healthy human oral cavities and gut. *Nat. Commun.* **11**, 693 (2020).

308. Rose, G. *et al.* Antibiotic resistance potential of the healthy preterm infant gut microbiome. *PeerJ* **5**, e2928 (2017).

309. Christoff, A. P. *et al.* One year cross-sectional study in adult and neonatal intensive care units reveals the bacterial and antimicrobial resistance genes profiles in patients and hospital surfaces. *PLOS ONE* **15**, e0234127 (2020).

310. Pärnänen, K. *et al.* Maternal gut and breast milk microbiota affect infant gut antibiotic resistome and mobile genetic elements. *Nat. Commun.* **9**, (2018).

311. Sun, J. *et al.* Environmental remodeling of human gut microbiota and antibiotic resistome in livestock farms. *Nat. Commun.* **11**, 1–11 (2020).

312. Clemente, J. C. *et al.* The microbiome of uncontacted Amerindians. *Sci. Adv.* **1**, e1500183 (2015).

313. Seville, L. A. *et al.* Distribution of Tetracycline and Erythromycin Resistance Genes Among Human Oral and Fecal Metagenomic DNA. *Microb. Drug Resist.* **15**, 159–166 (2009).

314. Gasparrini, A. J. *et al.* Persistent metagenomic signatures of early-life hospitalization and antibiotic treatment in the infant gut microbiota and resistome. *Nat. Microbiol.* 1–13 (2019) doi:10.1038/s41564-019-0550-2.

315. Zaura, E. *et al.* Same Exposure but Two Radically Different Responses to Antibiotics: Resilience of the Salivary Microbiome versus Long-Term Microbial Shifts in Feces. *mBio* **6**, e01693-15 (2015).

316. Mah, T.-F. C. & O'Toole, G. A. Mechanisms of biofilm resistance to antimicrobial agents. *Trends Microbiol.* **9**, 34–39 (2001).

317. Lee, K. *et al.* Mobile resistome of human gut and pathogen drives anthropogenic bloom of antibiotic resistance. *Microbiome* **8**, 2 (2020).

318. Che, Y. *et al.* Mobile antibiotic resistome in wastewater treatment plants revealed by Nanopore metagenomic sequencing. *Microbiome* **7**, 44 (2019).

319. Ravi, A. *et al.* The commensal infant gut meta-mobilome as a potential reservoir for persistent multidrug resistance integrons. *Sci. Rep.* **5**, 15317 (2015).

320. Bengtsson-Palme, J., Boulund, F., Fick, J., Kristiansson, E. & Larsson, D. G. J. Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in India. *Front. Microbiol.* **5**, (2014).

321. Bickhart, D. M. *et al.* Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation. *Genome Biol.* **20**, 153 (2019).

322. Modi, S. R., Lee, H. H., Spina, C. S. & Collins, J. J. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* **499**, 219–222 (2013).

323. 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**, (2018).

324. Vaga, S. *et al.* Compositional and functional differences of the mucosal microbiota along the intestine of healthy individuals. *Sci. Rep.* **10**, 14977 (2020).

325. Feng, J. *et al.* Antibiotic Resistome in a large-scale healthy human gut microbiota deciphered by metagenomic and network analyses. *Environ. Microbiol.* n/a-n/a doi:10.1111/1462-2920.14009.

326. Rahman, S. F., Olm, M. R., Morowitz, M. J. & Banfield, J. F. Machine Learning Leveraging Genomes from Metagenomes Identifies Influential Antibiotic Resistance Genes in the Infant Gut Microbiome. *mSystems* **3**, e00123-17 (2018).

327. Hu, Y. *et al.* Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nat. Commun.* **4**, 2151 (2013).

328. Seville, L. A. *et al.* Distribution of tetracycline and erythromycin resistance genes among human oral and fecal metagenomic DNA. *Microb. Drug Resist. Larchmt. N* **15**, 159–166 (2009).

329. Warinner, C. *et al.* Pathogens and host immunity in the ancient human oral cavity. *Nat. Genet.* **46**, 336–344 (2014).

330. Knoll, B., Tleyjeh, I. M., Steckelberg, J. M., Wilson, W. R. & Baddour, L. M. Infective Endocarditis Due to Penicillin-Resistant Viridans Group Streptococci. *Clin. Infect. Dis.* **44**, 1585–1592 (2007).

331. Zhang, X. *et al.* The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat. Med.* **21**, 895–905 (2015).

332. Brito, I. L. *et al.* Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **535**, 435–439 (2016).

333. Lassalle, F. *et al.* Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet. *Mol. Ecol.* **27**, 182–195 (2018).

334. Voigt, A. Y. *et al.* Temporal and technical variability of human gut metagenomes. *Genome Biol.* **16**, 73 (2015).

335. Zeller, G. *et al.* Potential of fecal microbiota for early‐stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).

336. Human Microbiome Project Consortium. A framework for human microbiome research. *Nature* **486**, 215–221 (2012).

337. Thomas, M. *et al.* Metagenomic characterization of the effect of feed additives on the gut microbiome and antibiotic resistome of feedlot cattle. *Sci. Rep.* **7**, 12257 (2017).

338. Ma, L. *et al.* Catalogue of antibiotic resistome and host-tracking in drinking water deciphered by a large scale survey. *Microbiome* **5**, 154 (2017).

339. Noyes, N. R. *et al.* Characterization of the resistome in manure, soil and wastewater from dairy and beef production systems. *Sci. Rep.* **6**, 24645 (2016).

340. Witherden, E. A., Bajanca-Lavado, M. P., Tristram, S. G. & Nunes, A. Role of inter-species recombination of the ftsI gene in the dissemination of altered

penicillin-binding-protein-3-mediated resistance in *Haemophilus influenzae* and *Haemophilus haemolyticus*. *J. Antimicrob. Chemother.* **69**, 1501–1509 (2014).

341. Chaffanel, F., Charron-Bourgoin, F., Libante, V., Leblond-Bourget, N. & Payot, S. Resistance Genes and Genetic Elements Associated with Antibiotic Resistance in Clinical and Commensal Isolates of *Streptococcus salivarius*. *Appl. Environ. Microbiol.* **81**, 4155–4163 (2015).
342. Criscuolo, A. & Brisse, S. AlienTrimmer: A tool to quickly and accurately trim off multiple short contaminant sequences from high-throughput sequencing reads. *Genomics* **102**, 500–506 (2013).
343. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma. Oxf. Engl.* **25**, 2078–2079 (2009).
344. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma. Oxf. Engl.* **26**, 841–842 (2010).
345. Jonsson, V., Österlund, T., Nerman, O. & Kristiansson, E. Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. *BMC Genomics* **17**, 78 (2016).
346. Viechtbauer, W. Conducting Meta-Analyses in R with the metafor Package. *J. Stat. Softw.* **36**, 1–48 (2010).
347. Bengtsson-Palme, J., Larsson, D. G. J. & Kristiansson, E. Using metagenomics to investigate human and environmental resistomes. *J. Antimicrob. Chemother.* **72**, 2690–2703 (2017).
348. Bengtsson-Palme, J. The diversity of uncharacterized antibiotic resistance genes can be predicted from known gene variants—but not always. *Microbiome* **6**, 125 (2018).
349. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
350. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).

351. Chang, J. *et al.* Sale of antibiotics without a prescription at community pharmacies in urban China: a multicentre cross-sectional survey. *J. Antimicrob. Chemother.* **72**, 1235–1242 (2017).

352. Auta, A. *et al.* Global access to antibiotics without prescription in community pharmacies: A systematic review and meta-analysis. *J. Infect.* (2018) doi:10.1016/j.jinf.2018.07.001.

353. Liu, X., Steele, J. C. & Meng, X.-Z. Usage, residue, and human health risk of antibiotics in Chinese aquaculture: A review. *Environ. Pollut. Barking Essex 1987* **223**, 161–169 (2017).

354. Stanton, T. B. A call for antibiotic alternatives research. *Trends Microbiol.* **21**, 111–113 (2013).

355. Hendriksen, R. S. *et al.* Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nat. Commun.* **10**, 1124 (2019).

356. Roldán, S., Herrera, D. & Sanz, M. Biofilms and the tongue: therapeutical approaches for the control of halitosis. *Clin. Oral Investig.* **7**, 189–197 (2003).

357. Piddock, L. J. V. Assess drug-resistance phenotypes, not just genotypes. *Nat. Microbiol.* **1**, 16120 (2016).

358. Greub, G. Culturomics: a new approach to study the human microbiome. *Clin. Microbiol. Infect.* **18**, 1157–1159 (2012).

359. Mullany, P. Functional metagenomics for the investigation of antibiotic resistance. *Virulence* **5**, 443–447 (2014).

360. Lanza, V. F. *et al.* In-depth resistome analysis by targeted metagenomics. *Microbiome* **6**, 11 (2018).

361. Tang, Q. *et al.* Current Sampling Methods for Gut Microbiota: A Call for More Precise Devices. *Front. Cell. Infect. Microbiol.* **10**, (2020).

362. Bendall, M. L. *et al.* Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J.* **10**, 1589–1601 (2016).

363. Leggett, R. M. & MacLean, D. Reference-free SNP detection: dealing with the data deluge. *BMC Genomics* **15**, S10 (2014).

364. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44**, 226–232 (2012).

365. Leggett, R. M. *et al.* Identifying and Classifying Trait Linked Polymorphisms in Non-Reference Species by Walking Coloured de Bruijn Graphs. *PLoS ONE* **8**, (2013).

366. Nijkamp, J. F., Pop, M., Reinders, M. J. T. & de Ridder, D. Exploring variation-aware contig graphs for (comparative) metagenomics using MaryGold. *Bioinformatics* **29**, 2826–2834 (2013).

367. Fernández, L., Rodríguez, A. & García, P. Phage or foe: an insight into the impact of viral predation on microbial communities. *ISME J.* 1 (2018) doi:10.1038/s41396-018-0049-5.

368. Clooney, A. G. *et al.* Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. *Cell Host Microbe* **26**, 764-778.e5 (2019).

369. Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4306.

370. Koonin, E. V. & Yutin, N. The crAss-like Phage Group: How Metagenomics Reshaped the Human Virome. *Trends Microbiol.* **28**, 349–359 (2020).

371. Yutin, N. *et al.* Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat. Microbiol.* **3**, 38–46 (2018).

372. Edwards, R. A. *et al.* Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat. Microbiol.* **4**, 1727–1736 (2019).

373. Yuan, Y. & Gao, M. Jumbo Bacteriophages: An Overview. *Front. Microbiol.* **8**, (2017).

374. Devoto, A. E. *et al.* Megaphages infect *Prevotella* and variants are widespread in gut microbiomes. *Nat. Microbiol.* 1 (2019) doi:10.1038/s41564-018-0338-9.

375. Pride, D. T. *et al.* Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. *ISME J.* **6**, 915–926 (2012).

376. Wang, J., Gao, Y. & Zhao, F. Phage–bacteria interaction network in human oral microbiome. *Environ. Microbiol.* **18**, 2143–2158 (2016).

377. Hayes, S., Mahony, J., Nauta, A. & van Sinderen, D. Metagenomic Approaches to Assess Bacteriophages in Various Environmental Niches. *Viruses* **9**, 127 (2017).

378. Jang, H. B. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).

379. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **36**, D281–D288 (2008).

380. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).

381. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32**, 11–16 (2004).

382. Fernandes, A. P. & Holmgren, A. Glutaredoxins: Glutathione-Dependent Redox Enzymes with Functions Far Beyond a Simple Thioredoxin Backup System. *Antioxid. Redox Signal.* **6**, 63–74 (2004).

383. Rands, C. M. *et al.* ACI-1 beta-lactamase is widespread across human gut microbiomes in *Negativicutes* due to transposons harboured by tailed prophages. *Environ. Microbiol.* **20**, 2288–2300 (2018).

384. Abril, C., Brodard, I. & Perreten, V. Two novel antibiotic resistance genes, tet(44) and ant(6)-Ib, are located within a transferable pathogenicity island in *Campylobacter fetus* subsp. fetus. *Antimicrob. Agents Chemother.* **54**, 3052–3055 (2010).

385. Achard, A., Villers, C., Pichereau, V. & Leclercq, R. New lnu(C) gene conferring resistance to lincomycin by nucleotidylation in *Streptococcus agalactiae* UCN36. *Antimicrob. Agents Chemother.* **49**, 2716–2719 (2005).

386. Li, B. *et al.* Detection and new genetic environment of the pleuromutilin-lincosamide-streptogramin A resistance gene lsa(E) in methicillin-resistant *Staphylococcus aureus* of swine origin. *J. Antimicrob. Chemother.* **68**, 1251–1255 (2013).

387. Kim, H. B. *et al.* oqxAB encoding a multidrug efflux pump in human clinical isolates of *Enterobacteriaceae*. *Antimicrob. Agents Chemother.* **53**, 3582–3584 (2009).

388. Novotna, G. & Janata, J. A New Evolutionary Variant of the Streptogramin A Resistance Protein, Vga(A)LC, from *Staphylococcus haemolyticus* with Shifted Substrate Specificity towards Lincosamides. *Antimicrob. Agents Chemother.* **50**, 4070–4076 (2006).

389. Al-Shayeb, B. *et al.* Clades of huge phages from across Earth's ecosystems. *Nature* 1–7 (2020) doi:10.1038/s41586-020-2007-4.

390. Romsang, A., Leesukon, P., Duangnkern, J., Vattanaviboon, P. & Mongkolsuk, S. Mutation of the gene encoding monothiol glutaredoxin (GrxD) in *Pseudomonas aeruginosa* increases its susceptibility to polymyxins. *Int. J. Antimicrob. Agents* **45**, 314–318 (2015).

391. Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol. Rev.* **40**, 258–272 (2016).

392. Hansen, M. F., Svenningsen, S. L., Røder, H. L., Middelboe, M. & Burmølle, M. Big Impact of the Tiny: Bacteriophage–Bacteria Interactions in Biofilms. *Trends Microbiol.* **0**, (2019).

393. Allaband, C. *et al.* Microbiome 101: Studying, Analyzing, and Interpreting Gut Microbiome Data for Clinicians. *Clin. Gastroenterol. Hepatol. Off. Clin. Pract. J. Am. Gastroenterol. Assoc.* **17**, 218–230 (2019).

394. Lepage, P. *et al.* Dysbiosis in inflammatory bowel disease: a role for bacteriophages? *Gut* **57**, 424–425 (2008).

395. Hoyles, L. *et al.* Characterization of virus-like particles associated with the human faecal and caecal microbiota. *Res. Microbiol.* **165**, 803–812 (2014).

396. Knowles, B. *et al.* Lytic to temperate switching of viral communities. *Nature* **531**, 466–470 (2016).

397. Thingstad, T. F. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol. Oceanogr.* **45**, 1320–1328 (2000).

398. Jenkinson, H. F. & Lamont, R. J. Oral microbial communities in sickness and in health. *Trends Microbiol.* **13**, 589–595 (2005).

399. Paula, M. O., Gaetti-Jardim Júnior, E. & Avila-Campos, M. J. Plasmid profile in oral *Fusobacterium nucleatum* from humans and Cebus apella monkeys. *Rev. Inst. Med. Trop. Sao Paulo* **45**, 5–9 (2003).

400. Arredondo-Alonso, S., Willems, R. J., van Schaik, W. & Schürch, A. C. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb. Genomics* **3**, e000128 (2017).

401. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

402. Ares-Arroyo, M. *et al.* PCR-Based Analysis of ColE1 Plasmids in Clinical Isolates and Metagenomic Samples Reveals Their Importance as Gene Capture Platforms. *Front. Microbiol.* **9**, (2018).

403. Chen, C.-Y., Lindsey, R. L., Strobaugh, T. P., Frye, J. G. & Meinersmann, R. J. Prevalence of ColE1-Like Plasmids and Kanamycin Resistance Genes in *Salmonella enterica* Serovars. *Appl. Environ. Microbiol.* **76**, 6707–6714 (2010).

404. Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**, D517–D525 (2020).

405. Clark, D. P. & Pazdernik, N. J. *Molecular Biology*. vol. Second Edition (Elsevier, 2013).

406. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2012).

407. Khan, Z., Bloom, J. S., Kruglyak, L. & Singh, M. A practical algorithm for finding maximal exact matches in large sequence datasets using sparse suffix arrays. *Bioinformatics* **25**, 1609–1616 (2009).

408. Myers, E. W. *et al.* A Whole-Genome Assembly of Drosophila. *Science* **287**, 2196–2204 (2000).

409. Khiste, N. & Ilie, L. E-MEM: efficient computation of maximal exact matches for very large genomes. *Bioinformatics* **31**, 509–514 (2015).

410. Fernandes, F. & Freitas, A. T. slaMEM: efficient retrieval of maximal exact matches using a sampled LCP array. *Bioinformatics* **30**, 464–471 (2014).

411. Vyverman, M., De Baets, B., Fack, V. & Dawyndt, P. essaMEM: finding maximal exact matches using enhanced sparse suffix arrays. *Bioinformatics* **29**, 802–804 (2013).

412. Tommaso, P. D. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316 (2017).

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

414. Liu, Y. *et al.* Structural Insights Into the Transcriptional Regulation of HigBA Toxin–Antitoxin System by Antitoxin HigA in Pseudomonas aeruginosa. *Front. Microbiol.* **10**, (2020).

415. Achaz, G., Coissac, E., Netter, P. & Rocha, E. P. C. Associations between inverted repeats and the structural evolution of bacterial genomes. *Genetics* **164**, 1279–1289 (2003).

416. Lin, C.-T., Lin, W.-H., Lyu, Y. L. & Whang-Peng, J. Inverted repeats as genetic elements for promoting DNA inverted duplication: implications in gene amplification. *Nucleic Acids Res.* **29**, 3529–3538 (2001).

417. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.* 1–10 (2020) doi:10.1038/s41587-020-0603-3.

418. Dewhirst, F. E. *et al.* The Human Oral Microbiome. *J. Bacteriol.* **192**, 5002–5017 (2010).

419. Fritz, A. *et al.* CAMISIM: simulating metagenomes and microbial communities. *Microbiome* **7**, 17 (2019).

420. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).

421. Tierney, B. T. *et al.* The Landscape of Genetic Content in the Gut and Oral Human Microbiome. *Cell Host Microbe* **26**, 283-295.e8 (2019).

422. Andersson, D. I. & Hughes, D. Microbiological effects of sublethal levels of antibiotics. *Nat. Rev. Microbiol.* **12**, 465–478 (2014).

423. Liu, Y. *et al.* Correlation between Exogenous Compounds and the Horizontal Transfer of Plasmid-Borne Antibiotic Resistance Genes. *Microorganisms* **8**, 1211 (2020).

424. Møller, T. S. B. *et al.* Treatment with Cefotaxime Affects Expression of Conjugation Associated Proteins and Conjugation Transfer Frequency of an IncI1 Plasmid in Escherichia coli. *Front. Microbiol.* **8**, (2017).

425. Leclerc, Q. J., Lindsay, J. A. & Knight, G. M. Mathematical modelling to study the horizontal transfer of antimicrobial resistance genes in bacteria: current state of the field and recommendations. *J. R. Soc. Interface* **16**, 20190260 (2019).

426. Lopatkin, A. J. *et al.* Antibiotics as a selective driver for conjugation dynamics. *Nat. Microbiol.* **1**, 1–8 (2016).

427. Johnsen, P. J., Dubnau, D. & Levin, B. R. Episodic Selection and the Maintenance of Competence and Natural Transformation in *Bacillus subtilis*. *Genetics* **181**, 1521–1533 (2009).

428. Lindsay, J. A. *Staphylococcus aureus* genomics and the impact of horizontal gene transfer. *Int. J. Med. Microbiol. IJMM* **304**, 103–109 (2014).

