EXPLORING THE HUMAN GUT MICROBIOTA AS A RESERVOIR OF ANTIBIOTIC RESISTANCE GENES

by

GREGORY ELLIOTT McCALLUM

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

Institute of Microbiology and Infection
College of Medical and Dental Sciences
University of Birmingham
November 2022

ABSTRACT

The human gut harbours a complex microbial ecosystem, termed the gut microbiome, that includes hundreds of bacterial species. Whilst most bacteria in the human gut have a commensal or mutualistic relationship with their host, the gut microbiome can also act as a reservoir for antimicrobial resistance genes (ARGs). Collectively, these ARGs are known as the gut resistome. Recent decades have seen a rise in multidrug-resistant infections caused by opportunistic pathogens originating from the gut microbiome. There is thus a need to characterise which bacterial species carry and transfer ARGs in the gut. Here, I explored the human gut resistome using chromosome conformation capture (3C) techniques to link ARGs to their bacterial hosts. Metagenomic 3C was implemented on a human faecal sample, and an analysis of my own and published datasets from 3C-based gut microbiome studies revealed that short reads mapping to repetitive elements causes problematic noise during analysis of 3C data. A bioinformatic workflow named H-LARGe (Host-Linkage to Antimicrobial Resistance Genes) was developed to reduce the impact of this noise and successfully link ARGs to their hosts. Next, a derivative of 3C, called Hi-C, was performed on four human faecal samples. Analysis of the data using an updated version of the H-LARGe workflow indicated that ARGs, including clinically important multiresistance genes, were widespread in commensal species from the gut microbiota. Following Hi-C analysis, the hosts of several ARGs were cultured and whole-genome sequenced using both short- and long-read technologies. These data provided genomic context for the ARGs and offered insights into the limitations of using Hi-C to link ARGs to their host in complex metagenomic samples. This highlighted the complementarity of Hi-C and culture-based approaches to fully characterise the gut resistome.

For Bill

ACKNOWLEDGMENTS

First and foremost, I would like to thank my supervisors, Professor Willem van Schaik and Dr Sarah Kuehne, for all their unfailing support and guidance during my PhD. Willem has helped me grow as a person and a scientist throughout this project, and I am extremely grateful for the time he has dedicated to me.

I would also like to thank all members of the Van Schaik and McNally lab groups, past and present, for all of their helpful discussions in the office and assistance with both lab work and bioinformatics. In particular, I'd like to give a special mention to Dr Rob Moran, who has been not only a great mentor and source of knowledge throughout my PhD, but also a real friend. One of the highlights of the last few years has been running Contamination Club with Stan, Rob, and Ross, and I hope we can continue ContamClub for years to come.

Even more importantly, I am thankful for the wonderful trips to the Staff House pub every week, where I have got to know so many great friends. Being surrounded by such an amazing group of people has made my PhD experience so good, and I feel very privileged. Even during the peak of the COVID-19 pandemic, virtual "pub trips" on Zoom every week to catch up with everyone helped keep me sane. I have loved being part of this research group and cannot imagine a better team to have been a member of.

The "McSchaiks", May 2022

Thank you to my friends and family, especially my parents, Carole and David, who have supported me in every aspect of my life.

Finally, I would like to thank my incredible partner, Emma, for her encouragement, support, motivation, and love that has kept me going throughout my PhD, especially during the difficult times of the COVID-19 pandemic.

"Do, or do not. There is no try."

-Yoda

DECLARATION OF AUTHORSHIP

Gregory Elliott McCallum was the first author and main contributor of the publication:

McCallum, G. E., Rossiter, A. E., Quraishi, M. N., Iqbal, T. H., Kuehne, S. A., and Schaik, W. van (2022) ‘Noise reduction strategies in metagenomic chromosome confirmation capture to link antibiotic resistance genes to microbial hosts’, *bioRxiv*, p. 2022.11.05.514866. doi: 10.1101/2022.11.05.514866.

Results from this publication are presented in Chapter 3 of this thesis.

CONTENTS

LIST OF FIGURES

Appendix Figures:

LIST OF TABLES

Appendix Tables:

LIST OF ABBREVIATIONS

× g	times gravity
°C	degrees Celsius
µL	microlitre
µM	micromolar
3C	chromosome conformation capture
4C	circular chromosome conformation capture
AMR	antimicrobial resistance
ANI	average nucleotide identity
ANOVA	analysis of variance
ARG	antibiotic resistance gene
β	beta
BHI	brain heart infusion
BLAST	Basic Local Alignment Search Tool
bp	base pairs
CDI	*Clostridioides difficile* infection
cfu	colony forming units
CHi-C	capture Hi-C
CRISPR	clustered regularly interspaced short palindromic repeats
Ct	cycle threshold
CTn	conjugative transposon
DNA	deoxyribonucleic acid
DTT	dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
epicPCR	Emulsion, Paired Isolation, and Concatenation PCR
ESBL	extended spectrum beta-lactamase
ExPEC	extraintestinal pathogenic *Escherichia coli*
g	gram(s)
GAM	Gifu Anaerobic Medium
GC	guanine-cytosine
GTDB	Genome Taxonomy Database
GTDB-Tk	Genome Taxonomy Database Toolkit
h	hour(s)
HAM-ART	Hi-C Assisted Metagenomics for Antimicrobial Resistance Tracking
HGT	horizontal gene transfer
H-LARGe	Host-Linkage to Antimicrobial Resistance Genes
ICE	integrative and conjugative element
IME	integrative and mobilizable element
ImmeDB	Intestinal microbiome mobile element database
IS	insertion sequence
kb	kilobase
kbp	kilobase pairs
LB	lysogeny broth
LCA	lowest common ancestor
M	molar
MAG	metagenome-assembled genome

MAMP	microorganism-associated molecular pattern
MAPQ	mapping quality
Mbp	megabase pair
MDR	multidrug-resistant
meta3C	metagenomic chromosome conformation capture
mg	milligram
mGAM	modified Gifu Anaerobic Medium
MGE	mobile genetic element
min	minutes
mL	millilitre
MLST	multilocus sequence typing
mM	millimolar
NCBI	National Centre for Biotechnology Information
ng	nanogram
NGS	next-generation sequencing
nM	nanomolar
nt	nucleotide(s)
OIL-PCR	One-step Isolation and Lysis PCR
PBS	phosphate-buffered saline
PCR	polymerase chain reaction
qPCR	quantitative polymerase chain reaction
R^2	R-squared
RefSeq	reference sequence
RNA	ribonucleic acid
RPKM	reads per kilobase per million mapped reads
rpm	revolutions per minute
rRNA	ribosomal RNA
RT	room temperature
s	second(s)
SCFA	short-chain fatty acid
SDS	sodium dodecyl sulphate
sp.	species
spp.	species (plural i.e. several species)
SRA	short read archive
ST	sequence type
TBE	Tris/Borate/EDTA
TE	Tris/EDTA
Tm	melting temperature
U	units
UHGG	Unified Human Gastrointestinal Genome
UTI	urinary tract infection
v	version
v/v	volume/volume
w/v	weight/volume
WGS	whole-genome sequencing
YCFA	yeast extract, casein hydrolysate, fatty acids

CHAPTER 1
INTRODUCTION

1.1 A short history of antibiotics

The serendipitous discovery of penicillin kickstarted the “Golden Age” of the discovery and development of drugs that kill or inhibit the growth of bacteria, known as antibiotics (Hutchings, Truman, and Wilkinson, 2019). On returning home from a holiday in September 1928, Alexander Fleming noticed a fungal contaminant growing on a culture plate of *Staphylococcus aureus* he had left on a windowsill. The fungal colony was surrounded by bacteria-free zones, and Fleming determined that the bacteria were being killed by a secretion of “mould juice”, which he later termed “penicillin” after the *Penicillium* mould that produced it (Fleming, 1929). It took another 12 years until penicillin was purified by a team in Oxford led by Howard Florey, Ernst Chain, and Norman Heatley (Chain *et al.*, 1940), leading to mass production and extensive use in World War II (Figure 1.1) (Gaynes, 2017). By 1945, penicillin was widely distributed (Aminov, 2010), and is still used today having saved countless lives worldwide.

Penicillin was not the first antibiotic to be employed for medical treatment of infection. The use of antibiotic-producing microbes to remedy disease dates back millennia, such as ancient Egyptians applying mouldy bread to infected wounds, documented in The Ebers Papyrus from 1550 BC (Haas, 1999; Hutchings, Truman, and Wilkinson, 2019). In more recent history, the first modern antimicrobial drug was the arsenic-based, synthetic compound, Salvarsan, developed by German physician Paul Ehrlich in 1909 to treat syphilis (Figure 1.1) (Gould, 2016). Salvarsan was followed by Prontosil in 1935 (Figure 1.1), a synthetic pro-drug developed by Gerhard Domagk that metabolised to sulphanilamide. This began the era of sulphonamide drugs, a class of effective, broad-spectrum, antibacterial drugs that were used worldwide to treat infections caused by Gram-positive bacteria, particularly streptococcal infections (Shambaugh, 1966). These drugs were largely superseded by penicillin once it became widely available a decade later, which commenced a 20-year golden age of natural product antibiotic discovery (Hutchings, Truman, and Wilkinson, 2019). Between 1940-1960, research groups around the world searched for novel antibiotics that, like penicillin, were naturally produced by other microorganisms (Figure 1.1). During this period, the majority of antibiotic classes in clinical use today were discovered (Aminov, 2010). These discoveries meant that previously fatal infections could be treated and cured, revolutionising modern medicine. However, it soon became apparent that resistance to antibiotics posed a threat (Gould, 2016).