429. Hamilton, H. L. & Dillard, J. P. Natural transformation of *Neisseria gonorrhoeae*: from DNA donation to homologous recombination. *Mol. Microbiol.* **59**, 376–385 (2006).

430. Monod, M., Mohan, S. & Dubnau, D. Cloning and analysis of ermG, a new macrolide-lincosamide-streptogramin B resistance element from *Bacillus sphaericus*. *J. Bacteriol.* **169**, 340–350 (1987).

431. Shoemaker, N. B., Vlamakis, H., Hayes, K. & Salyers, A. A. Evidence for Extensive Resistance Gene Transfer among Bacteroides spp. and among Bacteroides and Other Genera in the Human Colon. *Appl. Environ. Microbiol.* **67**, 561–568 (2001).

432. Cooper, A. J., Shoemaker, N. B. & Salyers, A. A. The erythromycin resistance gene from the *Bacteroides* conjugal transposon Tcr Emr 7853 is nearly identical to ermG from *Bacillus sphaericus*. *Antimicrob. Agents Chemother.* **40**, 506–508 (1996).

433. Notomi, T. *et al.* Loop-mediated isothermal amplification of DNA. *Nucleic Acids Res.* **28**, e63 (2000).

434. Rosini, R., Nicchi, S., Pizza, M. & Rappuoli, R. Vaccines Against Antimicrobial Resistance. *Front. Immunol.* **11**, (2020).

435. Petrovic Fabijan, A. *et al.* Safety of bacteriophage therapy in severe *Staphylococcus aureus* infection. *Nat. Microbiol.* **5**, 465–472 (2020).

436. Schooley, R. T. *et al.* Development and Use of Personalized Bacteriophage-Based Therapeutic Cocktails To Treat a Patient with a Disseminated Resistant *Acinetobacter baumannii* Infection. *Antimicrob. Agents Chemother.* **61**, (2017).

437. Digby-Bell, J., Williams, A., Irving, P. & Goldenberg, S. Successful faecal microbiota transplant for recurrent *Clostridium difficile* infection delivered by colonoscopy through a diverted ileostomy in a patient with severe perianal Crohn's disease. *Case Rep.* **2018**, bcr-2017-222958 (2018).

438. Mullish, B. H., Ghani, R., McDonald, J. a. K. & Marchesi, J. R. Faecal microbiota transplant for eradication of multidrug-resistant *Enterobacteriaceae*: a lesson in applying best practice? Re: 'A five-day course of oral antibiotics followed by faecal transplantation to eradicate carriage of multidrug-resistant *Enterobacteriaceae*: A Randomized Clinical Trial'. *Clin. Microbiol. Infect.* **25**, 912–913 (2019).

439. Citorik, R. J., Mimee, M. & Lu, T. K. Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. *Nat. Biotechnol.* **32**, 1141–1145 (2014).

440. Inouye, M. *et al.* SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med.* **6**, 90 (2014).

441. Zaheer, R. *et al.* Impact of sequencing depth on the characterization of the microbiome and resistome. *Sci. Rep.* **8**, 5890 (2018).

442. Clausen, P. T. L. C., Zankari, E., Aarestrup, F. M. & Lund, O. Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. *J. Antimicrob. Chemother.* **71**, 2484–2488 (2016).

Appendices

Appendix numbering represents the chapter number from which the Appendix is referenced. For example, Appendix 2A is from Chapter 2.

Appendix 2A: Choosing an ARG reference database

Before mapping the metagenomic data, an existing reference database of ARGs had to be chosen that was well curated and could generate a high specificity and sensitivity of ARG annotations in metagenomic data. Five known ARG databases were compared: CARD[197], ARGs from NCBI[201] and MEGARes[200], and databases from the ARG-ANNOT[199] and ResFinder[198] tools. A catalogue containing nucleotide sequences of oral genes (that was created by a collaborating research group at the Centre for Host-Microbiome Interactions, King's College London) (with filename of unique prefix *ORAL_CATALOG_2017_09*) was aligned against each ARG database using ABRicate v0.5 (downloaded on 23rd August 2017) (https://github.com/tseemann/abricate). Apart from MEGARes, all other databases were downloaded as part of the ABRicate software. The MEGARes database was downloaded separately on 8th September 2017. The ARGs that were identified from the oral gene catalogue for each ARG database were aligned again against every ARG database with the same method. This determines the overlap of ARG annotations identified from pairs of databases.

MEGAres generated the highest number of ARG annotations, followed by CARD, with fewest from ResFinder, NCBI and ARG-ANNOT **(Fig. 2A)**. Most of the ResFinder, NCBI and ARG-ANNOT hits were also annotated by MEGARes. Although many hits

were generated by mapping to MEGARes, at the time of this study MEGARes had not been updated since December 2016, whereas CARD is a curated resource that is updated almost monthly. In addition, the sequence ontologies[xxiv] of MEGARes were not well defined compared to those of CARD. For example, the CTX-M-1 ARG, has a more comprehensive description in CARD (https://card.mcmaster.ca/ontology/38264) than in the MEGARes database (https://megares.meglab.org/browse/betalactams/Class_A_betalactamases/CTX/). Specifically, the CARD entry for CTX-M-1 labels it as being part of the CTX-M β-lactamase gene family, the cephalosporin drug class and the resistance mechanism, antibiotic inactivation. In contrast, CTX-M_1 in the MEGARes database is labelled as being part of the CTX group, of a less specific β-lactams class with no cephalosporin sub-group and Class A β-lactamase mechanism with no explicitly labelled cell mechanism. Therefore, CARD was chosen as the ARG reference database of this study as it had a better sequence ontology and was regularly updated.

xxiv A sequence ontology is the system of naming features and characteristics of sequences that are consistent, machine-readable and searchable.

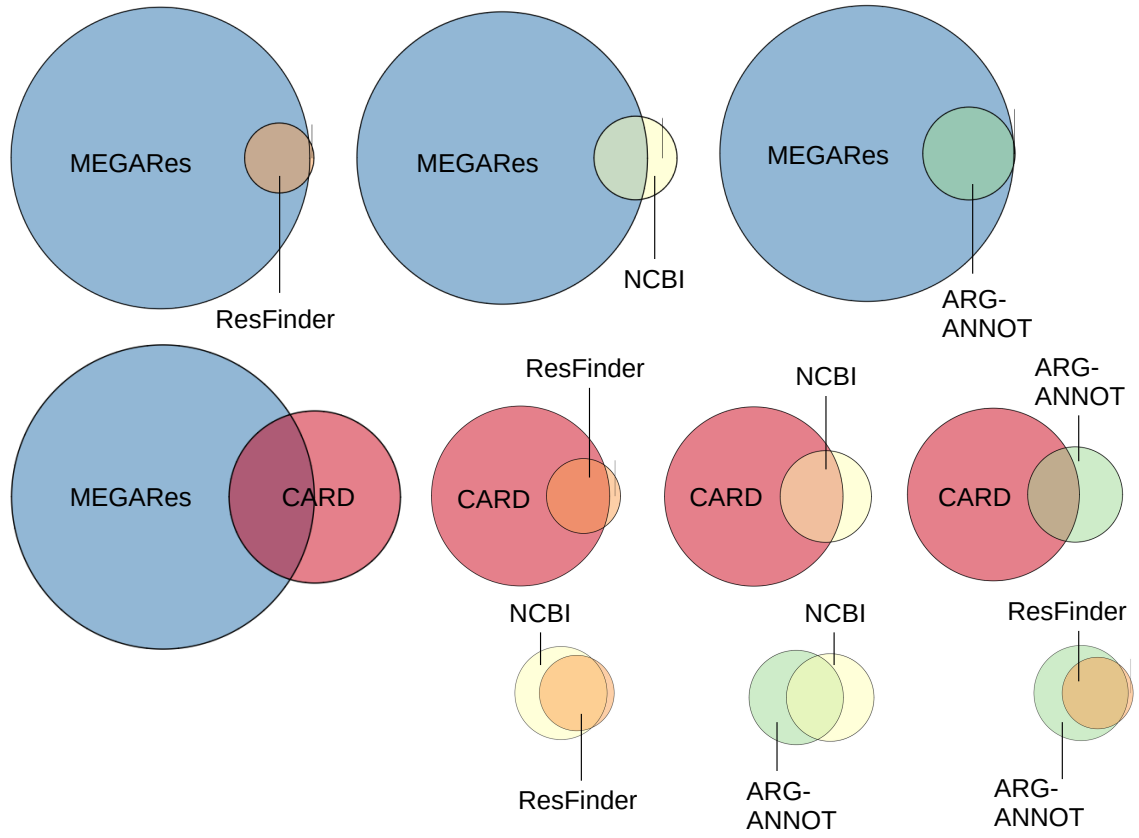

Figure 2A. Pairwise overlap of annotations from five ARG databases.

ARG-ANNOT (green: 3,729 annotations); CARD (red: 6,640 annotations); NCBI (yellow: 3,189 annotations); ResFinder (orange: 2,833 annotations); and MEGARes (blue: 13,869 annotations).

Appendix 2B: Selecting a mapping tool

To ensure that results could be replicated using different mapping software, three types of mapping software were compared: Bowtie2[195], SRST2[440] and KMA[203]. Bowtie2 is a very common tool that has been used widely in metagenomic resistome studies[310,355,441]. SRST2 is a Bowtie2-based mapping tool, not specifically built for metagenomics per se, but has been applied in at least one metagenomic resistome study[308]. KMA is a *k*-mer-based scoring algorithm that accounts for redundancy in the composition of sequences and has been recently applied in a metagenomic resistome study to circumvent such redundancy[42]. CARD contains substantial sequencing redundancy, i.e. many of its sequences are similar to each other, especially ARGs within ARG families that are variants of each other. Using CD-HIT, the nucleotide sequences from CARD were clustered in 938 clusters with greater than 90% identity (meaning at least 90% of the sequence for each read in a cluster was identical to every other read in the same cluster). One disadvantage with KMA is that it cannot calculate the raw read counts, which are required for Reads Per Kilobase Million (RPKM) normalisation. Therefore, a pipeline was used whereby ARGs were identified using KMA and then read counts were evaluated from these predicted ARGs using Bowtie2. All samples were mapped to CARD v3.0.0 with each method, Bowtie2, SRST2, KMA and KMA+Bowtie2, including a 90% breadth of coverage[xxv] threshold (See **Appendix 2C**: Selecting a coverage threshold). All methods produce similar distributions in the abundances of ARG classes **(Fig. 2B)**. The main differences in the results are that the KMA method generates a higher relative abundance of the lincosamide ARG class and lower relative

xxv The breadth of coverage is the percentage of bases of the reference sequence that are covered by reads (not to be confused with coverage depth). (Included in Glossary)

abundance of lincosamide/streptogramin/pleuromutilin ARG classes than other methods. These differences are likely a result of how the read depth is evaluated in KMA. Therefore, KMA+Bowtie2 was the chosen method in this study to account for sequence redundancy in CARD using KMA and also preserving counting of raw reads using Bowtie2.

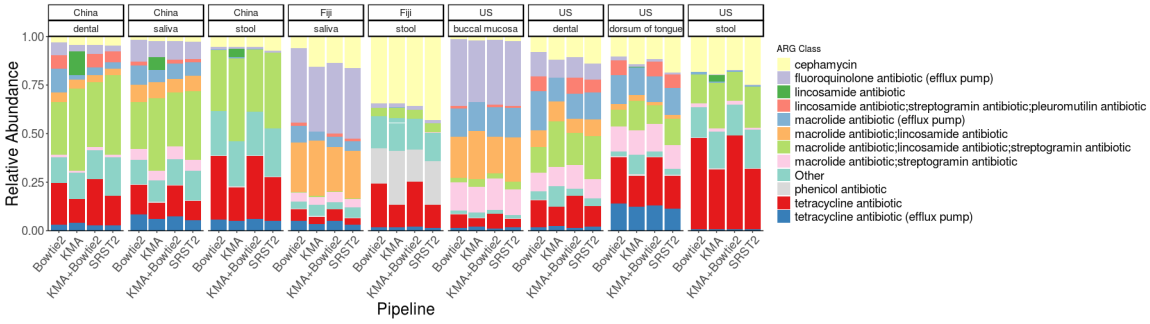

Figure 2B. Relative abundance of ARG classes using four read mapping tools.

Tools are Bowtie2, KMA, KMA+Bowtie2 and SRST2 used for all samples (n=1,174).

Appendix 2C: Selecting a breadth of coverage threshold

In this study, ARGs were filtered if they did not have above 90% breadth of coverage, which is a standard threshold used to map reads to ARG databases[442]. To verify this is an appropriate breadth of coverage threshold, coverages of ARGs from the KMA-Bowtie2 pipeline were compared against ARGs identified from assemblies of metagenomic reads. All metagenomes were assembled using SPAdes v3.9.0 including parameter --*meta* for metagenomic data. These assemblies were aligned against CARD v3.0.0 using BLASTn v2.7.1. All hits were filtered by an e-value[xxvi] ≤ 1e-50 and an identity ≥ 80. Multiple hits of different ARGs that overlapped each other by greater than 20% on an assembly are filtered to leave the hit with the lowest e-value and highest identity values. The presence or absence of ARGs in a metagenomic assembly was used as a proxy for ARG incidence in the sample. A breadth of coverage threshold of 90% was chosen because most ARGs that are found in an assembly are also identified in reads mapped to ARGs with a breadth of coverage greater than 90% **(Fig. 2C)**.

xxvi The e-value is the number of expected hits of similar quality that could be found just by chance. The smaller the e-value, the more significant the match. (Included in Glossary)

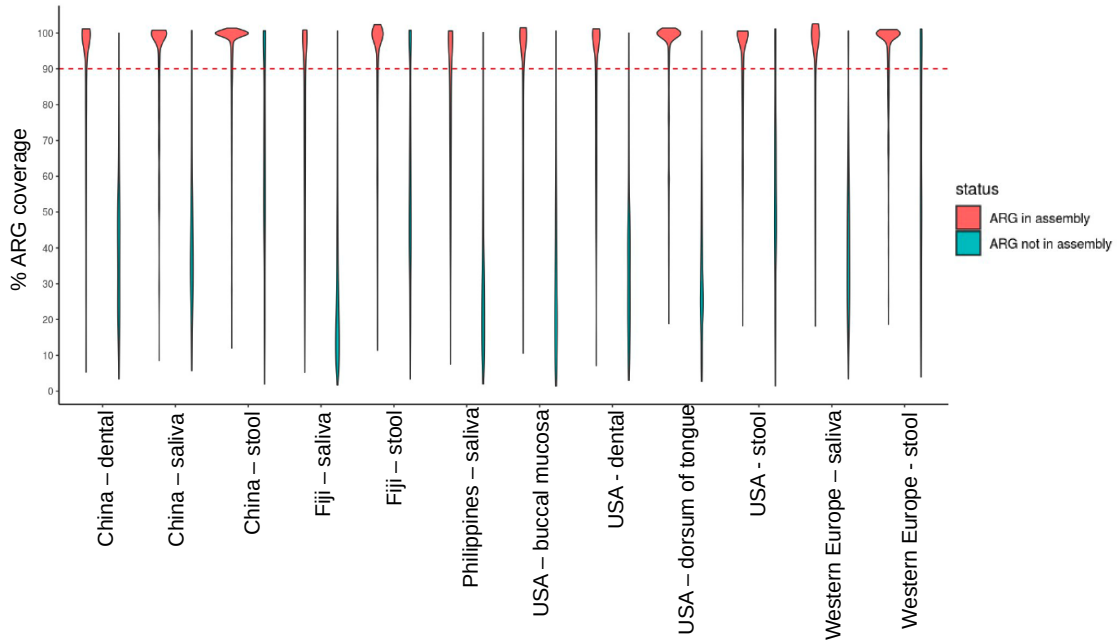

Figure 2C. Benchmarking of breadth of coverage threshold.

Showing the distribution of the percentage breadth of coverage of ARGs using the KMA-Bowtie2 pipeline for all samples (n = 1,174). Distributions are labelled by whether the equivalent ARG is present in (red) or absent (blue) from the metagenomic assemblies. Red dashed line at 90%.

Appendix 2D-M: Supplementary Materials from integrated paper

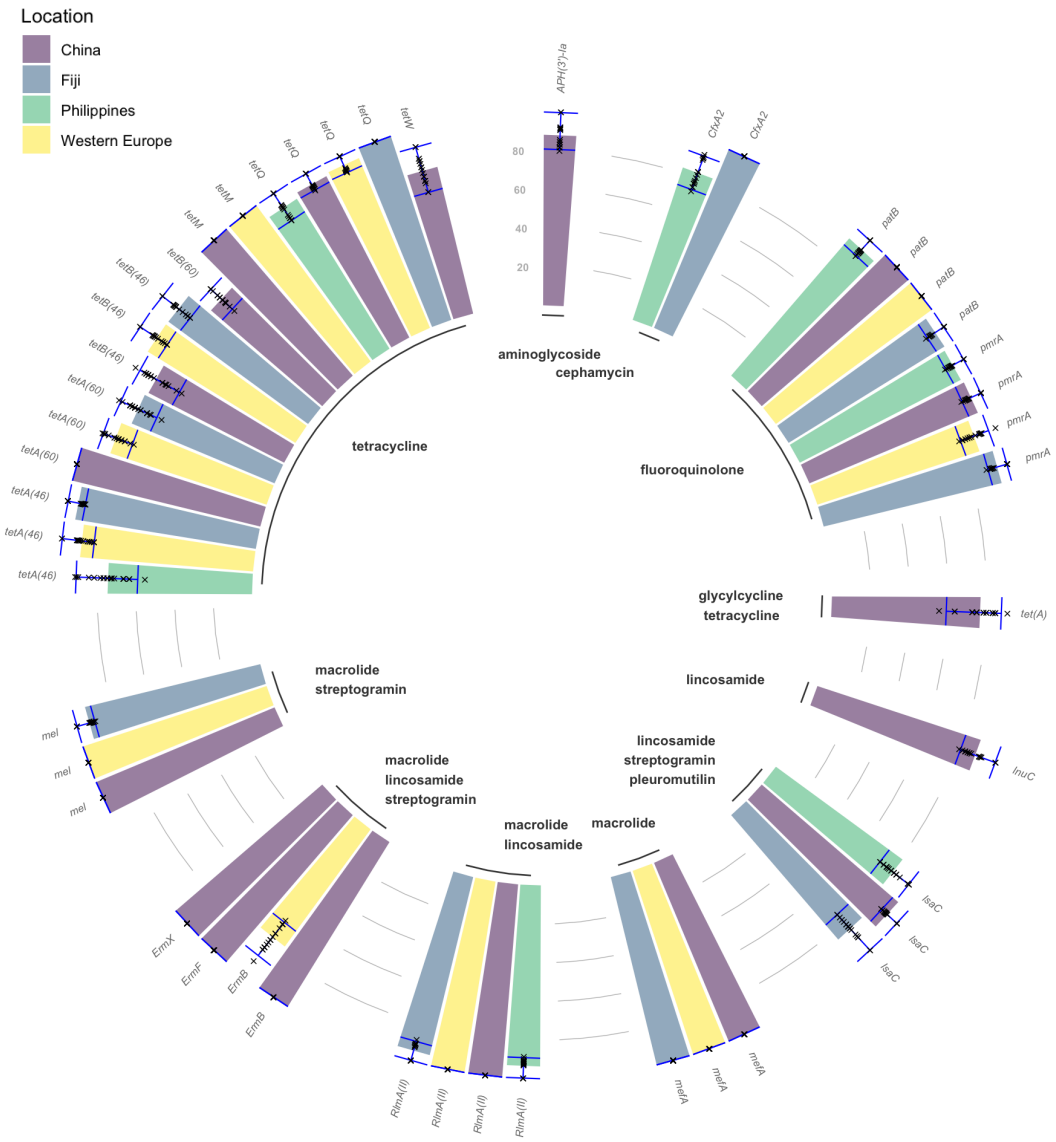

Figure 2Da. ARGs that are found ≥ 70% of saliva samples.