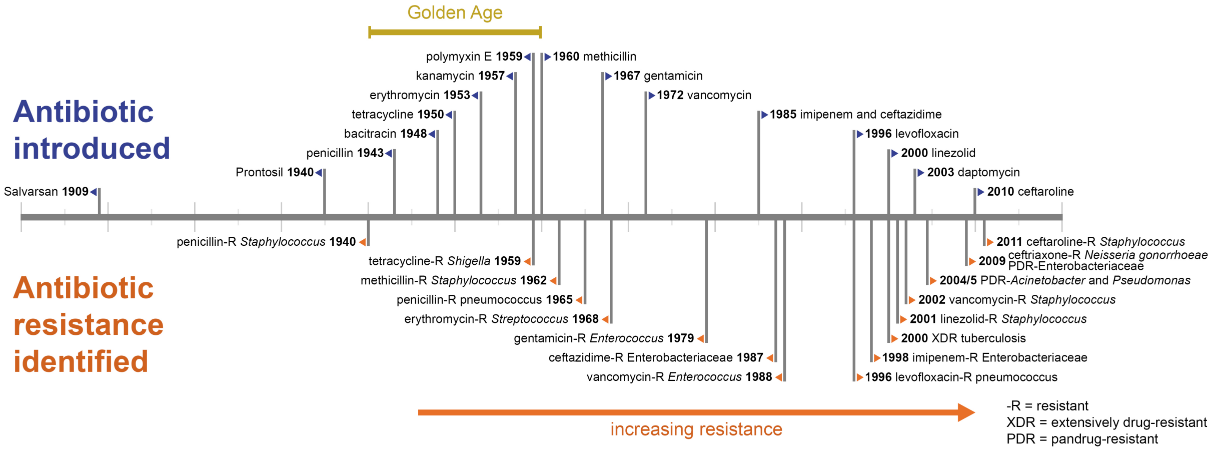

Golden Age
Antibiotic introduced
Salvarsan 1909
Prontosil 1940
penicillin 1943
bacitracin 1948
tetracycline 1950
erythromycin 1953
kanamycin 1957
polymyxin E 1959
1960 methicillin
1967 gentamicin
1972 vancomycin
1985 imipenem and ceftazidime
1996 levofloxacin
2000 linezolid
2003 daptomycin
2010 ceftaroline
Antibiotic resistance identified
penicillin-R *Staphylococcus* 1940
tetracycline-R *Shigella* 1959
methicillin-R *Staphylococcus* 1962
penicillin-R pneumococcus 1965
erythromycin-R *Streptococcus* 1968
gentamicin-R *Enterococcus* 1979
ceftazidime-R Enterobacteriaceae 1987
vancomycin-R *Enterococcus* 1988
1996 levofloxacin-R pneumococcus
1998 imipenem-R Enterobacteriaceae
2000 XDR tuberculosis
2001 linezolid-R *Staphylococcus*
2002 vancomycin-R *Staphylococcus*
2004/5 PDR-*Acinetobacter* and *Pseudomonas*
2009 ceftriaxone-R *Neisseria gonorrhoeae* PDR-Enterobacteriaceae
2011 ceftaroline-R *Staphylococcus*
increasing resistance
-R = resistant
XDR = extensively drug-resistant
PDR = pandrug-resistant

Figure 1.1. Timeline of notable antibiotics and resistance.
Timeline from 1900-2020. Top half highlights years that notable antibiotics were introduced. Bottom half highlights notable emergence of resistance. Period between 1940-1960 is highlighted as the golden age of natural product antibiotic discovery. Figure recreated and modified from (Centers for Disease Control and Prevention, 2013).

1.2 Antimicrobial resistance

The extensive use of antibiotics and other antimicrobial compounds since their introduction, and the overuse and misuse that followed, has resulted in the emergence of antimicrobial resistance (AMR) (Aminov, 2010). AMR describes the resistance of parasites, viruses, fungi, and bacteria to antimicrobial drugs (Prestinaci, Pezzotti, and Pantosti, 2015). The rise of AMR over the last century has led it to be one of the greatest global health threats facing humanity (World Health Organization, 2020). Very few new antibiotics are being developed, especially ones that are effective against critical-priority pathogens (Figure 1.1) (Theuretzbacher *et al.*, 2020), worsening the future outlook of the AMR crisis. As well as the treatment of life-threatening infections, antibiotics are vital in many areas of medicine, such as prophylactic use during invasive surgery or in immunocompromised patients (Blair *et al.*, 2015), and AMR may jeopardise all of these uses. It is estimated that over 1.27 million deaths globally in 2019 can be directly attributed to bacterial AMR (Murray *et al.*, 2022), and in Europe alone, 33,000 people die annually from antibiotic resistant infections (Cassini *et al.*, 2019). A seminal report into the global burden of AMR reported that by 2050, AMR could be attributed to 10 million deaths annually (O'Neill, 2016), although this is a worst-case scenario. In addition to these mortality statistics, AMR also places a significant burden on the economy, with multidrug-resistant (MDR) infections costing the European Union economy over €1.5 billion each year (ECDC, 2009). It is projected that by 2050, the global annual cost of AMR could reach as high as $300 billion-$1 trillion (Dadgostar, 2019). These figures highlight the severity of AMR affecting all countries, although it is important to note that the burden of AMR is, and will continue to be, disproportionately higher in low- and middle-income countries (Pokharel, Raut, and Adhikari, 2019).

The rise of AMR can be attributed to various human factors. Exposure to antimicrobial drugs, even if being appropriately used for medical treatment, will select for resistant microbes. However, whilst antibiotics remain vital for medical use, the abuse and overprescription of antibiotics has massively increased this selective pressure on bacteria (Viswanathan, 2014). Non-human use of antibiotics has also hugely added to the AMR crisis. Antibiotics are frequently used in veterinary medicine to treat infection in animals, often using the same antibiotic classes as human medicine (Meek, Vyas, and Piddock, 2015). In addition, antibiotics have been extensively used in animals for non-medical purposes. For example, the agriculture and aquaculture industry regularly supplement livestock feed with antibiotics, often at sub-inhibitory concentrations, for growth promotion and prevention of disease outbreak due to the poor sanitary conditions and overcrowding of farm animals (Manyi-Loh *et al.*, 2018). Domestic use of antimicrobial agents, such as cleaning products and soaps containing the biocide triclosan, also exacerbates the AMR crisis by selecting for biocide-resistant strains that also lose susceptibility to clinically used antibiotics (Webber *et al.*, 2015). Resistant strains of bacteria arise from all of these factors due to a variety of mechanisms (Darby *et al.*, 2022).

1.2.1 Mechanisms of resistance

Various classes of antibiotics exist to kill or stop the growth of bacteria using different mechanisms. Bacteria can be intrinsically resistant to certain classes of antibiotic. For instance, the presence of the outer membrane in Gram-negative bacteria naturally prevents vancomycin from penetrating the cell (Exner *et al.*, 2017). Furthermore, due to their rapid replication rate allowing for frequent mutations in their DNA, bacteria can evolve to become resistant to antibiotics through various mechanisms (Figure 1.2). For example, these mutations can result in genes encoding for major porins to be downregulated, resulting in decreased membrane permeability, or the overexpression of efflux pumps to remove antibiotics from the cell after entry before a lethal concentration is reached (Figure 1.2) (Blair *et al.*, 2015). Various types of efflux systems exist, and these can result in multidrug-resistance (Webber *et al.*, 2015). Another resistance mechanism utilised by bacteria is modifying the structure of drug targets to prevent antibiotic-target binding (Figure 1.2). Resistance to glycopeptides, like vancomycin, involves modification of part of the peptidoglycan cell wall synthesis pathway caused by mutations that substitute D-alanine-D-alanine to D-alanine-D-lactate or D-alanine-D-serine. These mutations significantly reduce the binding affinity of glycopeptide antibiotics to the cell wall precursors, resulting in resistance to the drugs (Ahmed and Baptiste, 2018). Bacteria can also evolve to inactivate antibiotics (Figure 1.2). In 1940, even before extensive use of penicillin, enzymatic degradation of penicillin had been observed (Figure 1.1) (Abraham and Chain, 1940), although the extent of this problem was not realised until years later (Aminov, 2010). These antibiotic-modifying enzymes, known as beta (β)-lactamases, hydrolyse penicillins, resulting in the inability to bind with the drug target-site (Blair *et al.*, 2015). The overuse of β-lactam antibiotics has led to β-lactam resistance genes becoming widespread globally (Castanheira, Simner, and Bradford, 2021). Importantly, as well as the ability to gain adaptive resistance through chromosomal mutations, bacteria are also capable of disseminating antibiotic resistance genes (ARGs) through horizontal gene transfer (HGT) (Soucy, Huang, and Gogarten, 2015).

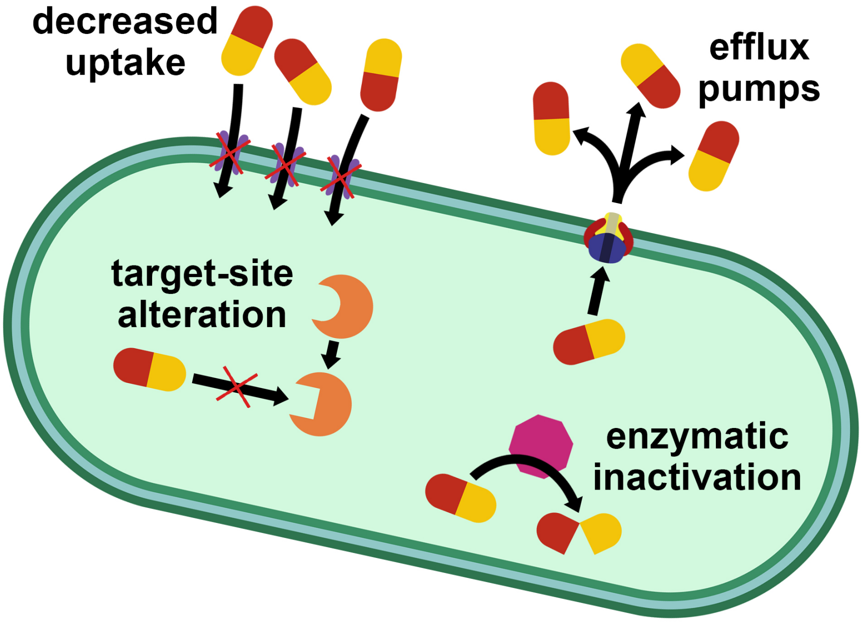


Figure 1.2. Mechanisms of antibiotic resistance.
Bacteria can gain resistance to antibiotics through various mechanisms. These include decreased uptake or accumulation of antibiotics through down regulation of porins or increased expression of efflux pumps to transport antibiotics out of the cell. Alteration of drug target-sites causes resistance by preventing antibiotic-target binding. Bacteria can also confer resistance by producing enzymes to inactivate the antibiotics and prevent them from working. These resistance mechanisms can be acquired through mutations in the genome, or acquisition of antimicrobial resistance genes via horizontal gene transfer.

1.2.2 Horizontal gene transfer

Vertical gene transfer occurs during replication when bacteria pass on genetic material to daughter cells. HGT describes the transfer of genetic material between organisms in the same generation. Multiple mechanisms of HGT exist. Firstly, physiologically competent bacteria can uptake extracellular DNA from their environment in a process known as transformation (Figure 1.3) (Johnston *et al.*, 2014). Bacteria must be in a state of competence in order to take up naked DNA via transformation. Natural competence has been documented in over 80 species of bacteria (Johnston *et al.*, 2014), and the transfer of ARGs mediated by natural transformation has been observed in several clinically important pathogens including *Streptococcus pneumoniae* (Janoir *et al.*, 1999) and *Neisseria meningitidis* (Bowler *et al.*, 1994).

Another mechanism, transduction, is the transfer of genetic material between bacteria mediated by viral intermediates known as bacteriophages (Figure 1.3). Bacteriophages are viruses that infect bacteria by inserting their DNA into the cell which then integrates into the bacterial genome where it undergoes replication and repackaging into new phages (Clokie *et al.*, 2011). During this process, bacterial DNA can be packaged into phages and subsequently transferred. There are several modes of transduction, with generalised and specialised being the predominant mechanisms (Soucy, Huang, and Gogarten, 2015). Generalised transduction is caused by mispackaging of DNA during the lytic cycle, resulting in sections of bacterial chromosomal or extrachromosomal DNA being incorporated into the phage capsid (Colavecchio *et al.*, 2017). Specialised transduction occurs during the excision of prophages from the bacterial chromosome when the chromosomal regions flanking the prophage are excised and packaged into the phage capsid along with the phage DNA (Colavecchio *et al.*, 2017). More recently, a third mechanism called lateral transduction has been described (Chen *et al.*, 2018). During lateral transduction, prophages enter replication whilst still integrated into the bacterial chromosome, generating multiple copies of phage DNA in the host genome. When excision then occurs, large regions of chromosomal DNA adjacent to the prophage, up to several hundred kilobases long, is packaged into the phage capsid along with the phage DNA (Chiang, Penadés, and Chen, 2019). During any of these mechanisms, any bacterial DNA packaged into the phage capsid will be horizontally transferred when the phage infects another cell and can then integrate into the recipient chromosome via homologous recombination. If the region of transferred DNA encodes an ARG, then the resistance gene will be transferred to the recipient bacterial cell (Colavecchio *et al.*, 2017)