China (n = 18), Fiji (n = 18), the Philippines (n = 18) and Western Europe (n = 18). The height of bars are the means and the error bars are 95% CIs of percentages extracted from bootstrapping samples 20 times shown by points.

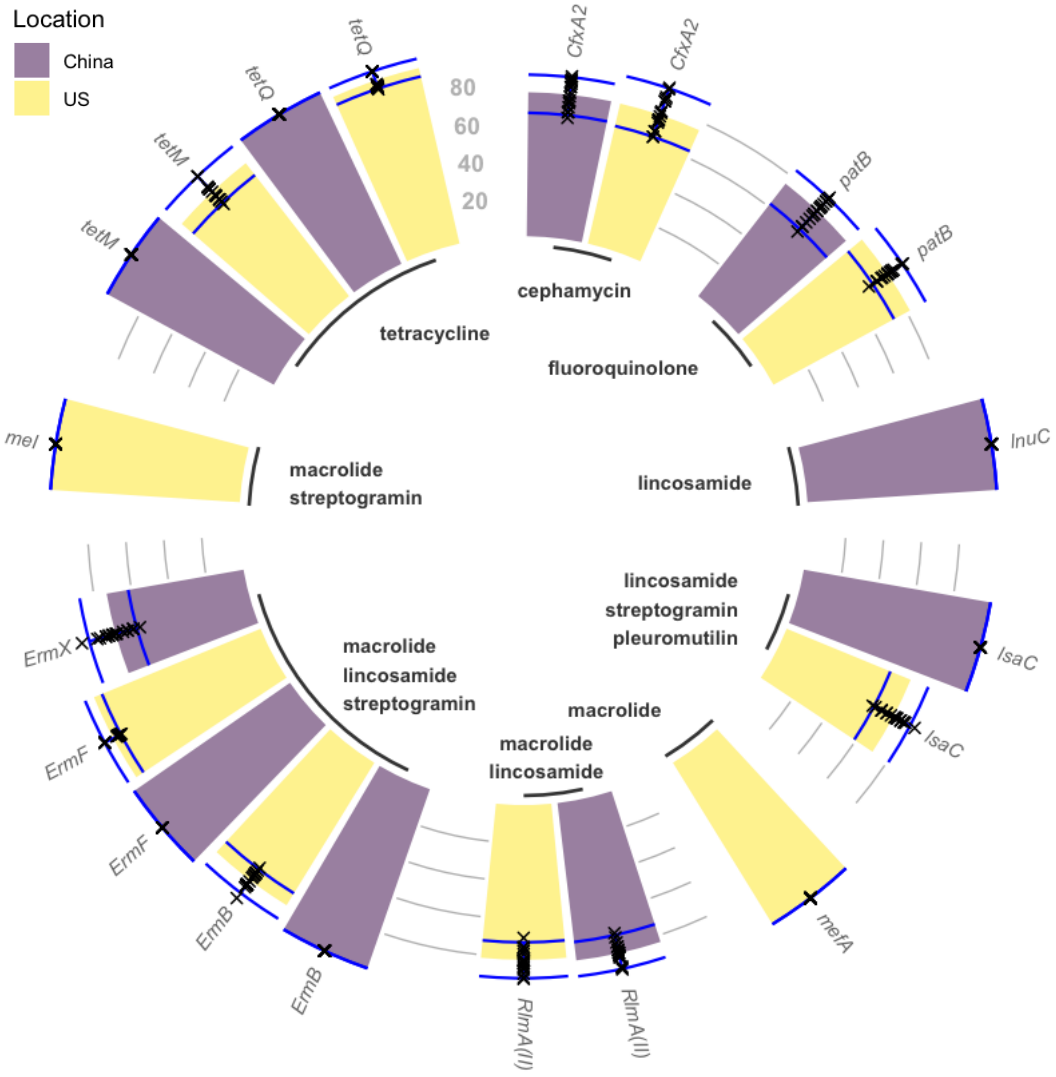

Figure 2Db. ARGs that are found ≥ 70% of dental plaque samples.

China (n = 18) and the USA (n = 18). The height of bars are the means and the error bars are 95% CIs of percentages extracted from bootstrapping samples 20 times shown by points.

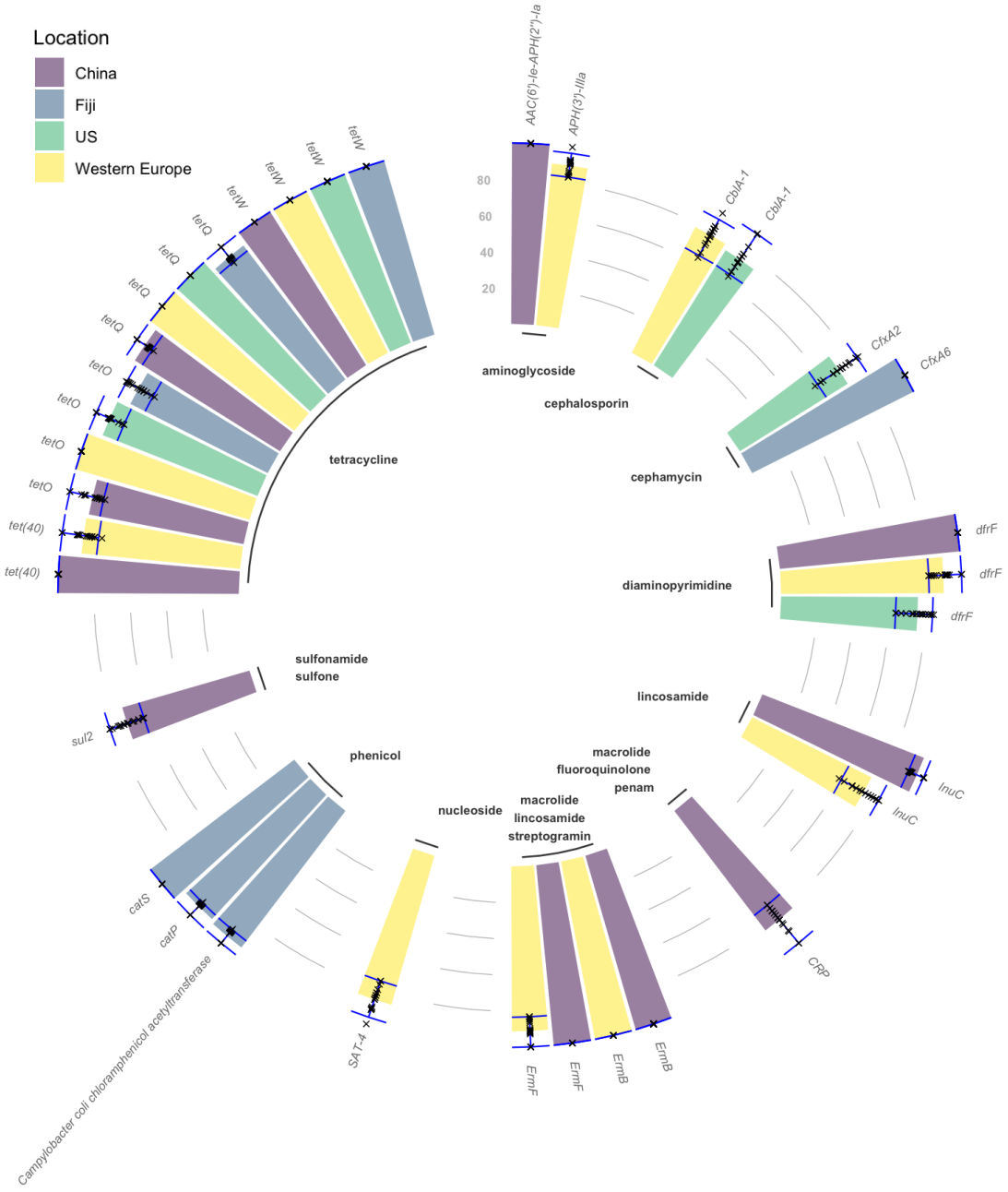

Figure 2Dc. ARGs that are found ≥ 70% of stool samples.

China (n = 18), Fiji (n = 18), the USA (n = 18) and Western Europe (n = 18). The height of bars are the means and the error bars are 95% CIs of percentages extracted from bootstrapping samples 20 times shown by points.

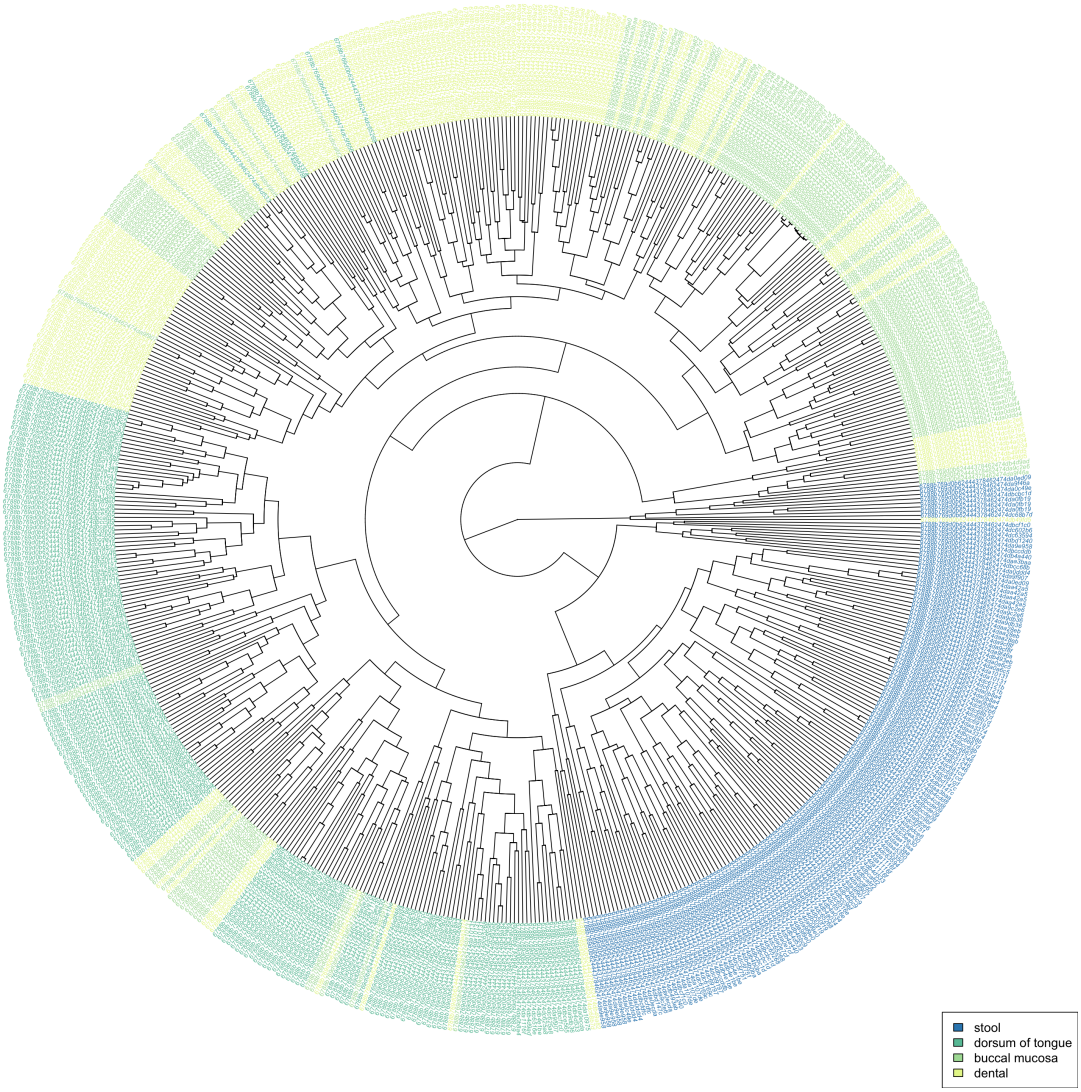

Figure 2E. Longitudinal USA samples clustered by ARG abundance profiles.

Hierarchical clustering of ARG abundance (log10[RPKM+1]) (complete method on Euclidean distance matrix) and labelled by body site: buccal mucosa: n = 55 (36 with two, 18 with three and 1 with six timepoints), dorsum of tongue: n = 69 (43 with two, 24 with three and 2 with four timepoints), dental plaque: n = 67 (43 with two, 20 with three, 1 with four and 3 with six timepoints), stool n = 57 (33 with two, 21 with three, 2 with four and 1 with six timepoints). These samples were collected within two years with individuals having had no antimicrobial treatment in that time.

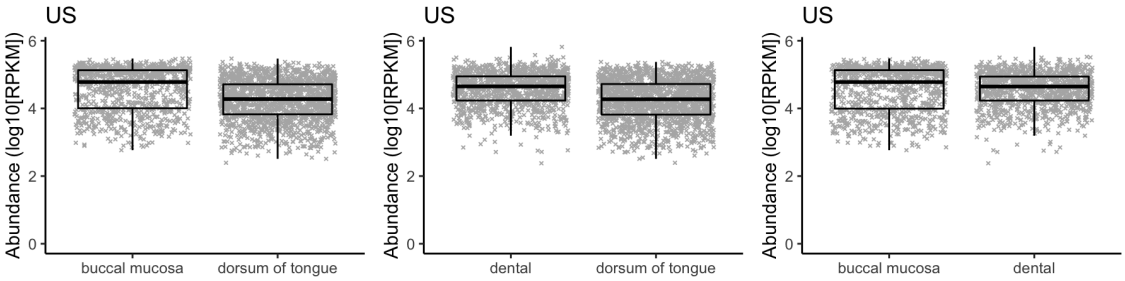

Figure 2F. Comparing ARG abundance between oral cavity sites.

Absolute abundance in log10 of RPKM of ARGs for paired samples of individuals from the USA (buccal mucosa and dorsum of tongue: n = 86, dental plaque and dorsum of tongue: n = 89, buccal mucosa and dental plaque: n = 86). Centre line is median, box limits are upper and lower quartiles, whiskers are 1.5x interquartile ranges and points beyond whiskers are outliers.

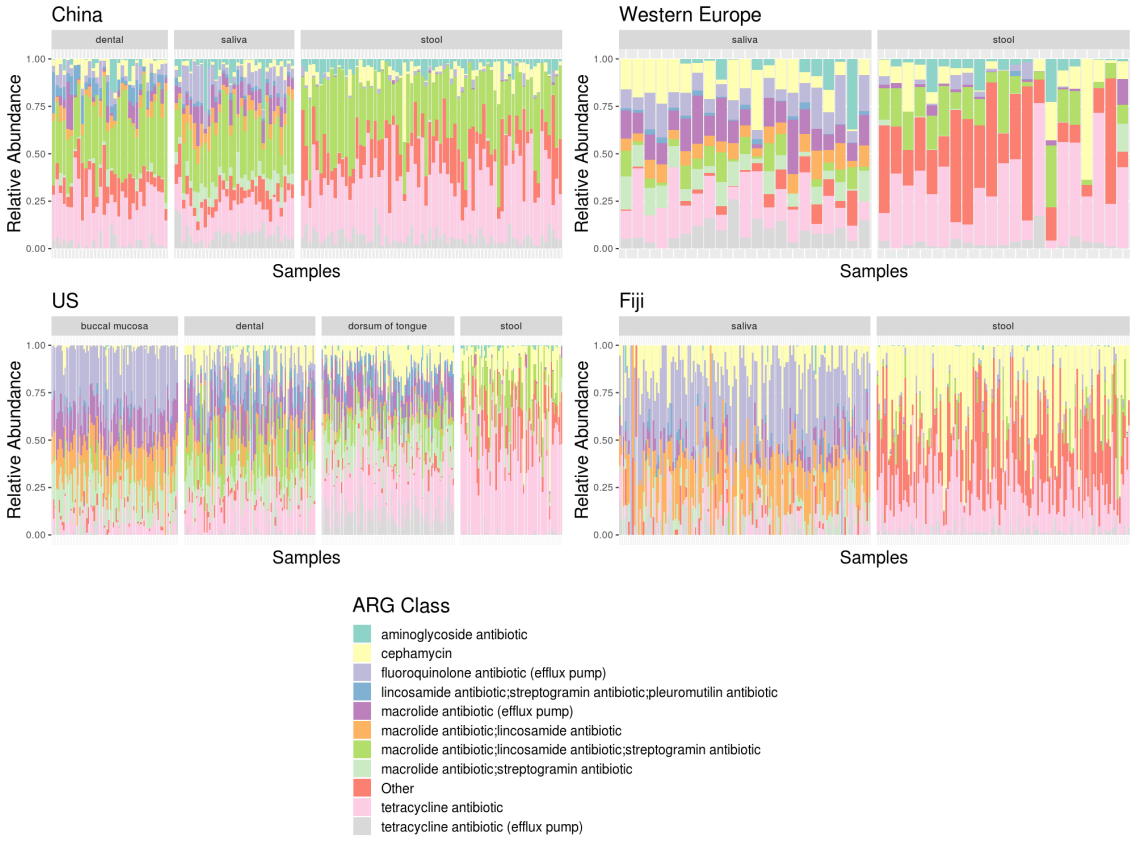

Figure 2G. Comparing ARG abundance of different body sites between individuals.

Relative abundance of reads labelled by the top ten most abundant ARG classes across all geographical locations or Other classes for each sample of individuals from China (saliva: n = 33, dental plaque: n = 32, stool: n = 72), Fiji (saliva: n = 136, stool: n = 137), the USA (buccal mucosa: n = 87, dental plaque: n = 90, dorsum of tongue: n = 91, stool: n = 70) and Western Europe (saliva: n = 21, stool: n = 21).

Figure 2Ha. ARG abundance shown for each ARG from China.

Heatmaps of abundance (log10[RPKM+1]) clustered by hierarchical clustering column-wise by sample (complete method on Euclidean distance matrix) and separated row-wise by ARG class (saliva: n = 33, dental plaque: n = 32, stool: n = 72).

Figure 2Hb. ARG abundance shown for each ARG from the USA.

Heatmaps of abundance (log10[RPKM+1]) clustered by hierarchical clustering column-wise by sample (complete method on Euclidean distance matrix) and separated row-wise by ARG class (including longitudinal samples, buccal mucosa: n = 164, dorsum of tongue: n = 188, dental plaque: n = 191, stool n = 156).

Figure 2Hc. ARG abundance shown for each ARG from Fiji.

Heatmaps of abundance (log10[RPKM+1]) clustered by hierarchical clustering column-wise by sample (complete method on Euclidean distance matrix) and separated row-wise by ARG class (saliva: n = 136, stool: n = 137).

Figure 2Hd. ARG abundance shown for each ARG from Western Europe.

Heatmaps of abundance (log10[RPKM+1]) clustered by hierarchical clustering column-wise by sample (complete method on Euclidean distance matrix) and separated row-wise by ARG class (saliva: n = 21, stool: n = 21).