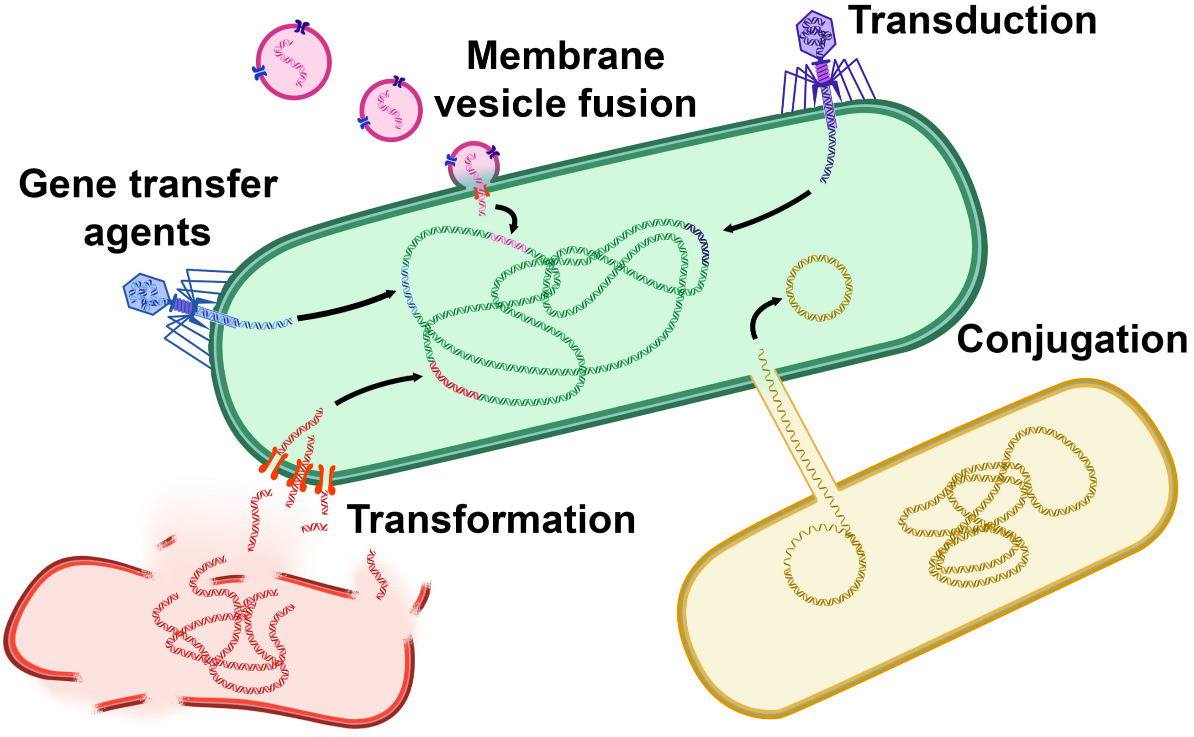


Figure 1.3. Mechanisms of horizontal gene transfer.
During transformation, naked DNA in the environment is taken up by competent bacteria. Transduction is the transfer of genetic material mediated by bacteriophages. Gene transfer agents are similar to bacteriophages but do not encode any of their own machinery. For membrane vesicle fusion, spherical structures composed of a lipid bilayer encompassing various cargo, including DNA, can fuse to recipient bacteria and transfer genetic material. During conjugation, plasmids and other mobile genetic elements transfer via a pilus formed during cell-cell contact. A modified, earlier version of this figure was published in (McInnes and McCallum *et al.*, 2020).

HGT can also be mediated by gene transfer agents, structures produced by bacteria that resemble bacteriophages (Figure 1.3). Unlike bacteriophages, gene transfer agents do not carry DNA that encodes their own machinery, and instead only carry fragments of their host cell genome (Von Wintersdorff *et al.*, 2016). The importance of gene transfer agents for the horizontal transfer of ARGs is not yet established, though several studies have demonstrated ARG transfer by gene transfer agents *in vitro* (Wall, Weaver, and Gest, 1975; Stanton *et al.*, 2008).

Membrane vesicle-mediated gene transfer has more recently been recognised as a mechanism for HGT of ARGs (Figure 1.3). Membrane vesicles are spherical structures with diameters between 10-500 nm that consist of a lumen surrounded by lipid-bilayers (Domingues and Nielsen, 2017). These vesicles are released by bacteria and can contain various components including DNA and RNA (Dell'annunziata *et al.*, 2021). Like gene transfer agents, the role membrane vesicles play in HGT of ARGs is understudied, although vesicle-mediated transfer of β-lactamase genes has been observed *in vitro* in clinical isolates of *Acinetobacter baumannii* (Rumbo *et al.*, 2011; Chatterjee *et al.*, 2017).

The final major mechanism of HGT is conjugation, where mobile genetic elements (MGEs) transfer during physical cell-cell contact by the formation of a conjugation pilus that single-stranded DNA can be passed through using a type IV secretion system (Figure 1.3) (Cabezón, de la Cruz, and Arechaga, 2017). MGEs include small, circular, extrachromosomal DNA molecules known as plasmids, which often carry ARGs, as well as elements that may be integrated into the chromosome such as integrative and conjugative elements (ICEs) (Frost *et al.*, 2005), or integrative and mobilizable elements (IMEs) (Guédon *et al.*, 2017). Plasmids are made up of a backbone of core genes controlling functions essential for the plasmid such as replication and mobilisation. Plasmids also often have a variety of nonessential genes that provide additional functions and fitness advantage to the bacterial host of the plasmid, known as accessory genes. As well as ARGs, plasmid accessory genes can include genes conferring virulence or resistance to heavy metals (Harrison and Brockhurst, 2012). Conjugative plasmids are also able to mediate transfer of DNA that is not self-transmissible, such as non-conjugative plasmids (Ramsay and Firth, 2017), or even chromosomal DNA that is mobilised and integrated into the plasmid by short, transposable DNA sequences called insertion sequence (IS) elements (Ebmeyer, Kristiansson, and Larsson, 2021). Conjugation of AMR plasmids allows the rapid dissemination of ARGs, which can cause major hospital outbreaks of MDR infections (Martin *et al.*, 2017; Phan *et al.*, 2018), and fast global emergence of novel ARGs (Lee *et al.*, 2016).

Resistance to antibiotics is present in all microbial ecosystems. Outside of clinical settings, HGT of ARGs frequently occurs in environmental habitats such as wastewater (Kizny Gordon *et al.*, 2017; Che *et al.*, 2019) and the soil microcosm (Aminov, 2011). Crucially, these complex microbial communities may act as a reservoir for ARGs (Kizny Gordon *et al.*, 2017), which can then facilitate outbreaks of MDR infections in humans (Breathnach *et al.*, 2012; Weingarten *et al.*, 2018). One of the largest and most diverse of these complex microbial ecosystems is the human gut microbiome (Thursby and Juge, 2017).

1.3 The human gut microbiome

The term ‘microbiome’ describes all microorganisms, genomes, and environmental conditions of a microbial habitat. The microorganisms present in a microbiome are collectively termed the ‘microbiota’ (Marchesi and Ravel, 2015). The human body is colonised by trillions of microorganisms that peacefully coexist in and on the host in various microbiomes. The gastrointestinal tract contains the largest population of bacteria inside humans that make up an ecosystem known as the human gut microbiota (Sekirov *et al.*, 2010). Recent years have seen a huge increase in research on the gut microbiome, aiming to uncover its composition, function, and roles in health and disease.

1.3.1 The human gut microbiome in health

The human gut microbiota typically harbours hundreds of bacterial species. The vast majority of these have a commensal or mutualistic relationship with their human host, whereby both sides benefit (Schluter and Foster, 2012). Bacteria from the phyla Firmicutes and Bacteroidetes account for around 90% of species in a healthy adult gut microbiota (Eckburg *et al.*, 2005). Other less abundant, but frequently detected, phyla of the intestinal microbial flora include Actinobacteria, Fusobacteria, Proteobacteria, and Verrucomicrobia (Rajilić-Stojanović and de Vos, 2014). Generally, it is apparent that these commensal bacteria play important key roles for human health (Figure 1.4). The gut microbiota serve an integral role in digestion by breaking down and metabolising substrates that would otherwise be ingestible by the host alone (Figure 1.4) (Shreiner, Kao, and Young, 2015; Oliphant and Allen-Vercoe, 2019). This not only aids digestion, providing essential nutrients and vitamins to the host, but also helps modulate the immune system. One of the major beneficial metabolites produced from bacterial fermentation of indigestible foods in the gut are short-chain fatty acids (SCFAs) (Figure 1.4) (Martín *et al.*, 2013). These include butyrate, an immunoregulatory molecule that reduces inflammation in the gut (Böcker *et al.*, 2003), and acetate, produced by commensal *Bifidobacterium* species, which is shown to provide protection against infection by *Escherichia coli* O157:H7 in mice (Fukuda *et al.*, 2011). As well as immunomodulation provided by SCFAs, the gut microbiota provide fundamental aid during development of the immune system (Di Tommaso, Gasbarrini, and Ponziani, 2021). Gut bacteria supply the immune system with highly conserved bacterial molecular signatures termed microorganism-associated molecular patterns (MAMPs) which stimulate innate immune cells, heightening future immune response when exposed to bacterial pathogens (Figure 1.4) (Negi *et al.*, 2019). The largest constituent of the human immune system, termed the mucosal immune system, is also aided by gut microbes as their presence strengthens the intestinal barrier by promoting mucus secretion (Petersson *et al.*, 2011). The gut microbiota also help prevent colonisation by pathogens through production of antimicrobial peptides and creating competition for nutrients (Figure 1.4) (Jandhyala *et al.*, 2015; Ducarmon *et al.*, 2019; Djukovic *et al.*, 2022).

1.3.2 The human gut microbiome in disease

Although largely serving to benefit human health, the gut microbiota can also contribute to disease (Figure 1.4). An imbalance in the gut microbiota caused by changes in abundance of the community members, known as dysbiosis, can result in illness (Figure 1.4) (Martín *et al.*, 2013). This can occur as a result of various exposures, including certain diets, medication, or ingesting pathogenic microbes (Carding *et al.*, 2015). Dysbiosis has been linked to many disorders such as inflammatory bowel disease (Frank *et al.*, 2007), colorectal cancer (Buc *et al.*, 2013), and even to conditions that have no obvious links to the gut, such as the immune-mediated skin disease psoriasis (Zhang *et al.*, 2021). These diseases can be a consequence of increased abundance of certain bacterial species or a reduction in immunomodulation due to the depletion of beneficial commensal bacteria (Martín *et al.*, 2013). For example, the reduced abundance of butyrate-producing gut commensal species, such as *Faecalibacterium prausnitzii*, has been strongly linked to Crohn’s disease by several studies (Cao, Shen, and Ran, 2014; Takahashi *et al.*, 2016).