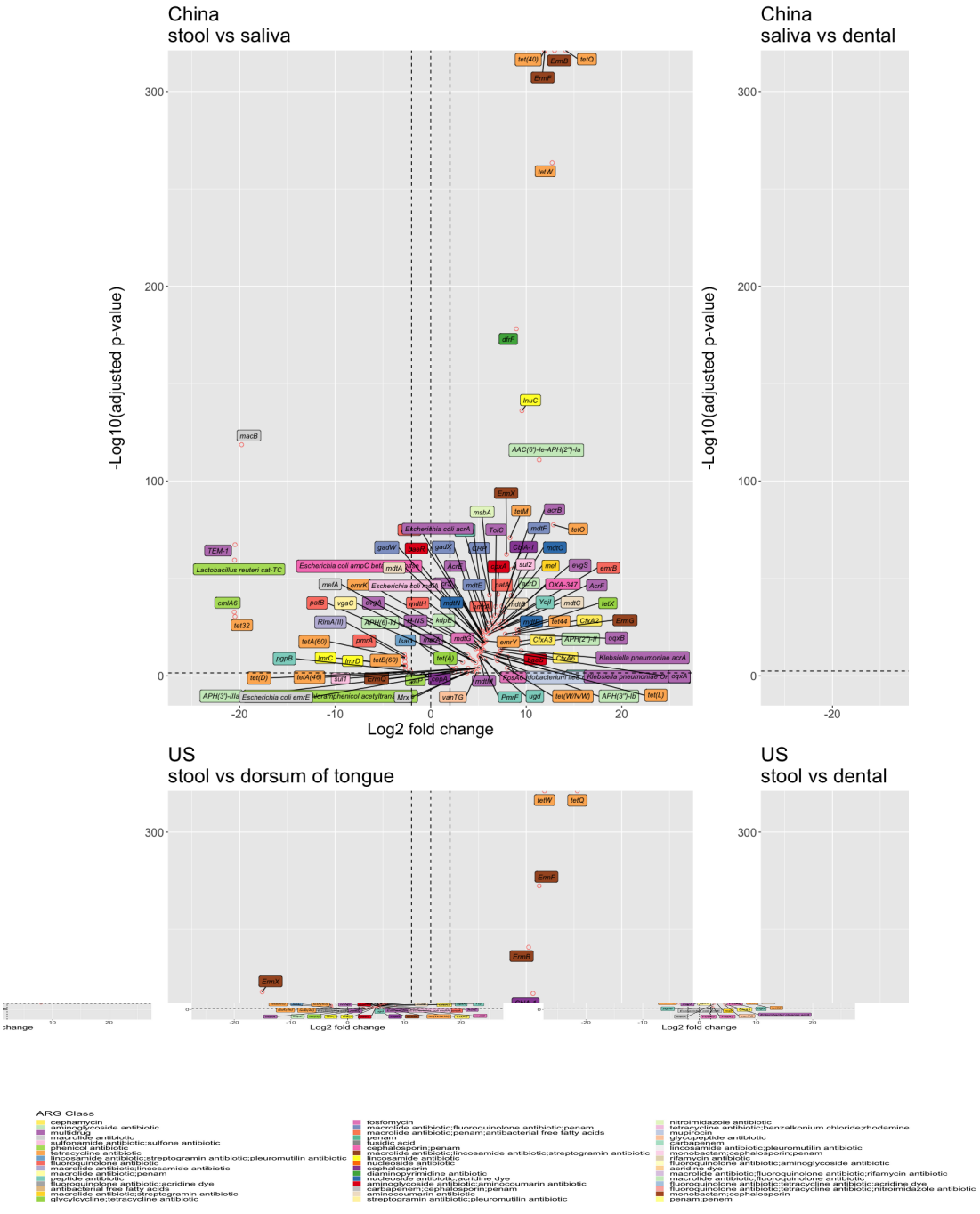

Figure 2Ia. Differential analysis of ARG abundance between stool and saliva samples from China. Volcano plot showing differential analysis, using DESeq2 package in R, between paired samples of adjusted p-value < 0.05 for individuals from China (n = 31).

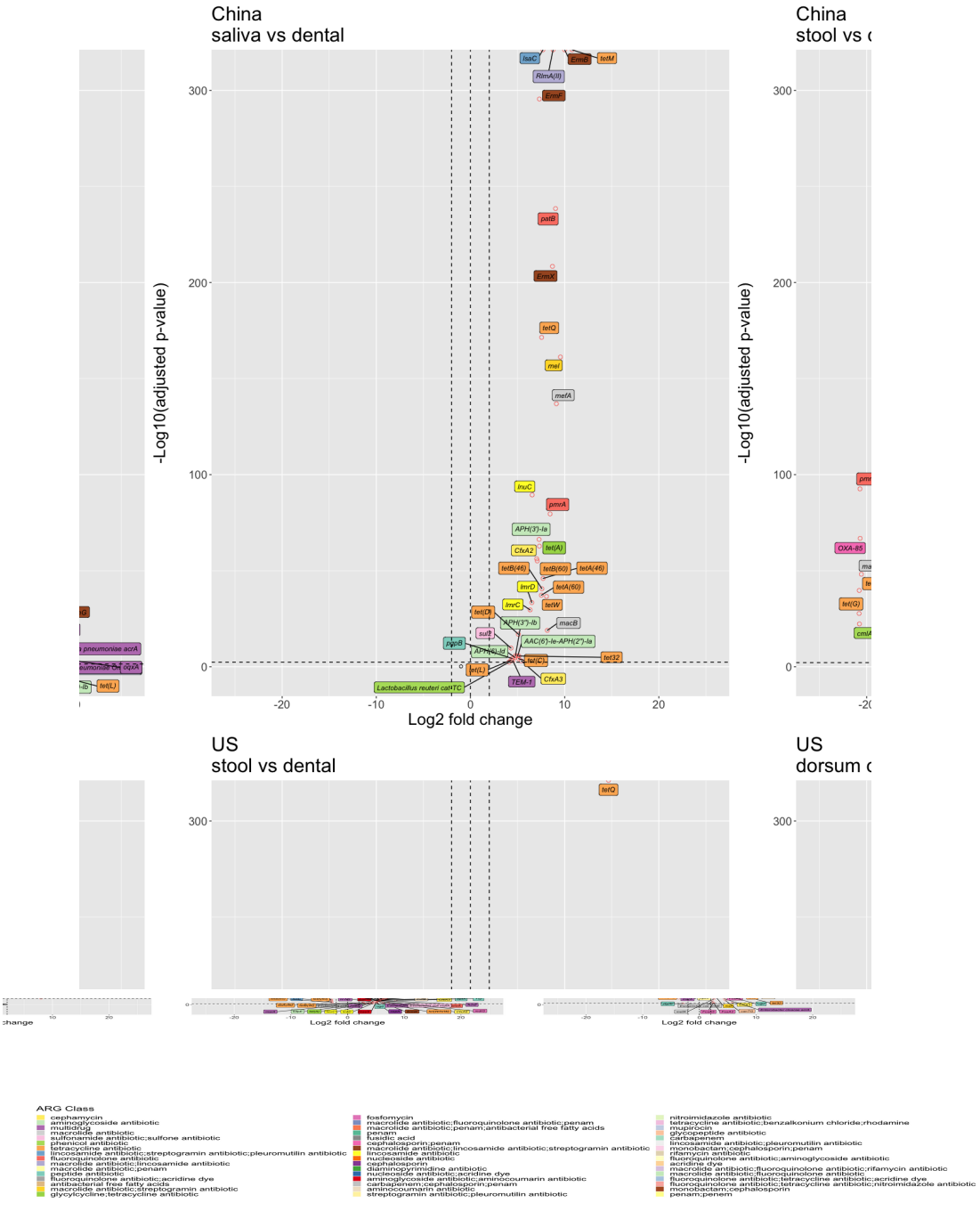

Figure 2Ib. Differential analysis of ARG abundance between saliva and dental samples from China. Volcano plot showing differential analysis, using DESeq2 package in R, between paired samples of adjusted p-value < 0.05 for individuals from China (n = 31).

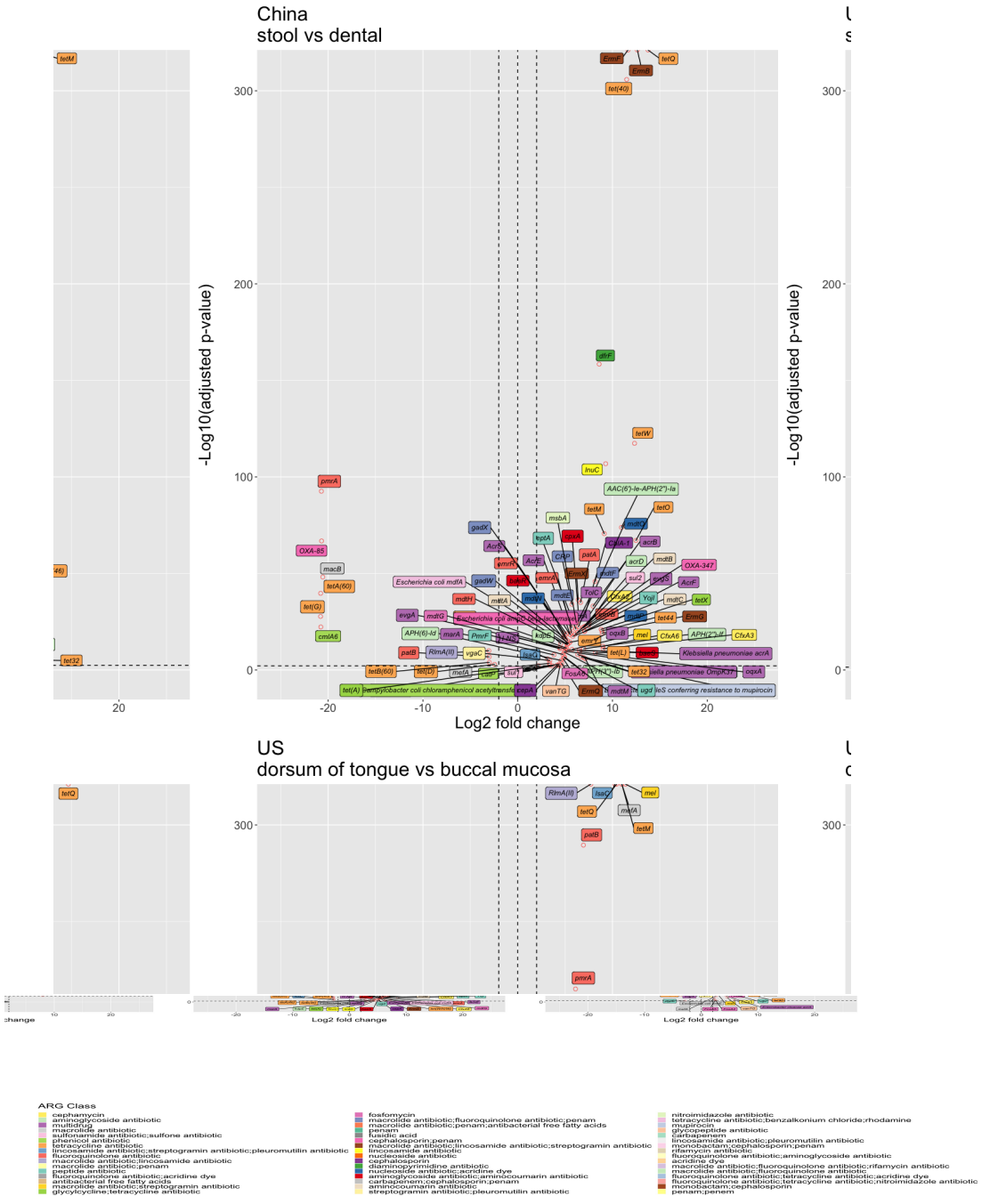

Figure 2Ic. Differential analysis of ARG abundance between stool and dental samples from China. Volcano plot showing differential analysis, using DESeq2 package in R, between paired samples of adjusted p-value < 0.05 for individuals from China (n = 30).

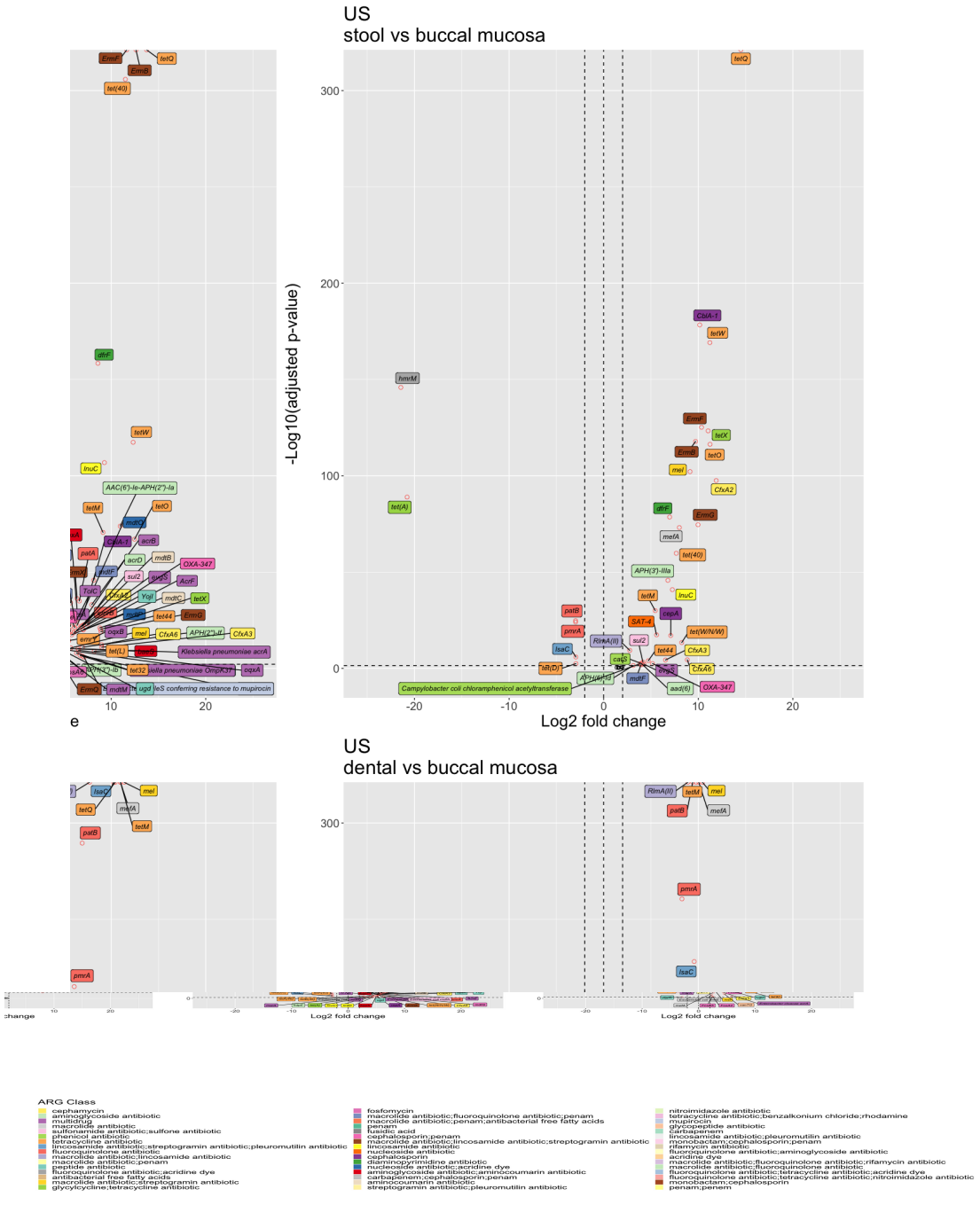

Figure 2Id. Differential analysis of ARG abundance between stool and buccal mucosa samples from the USA.

Volcano plot showing differential analysis, using DESeq2 package in R, between paired samples of adjusted p-value < 0.05 for individuals from the USA (n = 64).

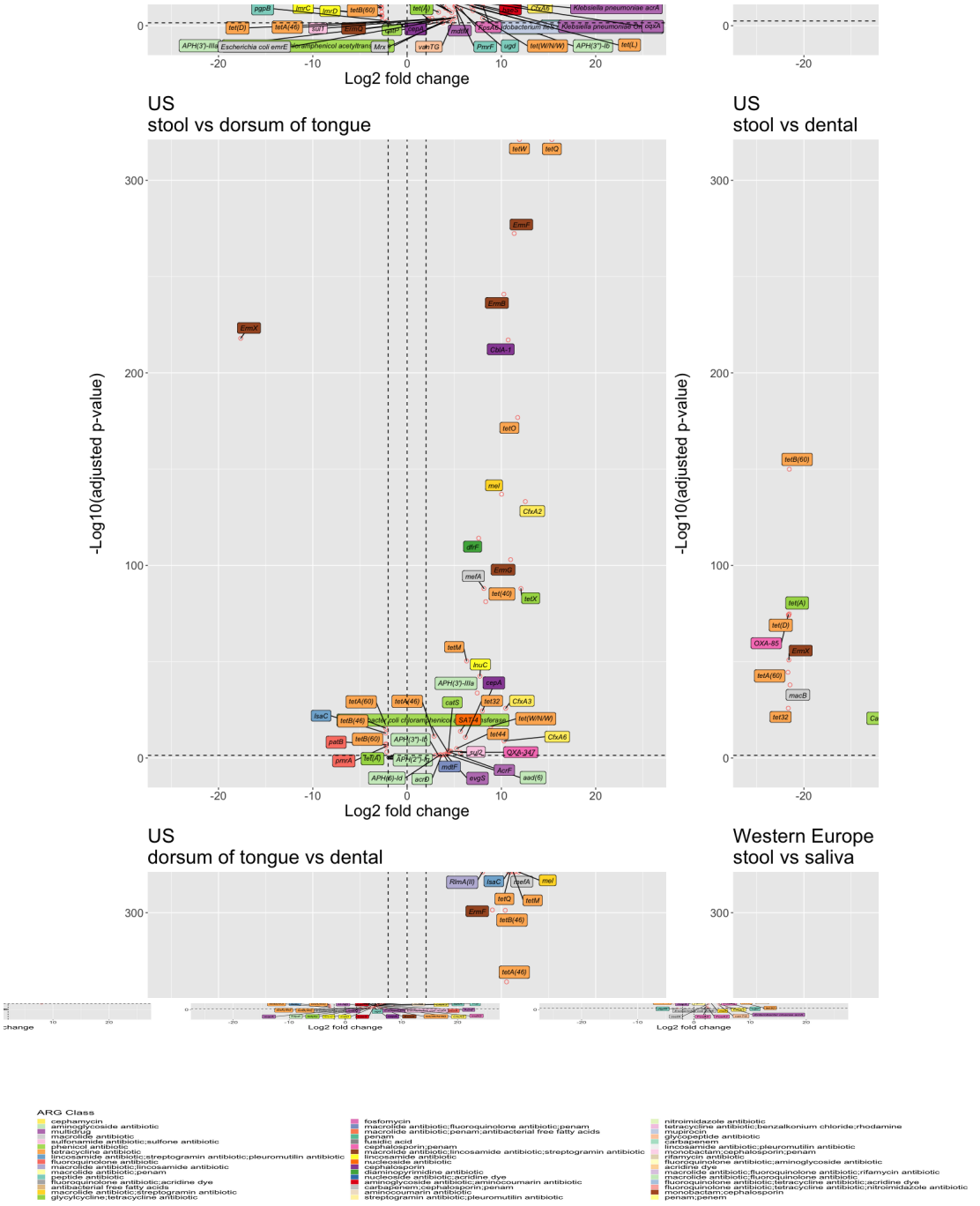

Figure 2Ie. Differential analysis of ARG abundance between stool and dorsum of tongue samples from the USA.

Volcano plot showing differential analysis, using DESeq2 package in R, between paired samples of adjusted p-value < 0.05 for individuals from the USA (n = 69).

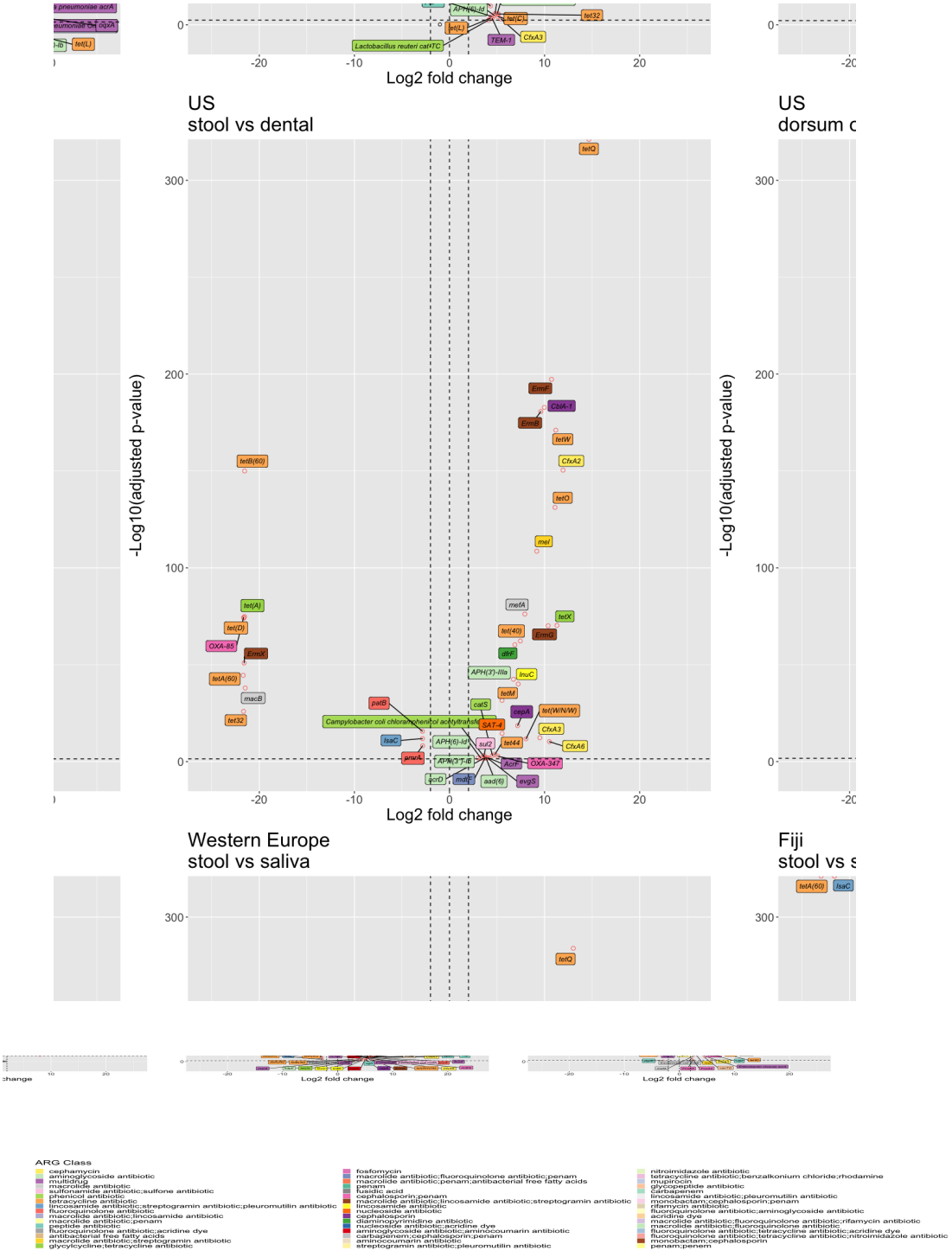

Figure 2If. Differential analysis of ARG abundance between stool and dental samples from the USA.

Volcano plot showing differential analysis, using DESeq2 package in R, between paired samples of adjusted p-value < 0.05 for individuals from the USA (n = 68).

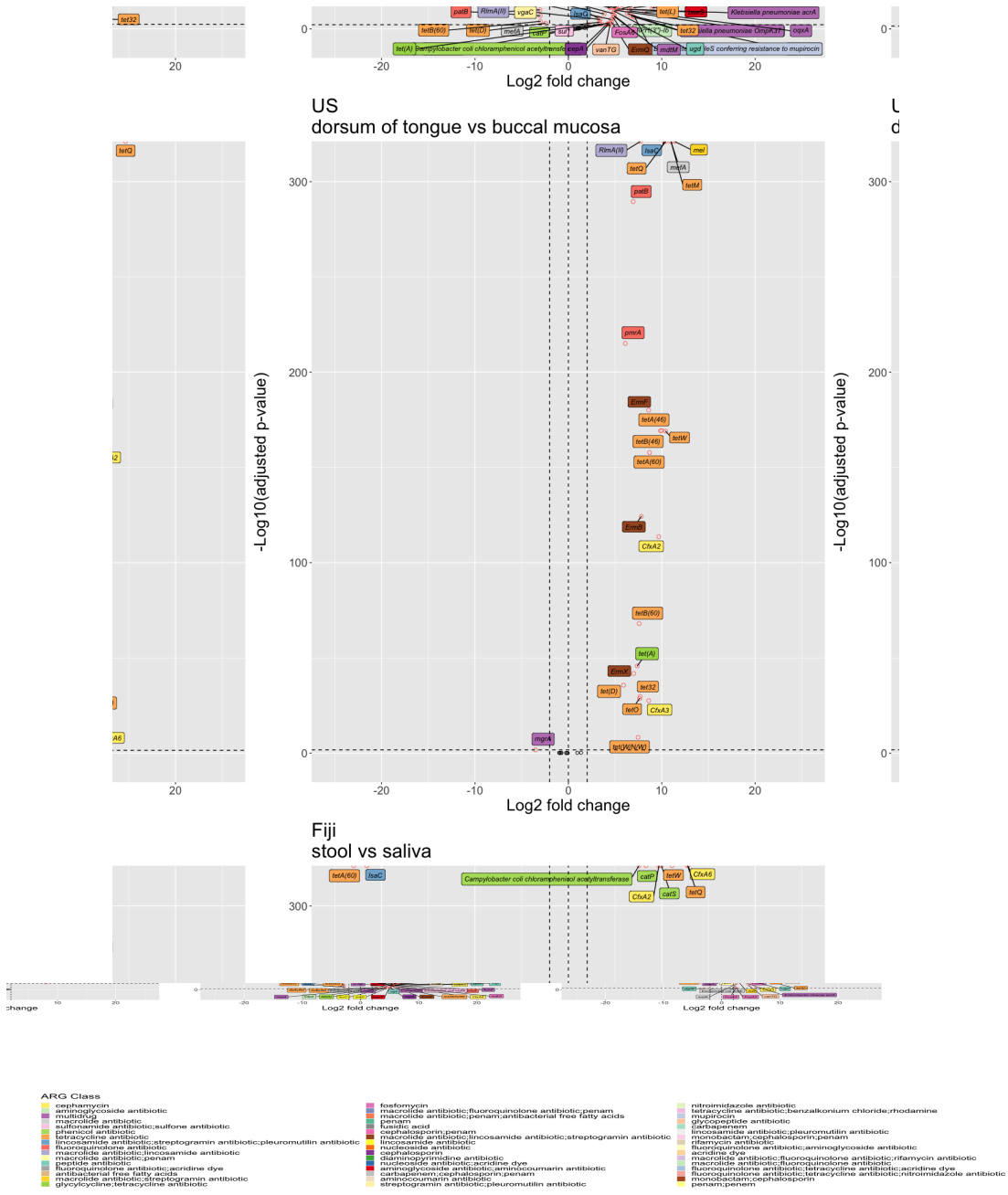

Figure 2Ig. Differential analysis of ARG abundance between dorsum of the tongue and buccal mucosa samples from the USA.

Volcano plot showing differential analysis, using DESeq2 package in R, between paired samples of adjusted p-value < 0.05 for individuals from the USA (n = 86).

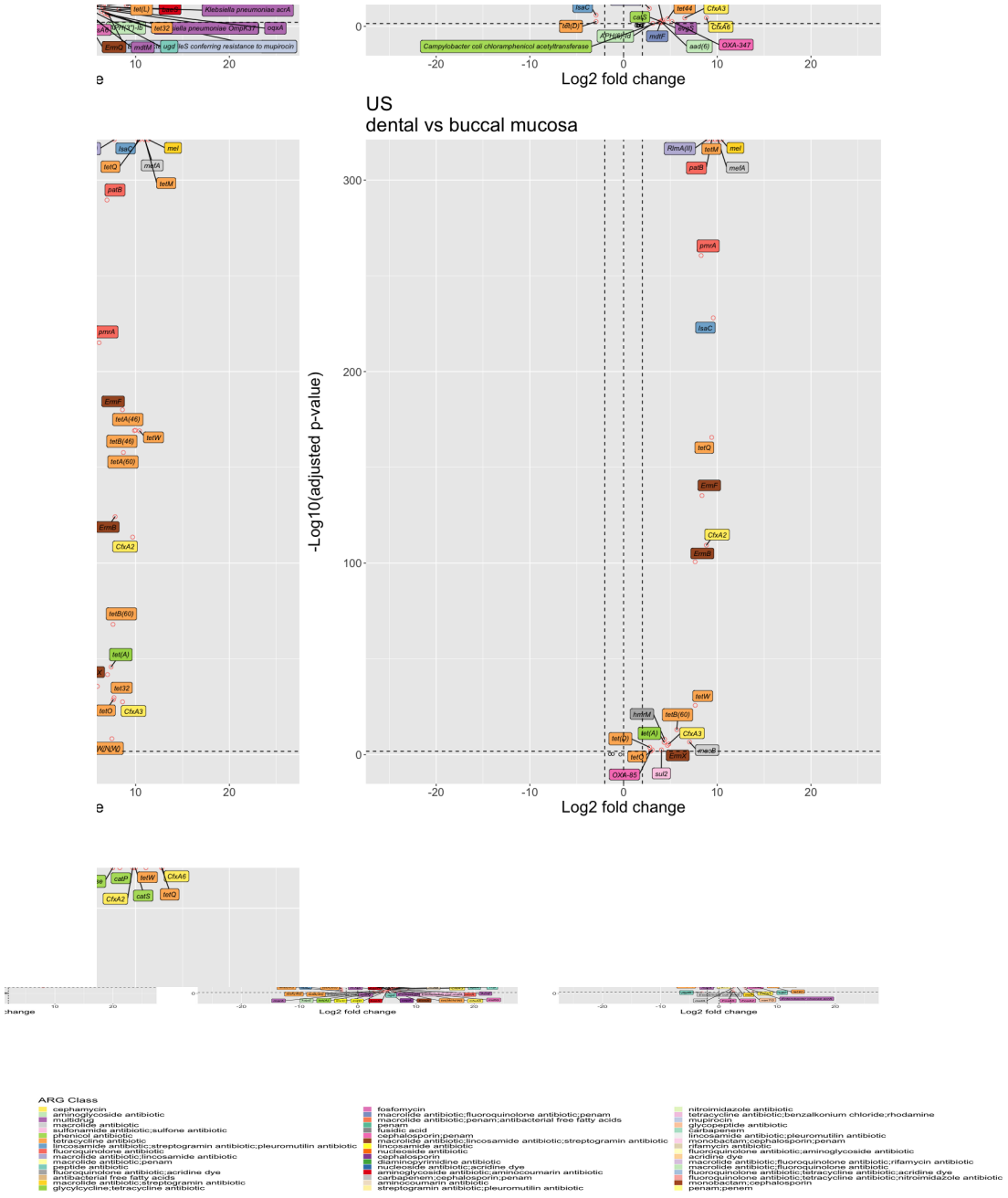

Figure 2Ih. Differential analysis of ARG abundance between dental and buccal mucosa samples from the USA.

Volcano plot showing differential analysis, using DESeq2 package in R, between paired samples of adjusted p-value < 0.05 for individuals from the USA (n = 86).

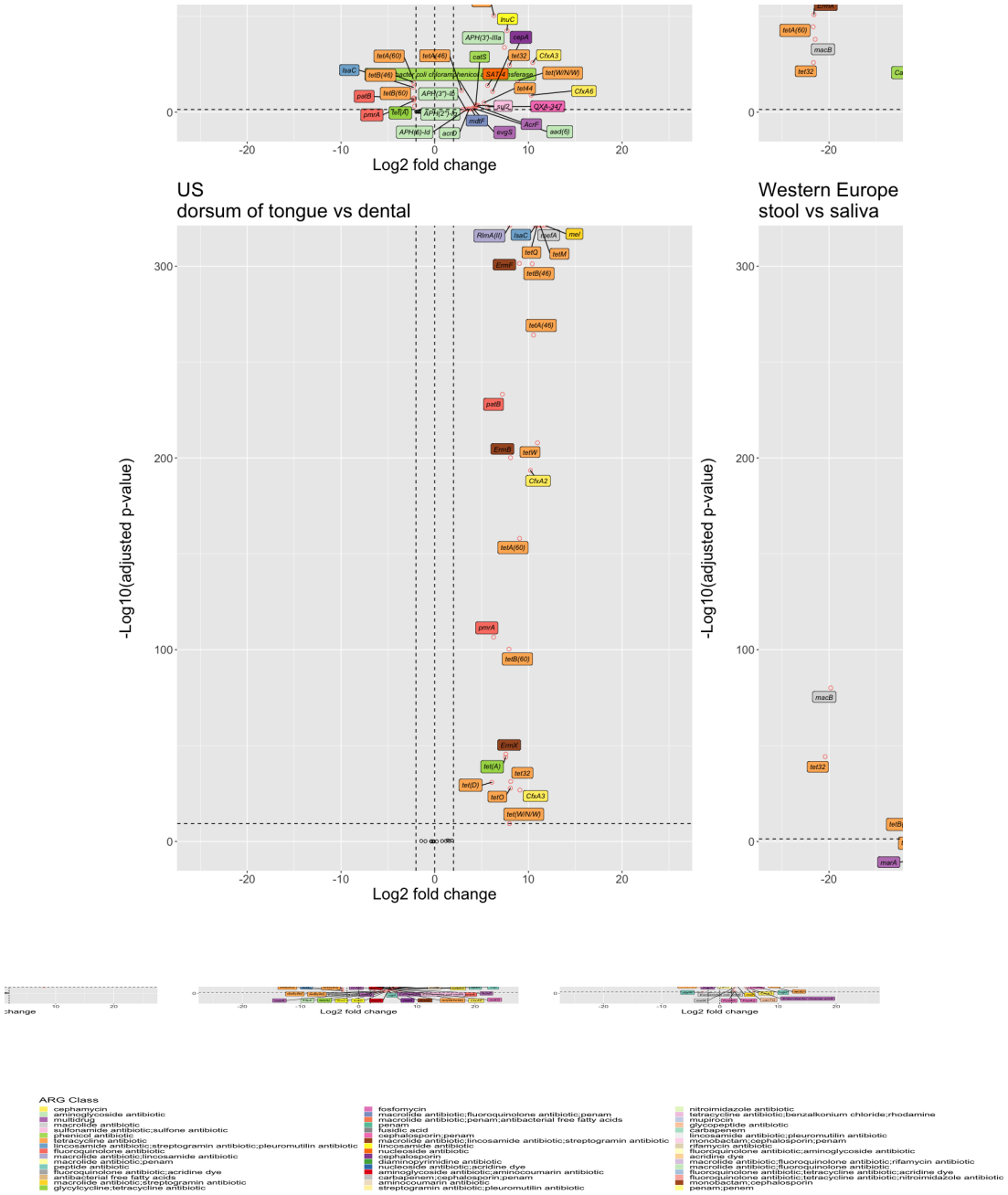

Figure 2Ii. Differential analysis of ARG abundance between dorsum of the tongue and dental samples from the USA.

Volcano plot showing differential analysis, using DESeq2 package in R, between paired samples of adjusted p-value < 0.05 for individuals from the USA (n = 89).

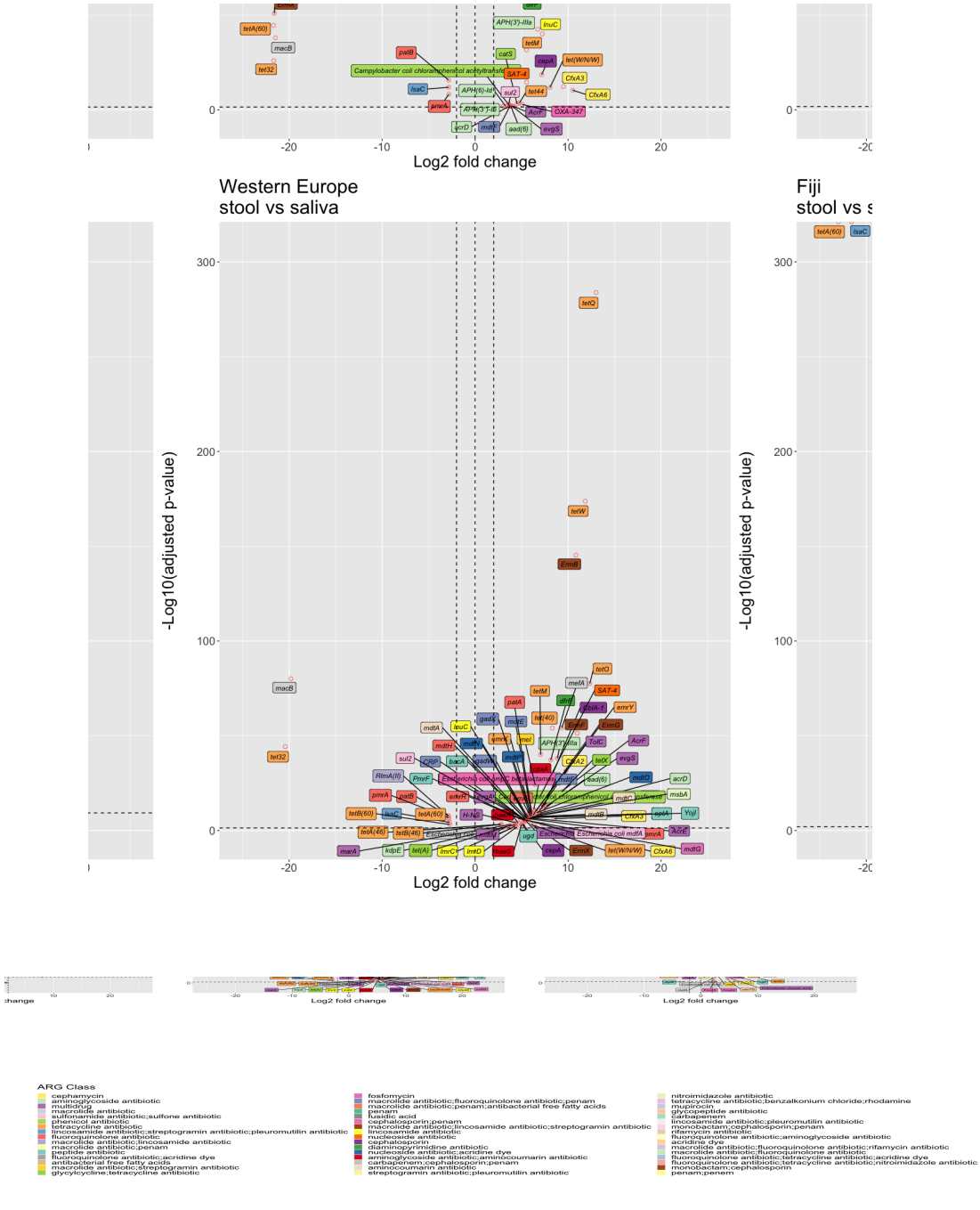

Figure 2Ij. Differential analysis of ARG abundance between stool and saliva samples from Western Europe.