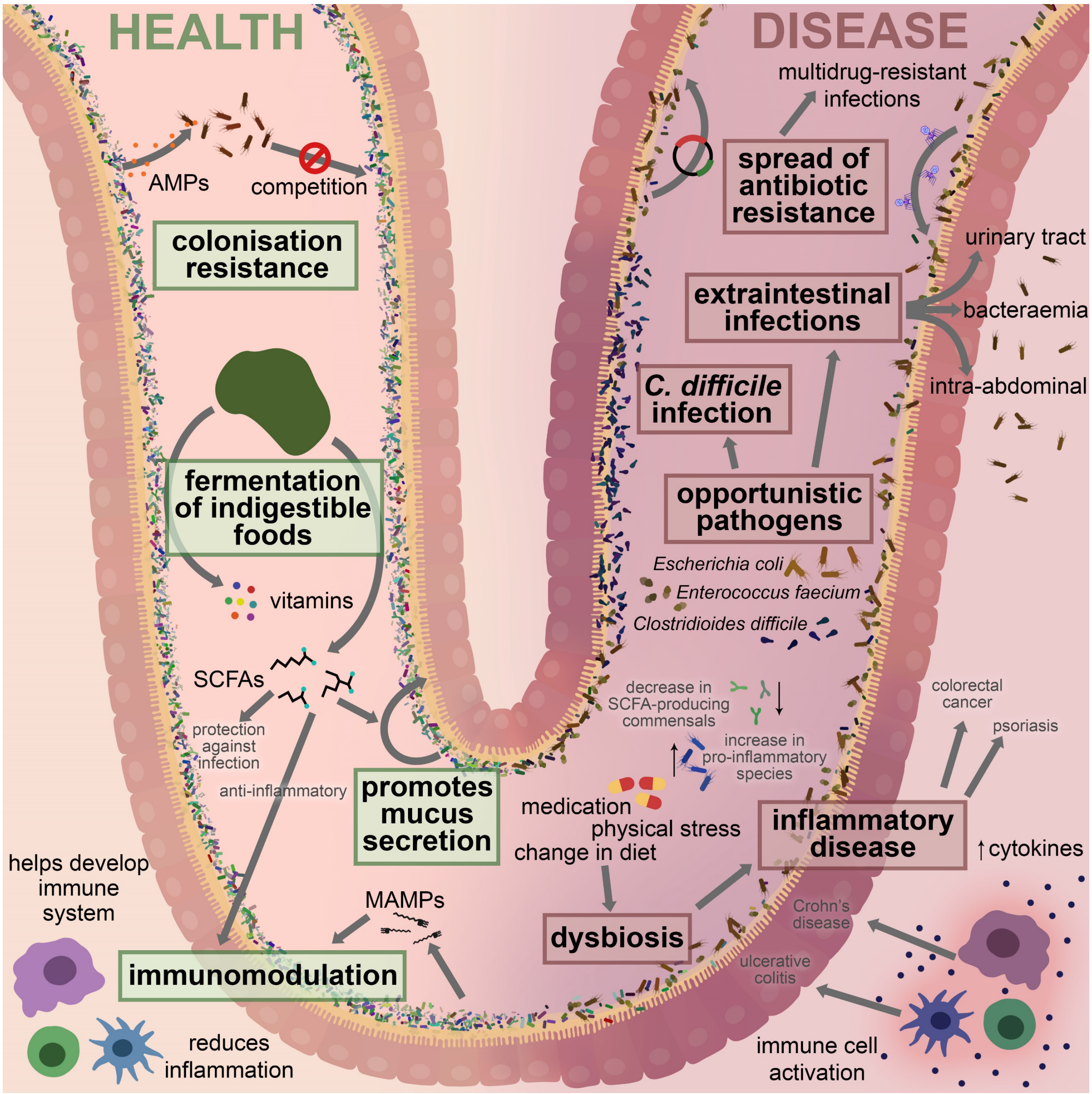


Figure 1.4. The role of the gut microbiota in health and disease.
Gut bacteria play a role in both human health (left) and disease (right). Gut commensals prevent colonisation by pathogenic bacteria through production of antimicrobial peptides (AMPs) and competition for nutrients. Indigestible foods are fermented by gut microbes to produce vitamins and short-chain fatty acids (SCFAs) that have immunoregulatory roles. The development of the immune system is aided by microorganism-associated molecular patterns (MAMPs). Dysbiosis of the gut microbiota can lead to a pro-inflammatory environment. The gut also harbours opportunistic pathogens that can cause infection. Antimicrobial resistance genes spread in the gut via horizontal gene transfer.

1.3.2.1 Opportunistic pathogens of the human gut microbiota

Some bacterial species that inhabit the gut are known as opportunistic pathogens. These are species that would not normally harm a healthy individual, but if the host becomes debilitated or immunocompromised, then the bacteria can take advantage and start to cause infections in these individuals (Price *et al.*, 2017). The most prominent opportunistic pathogens originating from the gut are species belonging to the family Enterobacteriaceae such as *Klebsiella pneumoniae* and *Escherichia coli*, as well as from the phylum Firmicutes such as *Enterococcus* species (spp.) and *Clostridioides difficile* . Opportunistic pathogens can have pro-inflammatory properties, such as strains of adhesion-invasive *E. coli*, which are associated with Crohn's disease (Darfeuille-Michaud *et al.*, 2004). Importantly, in addition to exacerbating inflammatory disease, opportunistic pathogens can cause infection (Figure 1.4).

Individuals that carry opportunistic pathogens in their gut are at increased risk of urinary tract infections (UTIs), as well as more severe infections such as bacteraemia (Figure 1.4). The most common cause of UTIs in humans is extraintestinal pathogenic *E. coli* (ExPEC) (Poolman and Wacker, 2016). These are strains of *E. coli* colonising the gut that possess virulence traits enabling them to invade sites outside of the intestinal tract, and are the leading cause of extraintestinal infections in humans worldwide (Smith, Fratamico, and Gunther, 2007; Manges *et al.*, 2019). Carriage of ExPEC strains predisposes individuals, largely women, to UTIs (Tchesnokova *et al.*, 2020).