Volcano plot showing differential analysis, using DESeq2 package in R, between paired samples of adjusted p-value < 0.05 for individuals from Western Europe (n = 21).

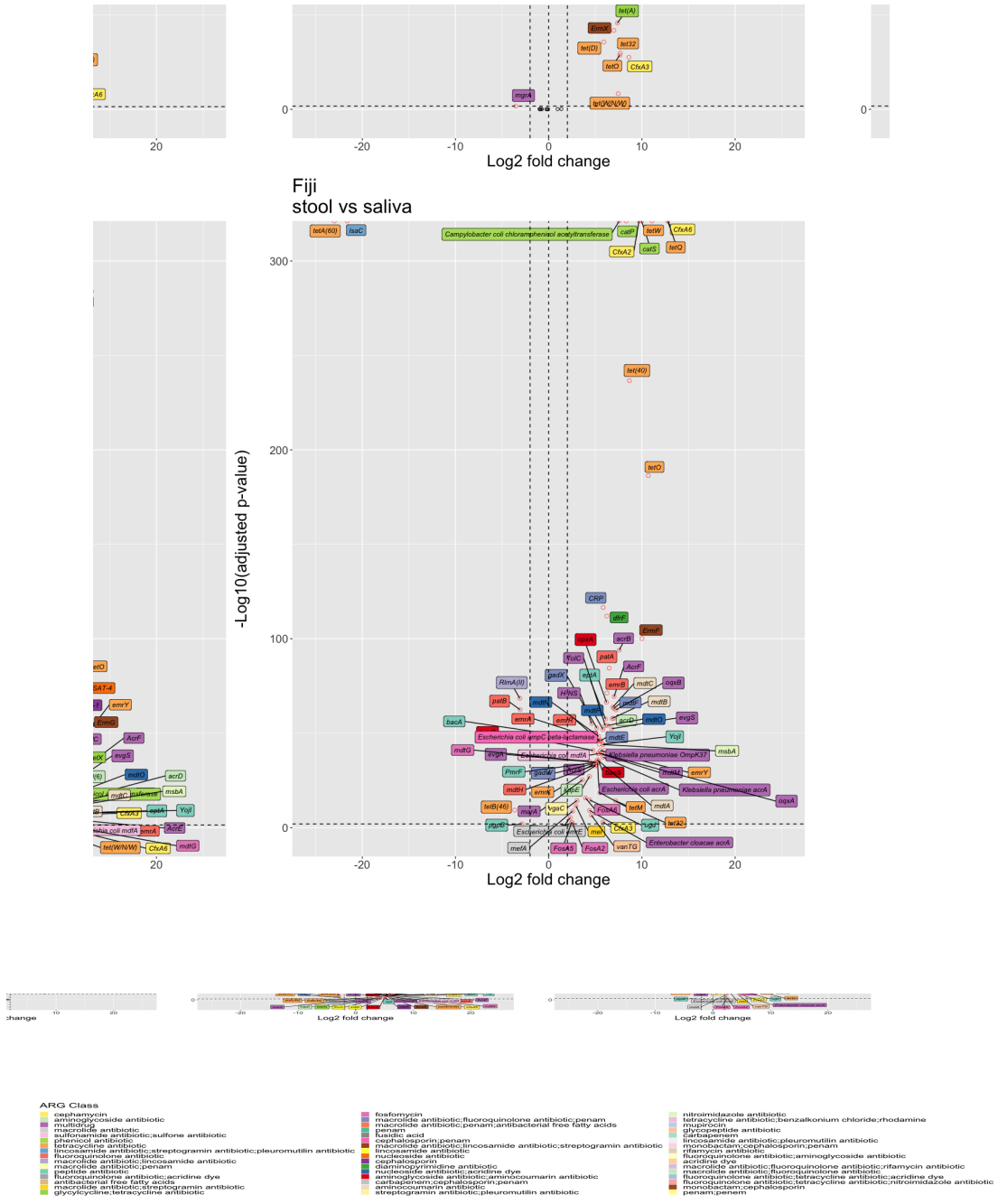

Figure 2Ik. Differential analysis of ARG abundance between stool and saliva samples from Fiji. Volcano plot showing differential analysis, using DESeq2 package in R, between paired samples of adjusted p-value < 0.05 for individuals from Fiji (n = 132).

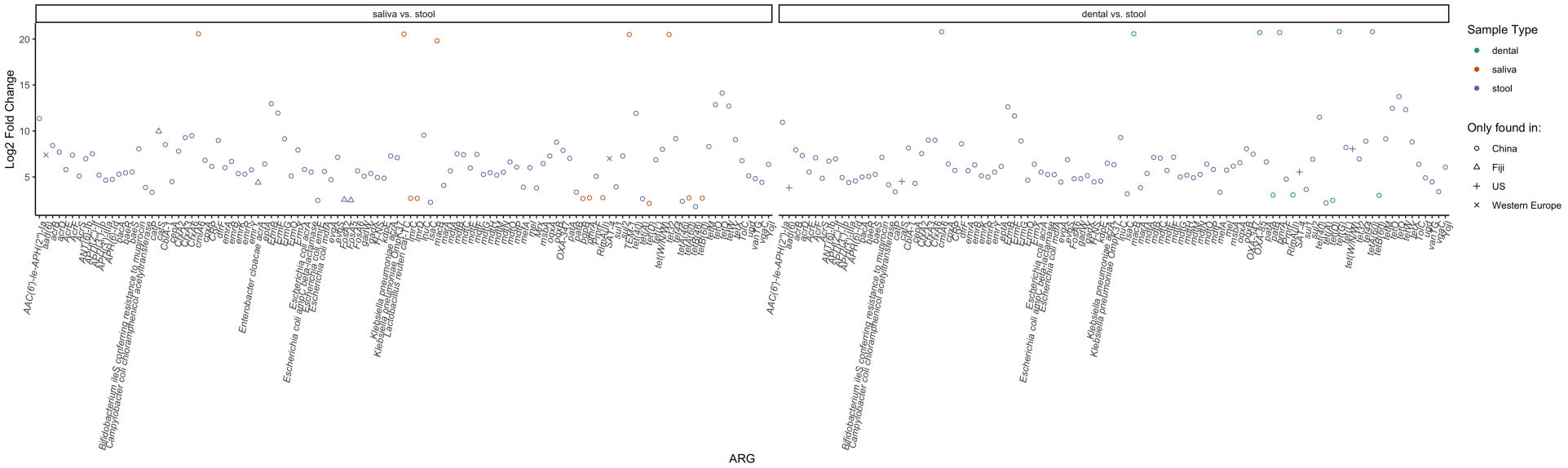

Figure 2J. Log2 fold change of ARGs exclusively found in one geographical location between paired samples.

ARGs selected where adjusted p-value < 0.05 from differential abundance analysis between paired samples of individuals from China (stool and saliva: n = 31, stool and dental plaque: n = 30), the USA (stool and dental plaque: n = 68), Fiji (saliva and stool: n = 137) and Western Europe (saliva and stool: n = 21).

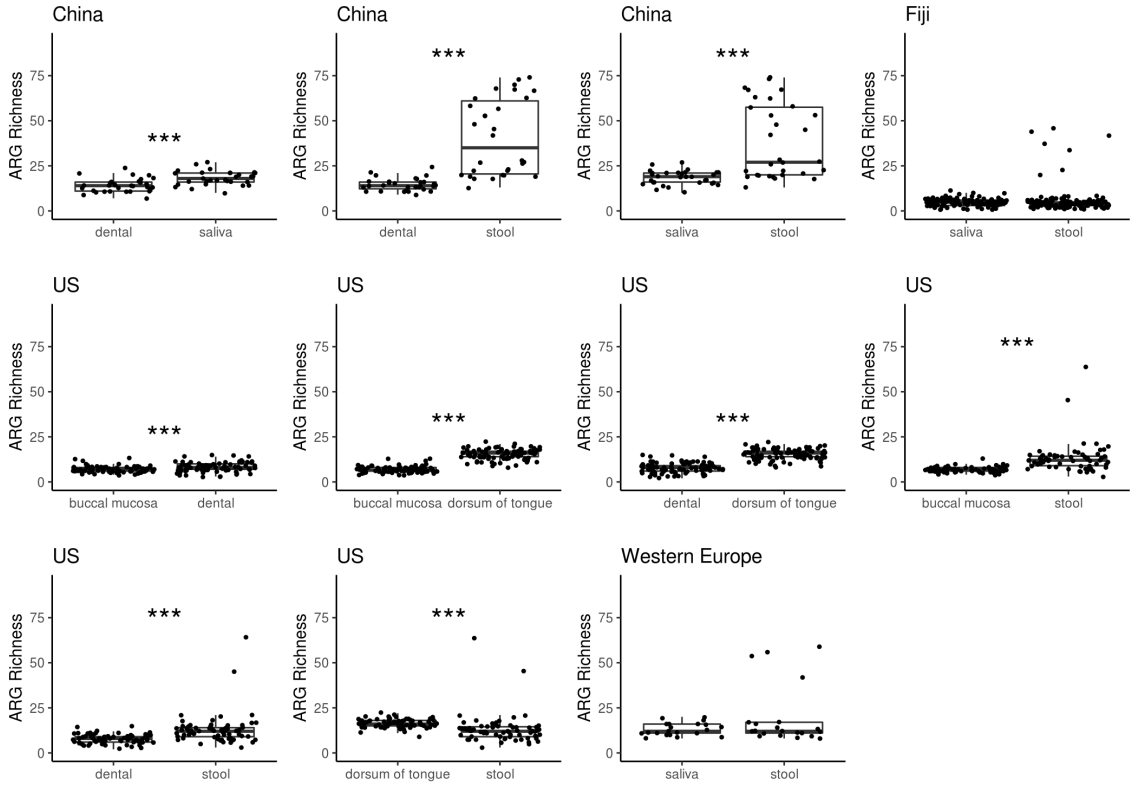

Figure 2K. Comparing ARG richness between paired body sites excluding ARGs that are part of or regulate an efflux pump complex.

ARG richness is defined as the number of unique ARGs that are not part of nor regulate an efflux pump complex for paired samples of individuals from China (dental plaque and saliva: n = 31, stool and dental plaque: n = 30, stool and saliva: n = 31), Fiji (saliva and stool: n = 128), the USA (buccal mucosa and dental plaque: n = 78, buccal mucosa and dorsum of tongue: n = 86, dental plaque and dorsum of tongue: n = 89, buccal mucosa and stool: n = 64, dental plaque and stool: n = 68, dorsum of tongue and stool: n = 67) and Western Europe (saliva and stool: n = 21) with Mann-Whitney, paired, two-sided t-test (p-value < 0.05 as *, < 0.01 as **, < 0.005 as ***). Centre line is median, box limits are upper and lower quartiles, whiskers are 1.5x interquartile ranges and points beyond whiskers are outliers.

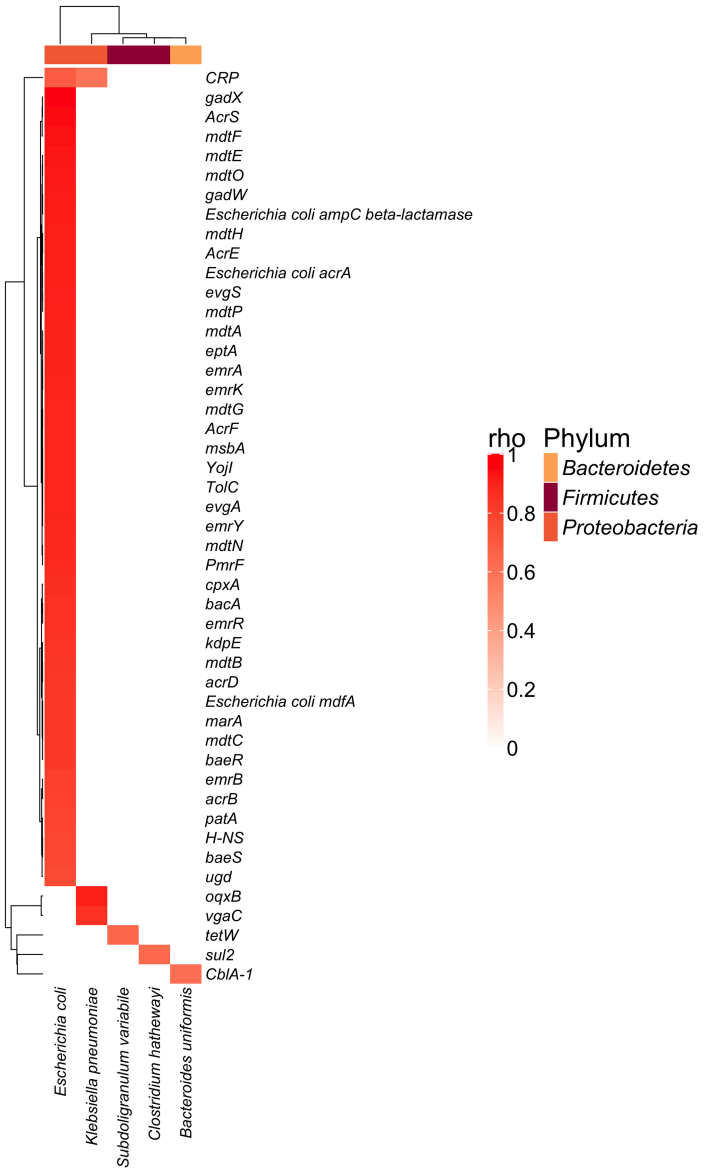

Figure 2L. Spearman's correlation of ARG and species abundance from China stool samples. Samples of individuals (n = 31). Rows and columns are clustered by hierarchical clustering of Euclidean distance. Columns are coloured by phylum. P-values are adjusted by Benjamini-Hochberg multiple test correction. *Rho* shown only where adjusted p-value < 0.05.

Table 2M. Coefficients and p-values from linear regression between ARG class abundance and antibiotic prescriptions in 2015 for each country and body site.

Antibiotic prescriptions measured in DDD Per 1000 Individuals. Coefficients are the r2 and intercept values. Linear regression fits the data into a linear model by equation $y = bx + a + e$, where b is the slope of the line and a is the intercept. e represents the residuals that are the difference between the data and the fitted model. The r2 value is the proportion of variation in the y (ARG class abundance) due to variation in the x (antibiotic prescriptions). The intercept value is the expected value of y when the value of x is 0. p-values > 0.05 indicate there is no statistically significant relationship between the ARG class abundance and antibiotic prescriptions in 2015.

r2 value	Intercept	p-value	Country	Body Site
0.0216	3.48E-07	0.728	China	dental
0.0376	1.07E-06	0.645	China	saliva
0.0253	2.54E-09	0.641	China	stool
0.00572	3.00E-09	0.847	Philippines	saliva
0.169	0.000244	0.418	USA	buccal mucosa
0.111	1.01E-07	0.382	USA	dental
0.0812	0.000117	0.536	USA	dorsum of tongue
0.0710	2.67E-08	0.428	USA	stool
0.124	1.56E-07	0.392	Western Europe	saliva
0.00950	1.02E-09	0.776	Western Europe	stool

Appendix 3: Chapter 3 Supplementary Figures

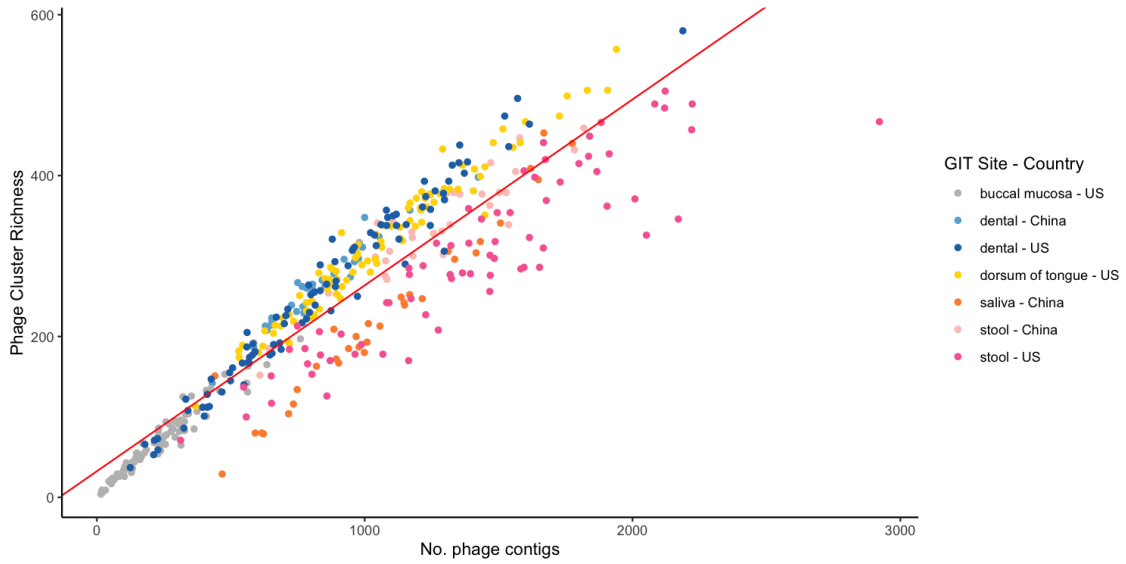

Figure 3A. Linear regression of Phage Cluster Richness against number of phage contigs.
For each sample from China and the USA (Adjusted r2 = 0.8618, p-value < $2.2x10^{-16}$)

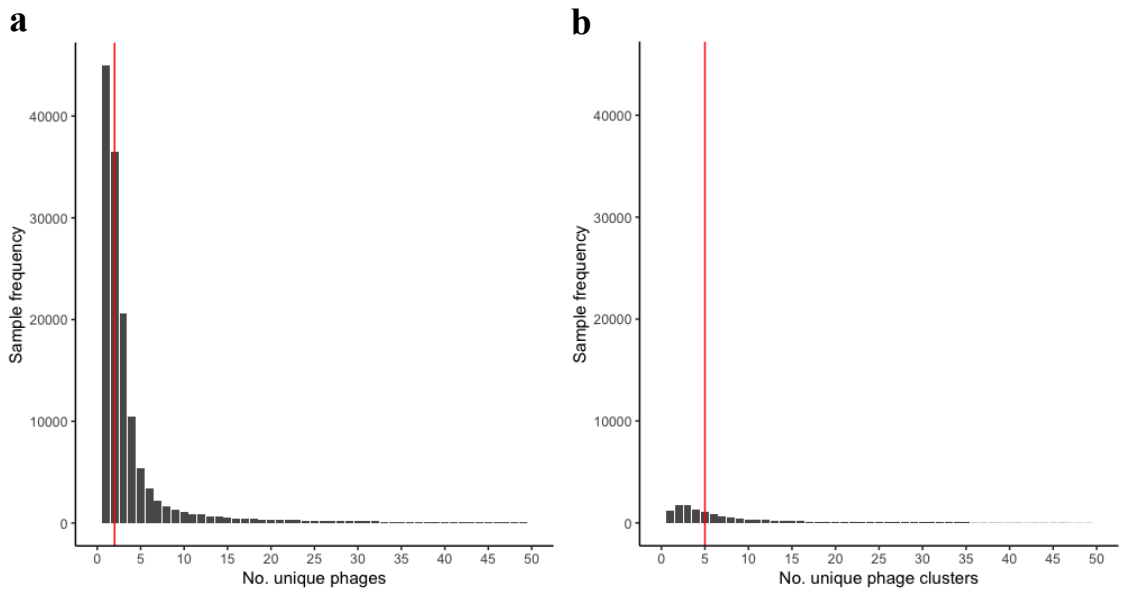

Figure 3B. Frequency of samples containing a number of unique a) phages and b) phage clusters.
Red line represents median number.

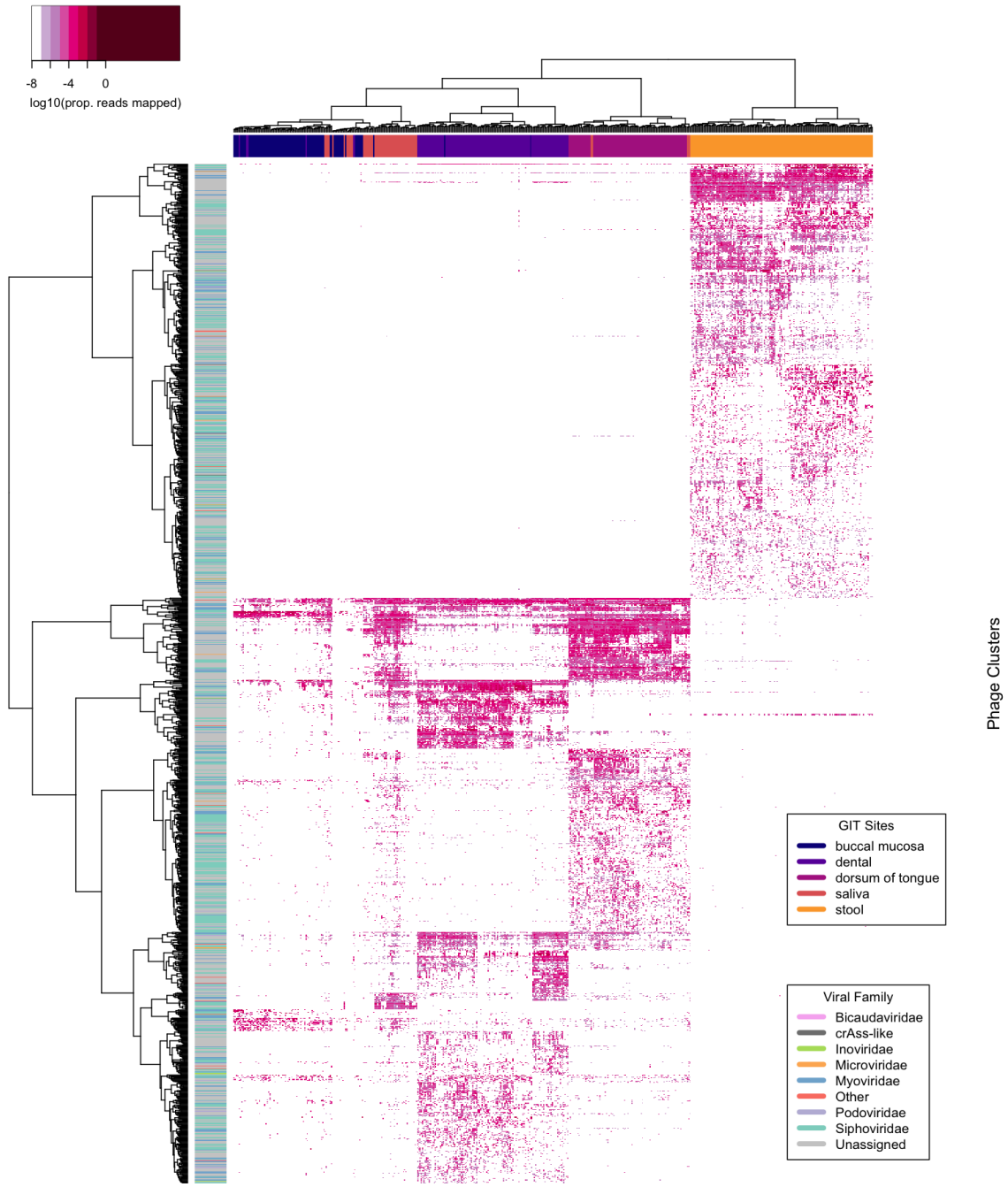

Figure 3C. Log10 of the proportion of reads mapped to differentially abundant phage clusters for each sample.

Samples and phage clusters clustered by hierarchical clustering. Phage clusters that were differentially abundant between groups were selected where Bonferroni-corrected adjusted p-values < 0.001 from Kruskal-Wallis Rank Sum test. USA buccal mucosa (n = 87), dorsum of tongue (n = 90), dental plaque (n = 90) and stool (n = 70); China dental plaque (n = 32), saliva (n = 33) and stool (n = 72); and Philippines saliva (n = 24).

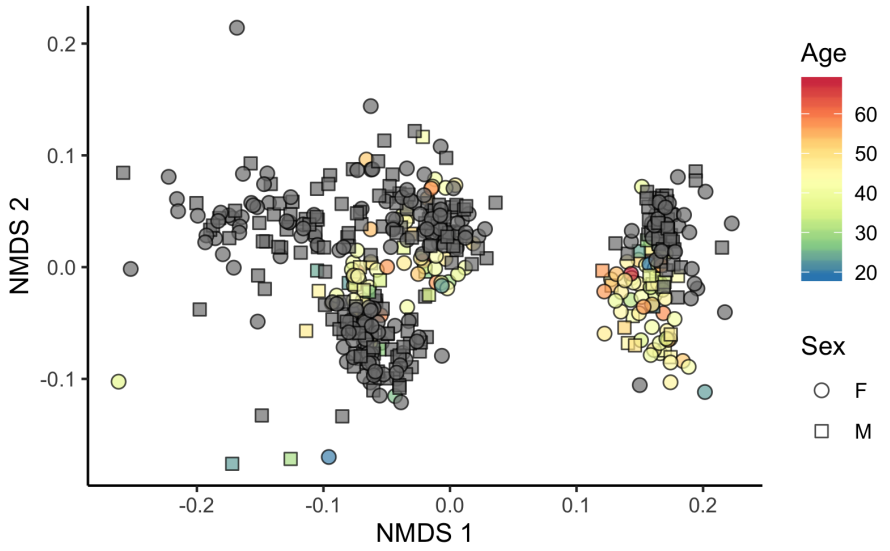

Figure 3D. Phage incidence and abundance profiles.

NMDS of Bray-Curtis dissimilarities between phage cluster incidence profiles of samples (excluding longitudinal USA) labelled by GIT site and geographical location.

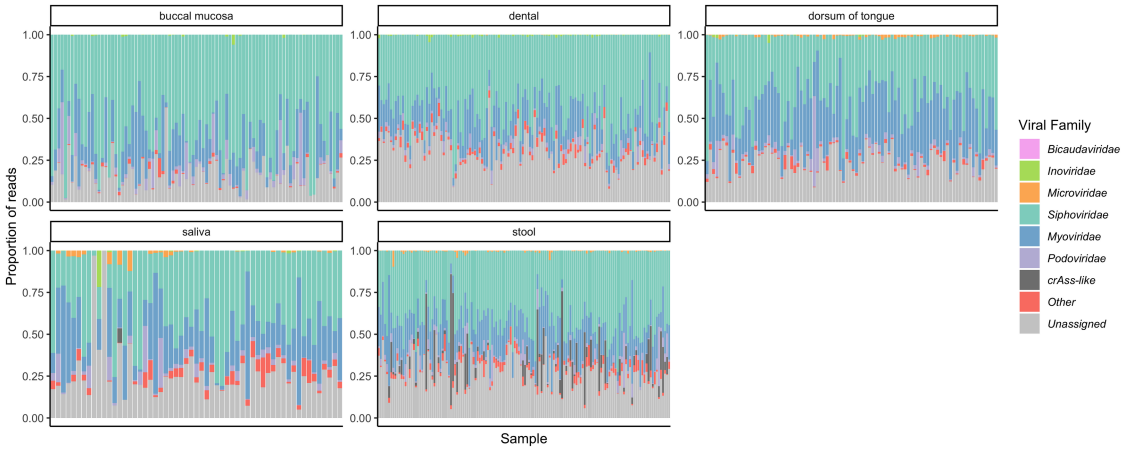

Figure 3E. Relative abundance of phage taxonomy across GIT sites.

Proportion of total reads that are mapped to phage clusters, coloured by viral family, for each sample (excluding USA longitudinal). "Other" represents non-phage viral families, *Alloherpesviridae*, *Ascoviridae*, *Baculoviridae*, *Flaviviridae*, *Herpesviridae*, *Iridoviridae*, *Marseilleviridae*, *Mimiviridae*, *Nudiviridae*, *Phycodnaviridae*, *Picornaviridae*, *Pithoviridae*, *Poxviridae* and *Retroviridae*. USA buccal mucosa (n = 87), dorsum of tongue (n = 90), dental plaque (n = 90) and stool (n = 70); China dental plaque (n = 32), saliva (n = 33) and stool (n = 72); and Philippines saliva (n = 24).

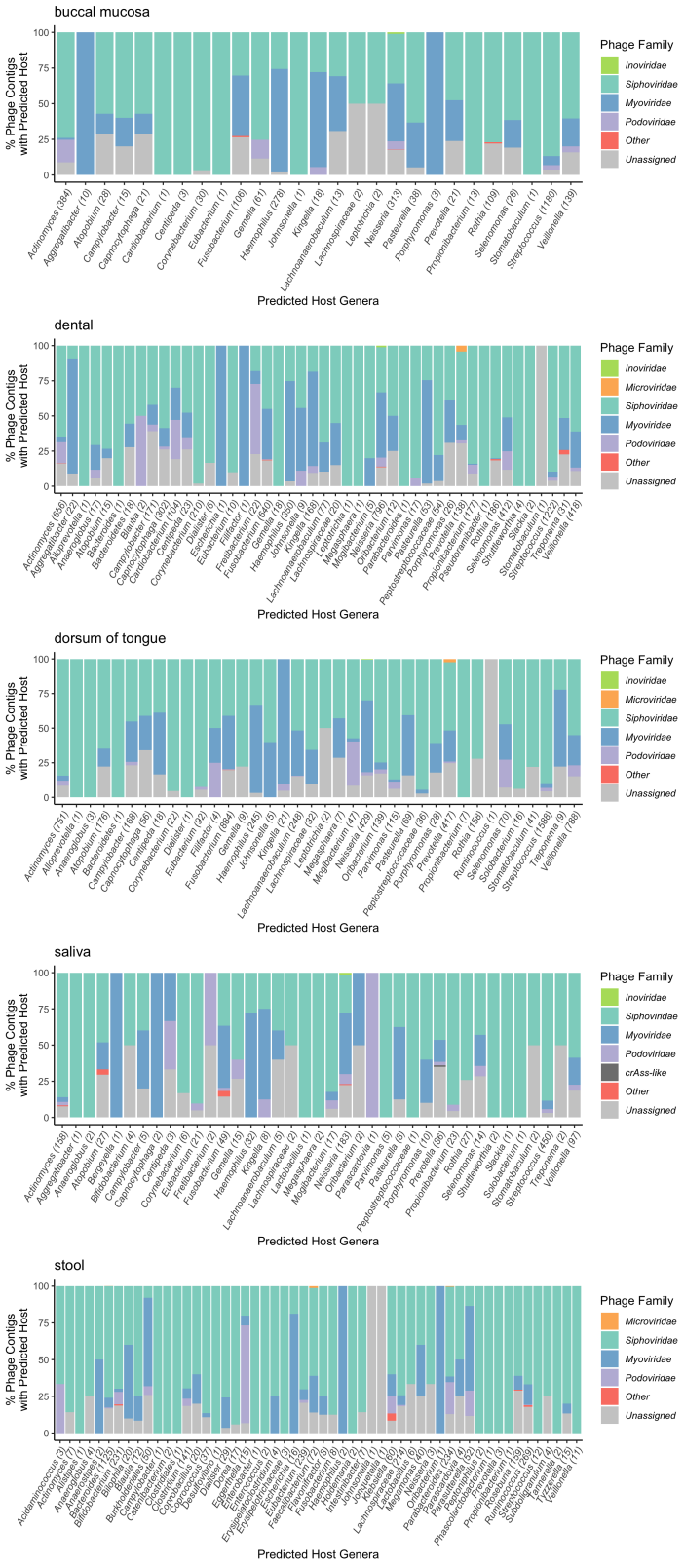

Figure 3F. Percentage of phage contigs of phage families with predicted bacteria hosts.

16,513 (out of 139,929) phage contigs were assigned to host genera. Bracketed number after genus name represent the total number of phage contigs assigned to a particular genus.

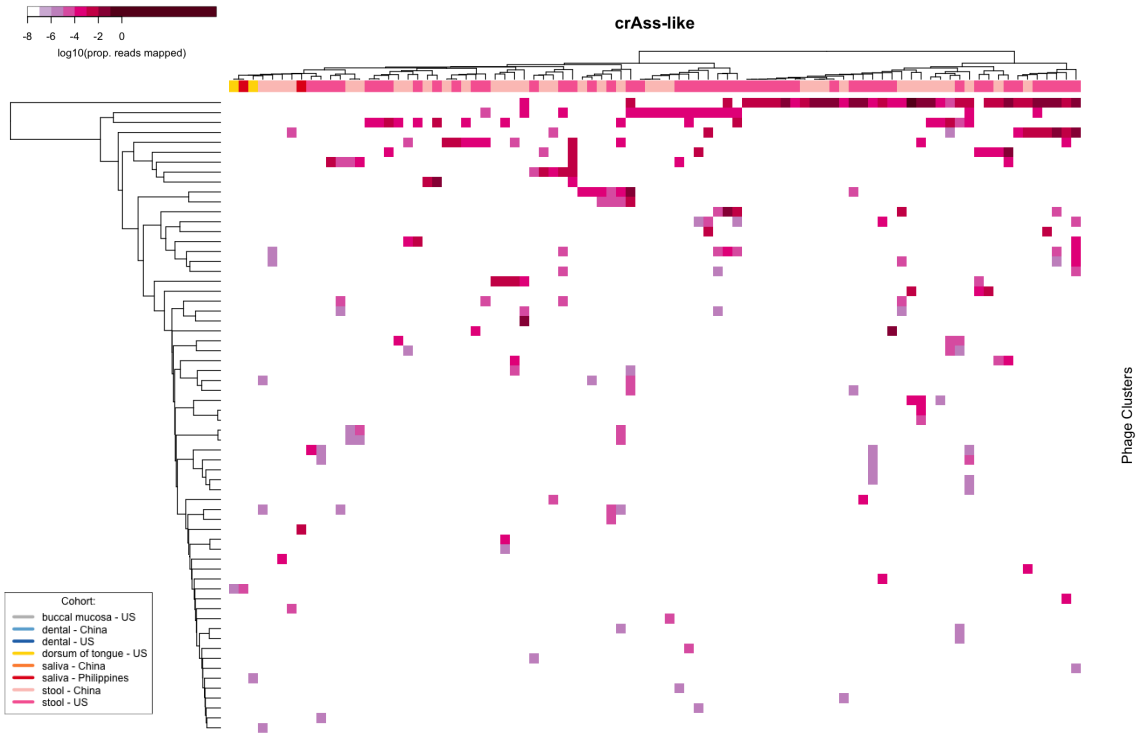

Figure 3Ga. Heatmap of log10 relative abundance of phage clusters for the crAss-like phage family. Dendrograms represent hierarchical clustering of samples (excluding USA longitudinal) colour coded by GIT site and country in the x axis and phage clusters in the y axis.

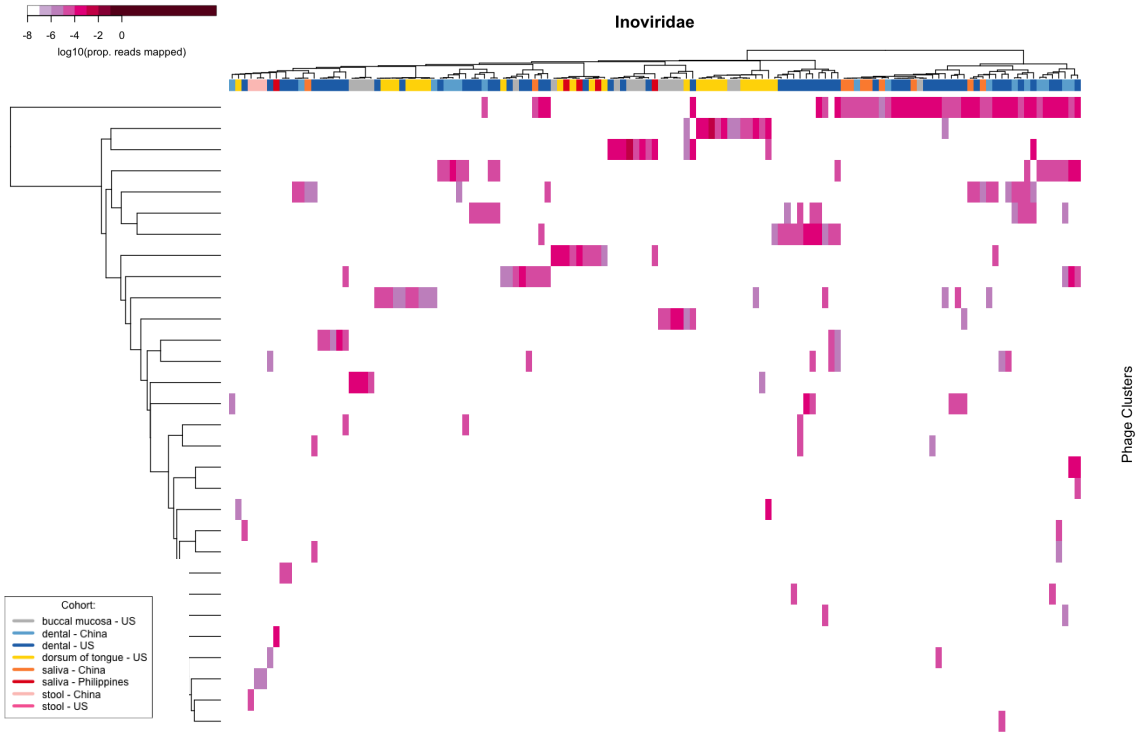

Figure 3Gb. Heatmap of log10 relative abundance of phage clusters for the *Inoviridae* phage family. Dendrograms represent hierarchical clustering of samples (excluding USA longitudinal) colour coded by GIT site and country in the x axis and phage clusters in the y axis.

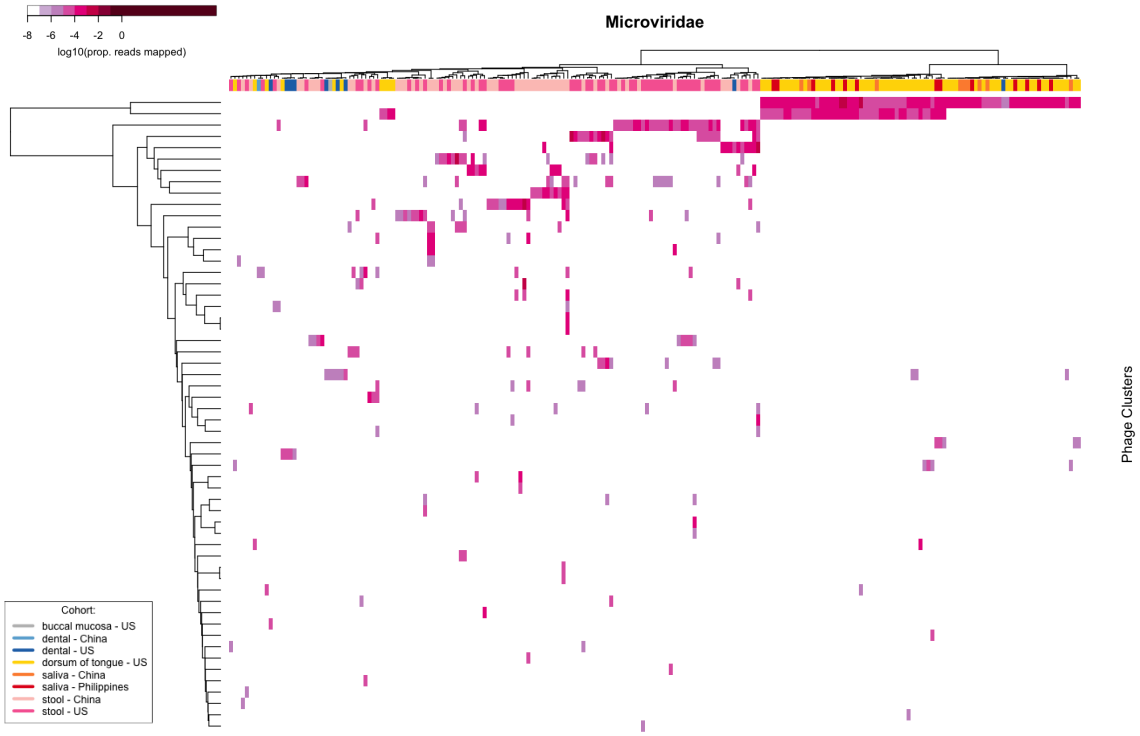

Figure 3Gc. Heatmap of log10 relative abundance of phage clusters for the *Microviridae* phage family.

Dendrograms represent hierarchical clustering of samples (excluding USA longitudinal) colour coded by GIT site and country in the x axis and phage clusters in the y axis.

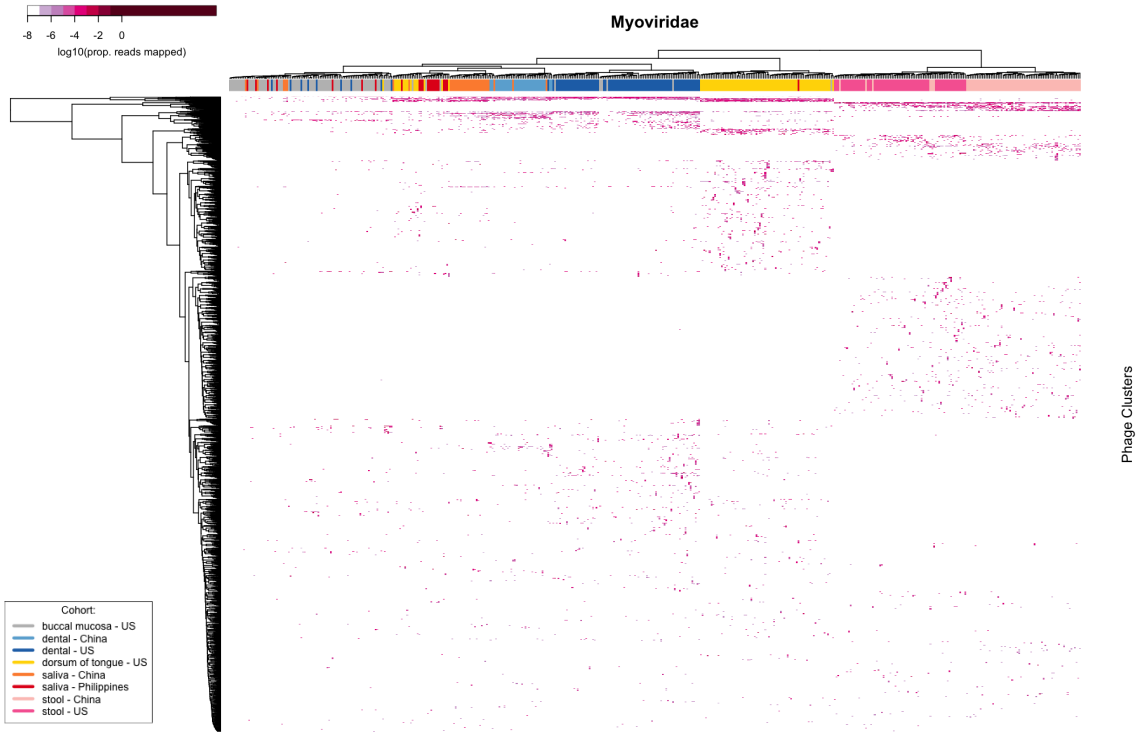

Figure 3Gd. Heatmap of log10 relative abundance of phage clusters for the *Myoviridae* phage family.

Dendrograms represent hierarchical clustering of samples (excluding USA longitudinal) colour coded by GIT site and country in the x axis and phage clusters in the y axis.

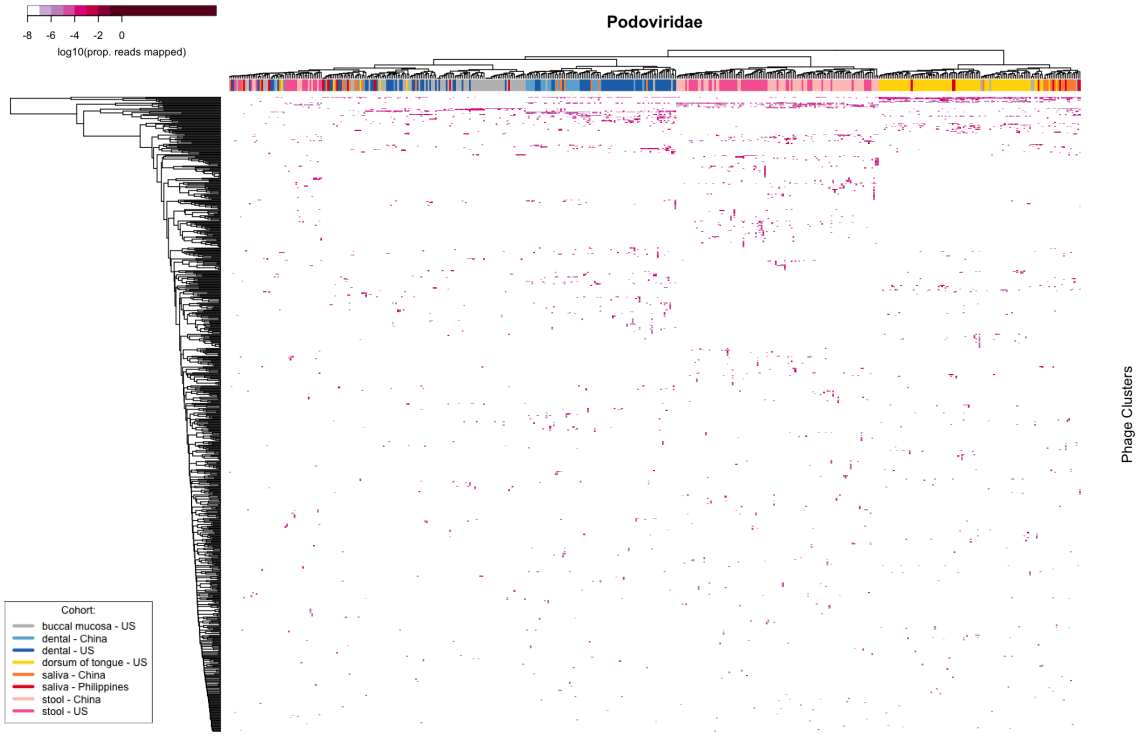

Figure 3Ge. Heatmap of log10 relative abundance of phage clusters for the *Podoviridae* phage family.

Dendrograms represent hierarchical clustering of samples (excluding USA longitudinal) colour coded by GIT site and country in the x axis and phage clusters in the y axis.

Figure 3Gf. Heatmap of log10 relative abundance of phage clusters for the *Siphoviridae* phage family.

Dendrograms represent hierarchical clustering of samples (excluding USA longitudinal) colour coded by GIT site and country in the x axis and phage clusters in the y axis.

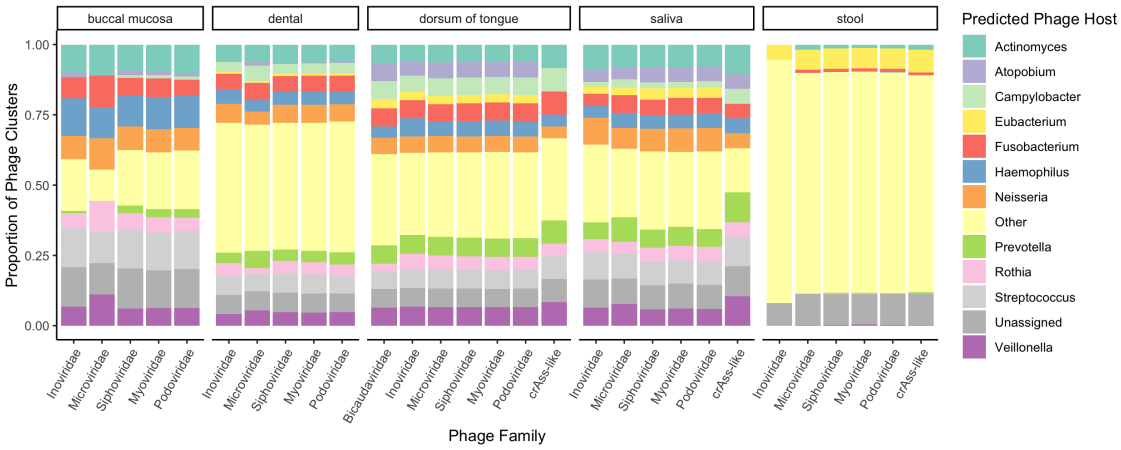

Figure 3H. Proportion of phage clusters with predicted phage hosts for each phage family and GIT site.

Buccal mucosa (n = 87 from the USA), dorsum of tongue (n = 90 from the USA), dental plaque (n = 90 from the USA and n = 32 from China), saliva (n = 33 from China and n = 24 from the Philippines) and stool (n = 70 from the USA and n = 72 from China) (Excluding longitudinal USA samples).

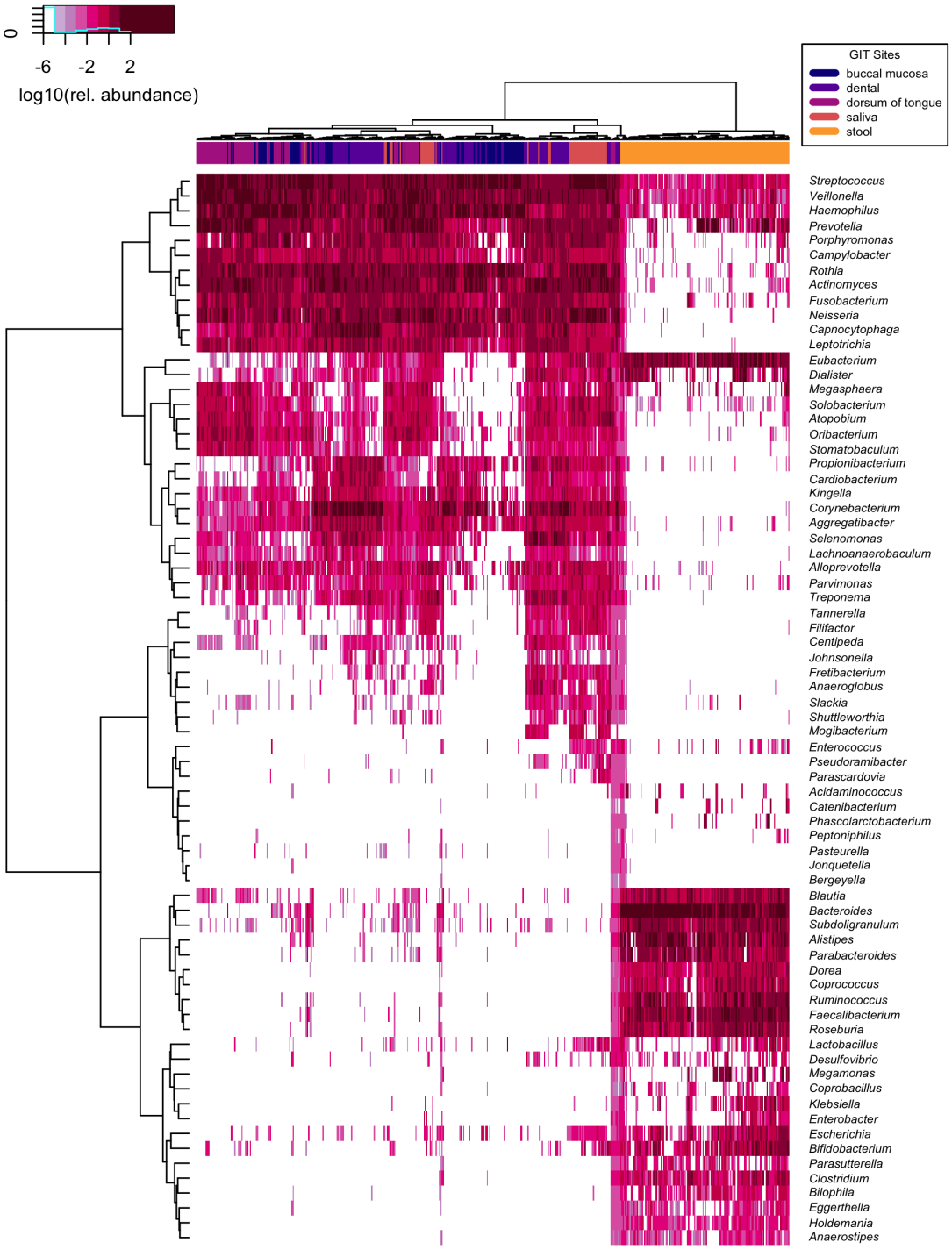

Figure 3I. Heatmap of log10 relative abundance of bacterial genera.

Dendrograms represent hierarchical clustering of genera in the y axis and samples in the x axis colour coded by GIT sites: buccal mucosa (n = 87 from the USA), dorsum of tongue (n = 90 from the USA), dental plaque (n = 90 from the USA and n = 32 from China), saliva (n = 33 from China and n = 24 from the Philippines) and stool (n = 70 from the USA and n = 72 from China). (Excluding longitudinal USA samples).

Figure 3J. Heatmap of log10 relative abundance of *Eubacterium*, *Haemophilus*, *Prevotella*, *Streptococcus* and *Veillonella* species.

Dendrograms represent hierarchical clustering of genera in the y axis and samples in the x axis colour coded by GIT sites: buccal mucosa (n = 87 from the USA), dorsum of tongue (n = 90 from the USA), dental plaque (n = 90 from the USA and n = 32 from China), saliva (n = 33 from China and n = 24 from the Philippines) and stool (n = 70 from the USA and n = 72 from China). (Excluding longitudinal USA samples).

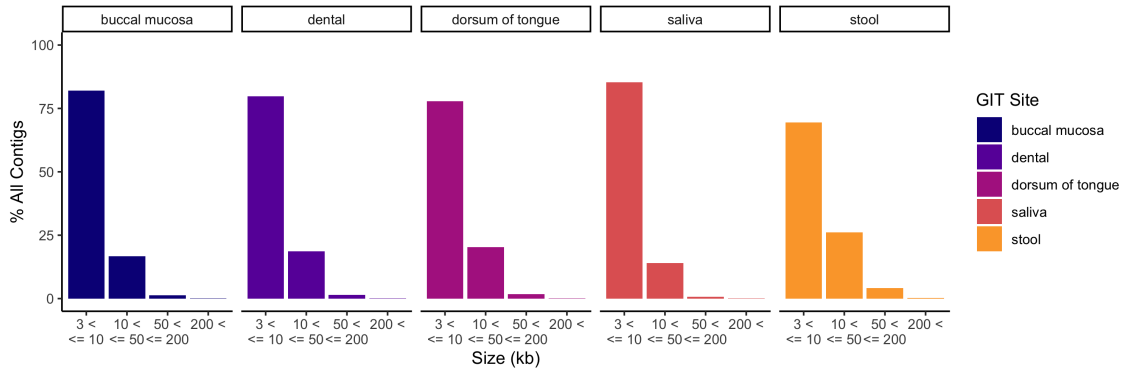

Figure 3K. Percentage of contigs with contig size for each GIT site.

Percentage of all contigs (phage and non-phage) that have genome sizes between 3 – 10, 10 – 50, 50 – 200 and over 200 kb for each GIT site.

Glossary

α-diversity: The α-diversity measures the variation of taxonomic composition within a sample.

β-diversity: The β-diversity measures the variation of taxonomic composition between samples, i.e. the ratio between regional and local diversity.

breadth of coverage: The breadth of coverage (also known as coverage) is the percentage of bases of the reference sequence that are covered by reads (not to be confused with coverage depth).

coverage depth: See read depth.

e-value: The e-value is the number of expected hits of similar quality that could be found just by chance. The smaller the e-value, the more significant the match.

homologous sequence: Homologous sequences are similar sequences that are related by evolutionary changes from a common ancestral sequence.

Gram-positive/Gram-negative: Bacteria can be classified into two categories: either Gram-positive or Gram-negative. Gram-negative bacteria have an outer membrane outside their cell wall, whereas Gram-positive bacteria do not. This means Gram-positive bacteria are more susceptible to cell wall targeting by antibiotics that Gram-negative bacteria.

in vitro: Experiments conducted on parts of cells after they are disrupted (in microbiology).

in vivo: Experiments conducted in live isolated cells (in microbiology).

***k*-medoids:** *k*-medoids is a clustering algorithm where data points are grouped into *k* clusters with a specified *k* value. Clusters are partitioned to minimise the distance between points within a cluster and a designated data point as the centre of the cluster.

***k*-mer:** A *k*-mer is a small sequence of length *k*.

lysogenic cycle: Viral genetic material is injected into the host cell and is replicated. Bacteriophage genetic material can integrates into the host genome to become a prophage.

lytic cycle: A virus infections the cell, replicates and lyses the cell.

MEM: A MEM (maximal exact match) between two sequences represents the maximum length of residues (nucleotides or amino acids) that match exactly, and cannot be extended in either direction without allowing for a mismatch.

paired-end sequencing: Paired-end sequencing is a type of Illumina sequencing technology that sequences both ends of the fragments.

phylogeny: Phylogeny is the evolutionary relationship of genetic or physical characteristics between species.

prophage: Bacteriophage DNA integrated in chromosomal DNA or existing as an extra-chromosomal plasmid.

read depth: The read depth (also known as coverage depth) is the number of unique reads that contain a particular nucleotide of the represented sequence.

reverse complement: The reverse complement of a DNA sequence is the reverse strand that has complementary base pairs. The complementary base pair rule states DNA base pairs are always paired A – T (adenine – thymine) and C – G (cytosine – guanine). This means the reverse complement of GCATGGA is TCCATGC, for example.

rolling-circle replication: In rolling circle replication, the double-stranded DNA is nicked. The 3’ end of the unnicked DNA is elongated and the 5’ end strand is displaced. Once replication is complete, the displaced DNA circularises and the second strand is synthesised. The 3’ end and 5’ end of both strands represent the configuration of bonds between carbon atoms of the DNA pentose backbone.

single-nucleotide polymorphism (SNP): A single-nucleotide polymorphism is a substitution of a single nucleotide at a specific position in a genome.

single-read sequencing: Single-read sequencing is a type of Illumina sequencing that sequences fragments from only one end.

strain: A genetic variant or subtype of a species.

taxonomic classification: Based on a hierarchical taxonomic rank of living organisms: domain, kingdom, phylum, class, order, family, genus, species. This excludes viruses that do not belong to a domain, but following a similar ranking system.

temperate phage: A phage that can replicate using both lytic and lysogenic cycles.

virulent phage: A phage that can only replicate by the lytic cycle.

END