Gut colonisation by opportunistic pathogens can be particularly problematic during treatment with antibiotics. The use of antibiotics to treat an infection is a common cause of dysbiosis in the gut, as the antibiotics inadvertently kill susceptible bacteria residing in the gut. This perturbation of the gut microbiota can allow resistant strains of gut-dwelling opportunistic pathogens to take over the microbiota, and often leads to hospital-acquired infection in intensive care unit patients who are colonised with opportunistic pathogens upon admission (Mehrad *et al.*, 2015; Gorrie *et al.*, 2017; Raplee *et al.*, 2021). Antibiotic use also increases risk of future UTI with antibiotic resistant opportunistic pathogens originating from the gut (Stracy *et al.*, 2022). Additionally, if a patient carries the opportunistic pathogen *C. difficile* prior to antibiotic treatment, antibiotic-caused disruption of the gut microbiota risks *C. difficile* infection (CDI), which can lead to severe colitis and possibly death (Mullish and Williams, 2018). The loss of commensal bacteria that protect against colonisation and infection allows the overgrowth of subdominant strains of *C. difficile* in the months following cessation of treatment (Hensgens *et al.*, 2012; Mullish and Williams, 2018). These infections can be difficult to treat and can commonly become recurrent. Recently, transfer of faecal matter from a healthy donor into the intestinal tract of a recipient, termed faecal microbiota transplantation, has been recognised as an effective treatment for recurrent CDI, with higher treatment success compared to antibiotics (Baunwall *et al.*, 2020).

Crucially, opportunistic pathogens in the gut frequently carry ARGs, and there has been a considerable rise of MDR nosocomial pathogens originating from the gut in recent decades (Vincent, 2003; Mehrad *et al.*, 2015; Guzman Prieto *et al.*, 2016). The majority of multidrug-resistance in Enterobacteriaceae can be attributed to the acquisition of plasmids carrying genes encoding for extended spectrum β-lactamases (ESBLs) (Coque, Baquero, and Canton, 2008). ESBLs are enzymes that are able to hydrolyse, and thus confer resistance to, most β-lactam antibiotics including penicillins, aztreonam, and cephalosporins (Castanheira, Simner, and Bradford, 2021). Plasmids that carry ESBL genes often also harbour other ARGs conferring resistance against other antibiotic classes including tetracyclines, aminoglycosides, and fluroquinolones (Nordmann, Dortet, and Poirel, 2012). Strains of MDR Enterobacteriaceae are now present worldwide, and the acquisition of intestinal MDR Enterobacteriaceae has been associated with global travel (van der Bij and Pitout, 2012; Arcilla *et al.*, 2017). A recent study showed that resident strains of *Bacteroides* and *Citrobacter* in an individual's gut microbiome pre-travel can provide colonisation resistance against MDR *E. coli* during travel (Davies *et al.*, 2022). Nevertheless, the spread of these MDR strains during travel poses serious health risks to travellers and has allowed MDR Enterobacteriaceae and associated AMR-plasmids to become widespread globally.

For opportunistic pathogens in the Enterococcaceae family such as *Enterococcus faecium* and *Enterococcus faecalis*, perhaps the most clinically concerning ARGs are those encoding for resistance to aminoglycoside and glycopeptide antibiotics (Miller, Munita, and Arias, 2014). Enterococci are intrinsically resistant to several antibiotics, including β-lactams, so acquisition of additional ARGs is causing the emergence of MDR clones of *Enterococcus* spp. that can cause life-threatening infections (Ahmed and Baptiste, 2018). Strains of *Enterococcus faecium* conferring high-level aminoglycoside resistance have become more prevalent worldwide in recent years (Diab *et al.*, 2019; Adamecz *et al.*, 2021). Although all *Enterococcus* strains have intrinsic low-level resistance to aminoglycosides, this high-level resistance is primarily achieved through acquisition of aminoglycoside resistance genes, and can be found on highly conjugative plasmids (Tanimoto and Ike, 2008). Resistance to the glycopeptide drug vancomycin has also become of serious concern as there has been a rapid rise in life-threatening vancomycin-resistant

Enterococcus infection and global spread (Markwart *et al.*, 2019). Vancomycin resistance is conferred by *van* operons often located on transposons that can mobilise and transfer between strains, including via plasmids (Palmer, Kos, and Gilmore, 2010; Arredondo-Alonso *et al.*, 2021). The rise of vancomycin-resistance *Enterococcus* has led to the need to use the relatively newly introduced antibiotic linezolid (Ahmed and Baptiste, 2018), for which resistance in *Enterococcus* spp. has rapidly emerged (Abbo *et al.*, 2019). Although prevalence of infection by *Enterococcus* resistant to both linezolid and vancomycin remains uncommon, the emergence of resistance to linezolid continues to increase (Bender *et al.*, 2018; Abbo *et al.*, 2019). This rise is, in part, owing to the horizontal transfer of ARGs conferring resistance to linezolid such as variants of the *cfr* gene and *optrA* (Lazaris *et al.*, 2017; Egan *et al.*, 2020).

Linezolid resistance is also rising in *C. difficile*, with an estimated 7% of strains harbouring *cfr*(C) (Candela *et al.*, 2017). Linezolid-resistant strains of *C. difficile* have also been found to harbour *cfr*(B) (Hansen and Vester, 2015), which has also been reported in *E. faecium* (Deshpande *et al.*, 2015). Cfr-type genes confer resistance to multiple antibiotics, with *cfr*(B) conferring resistance to drugs from the amphenicol, lincomycin, oxazolidinone, pleuromutilin, and streptogramin antibiotic classes (Hansen and Vester, 2015). Although these antibiotic classes are not currently used to treat CDI (Ooijevaar *et al.*, 2018; Sholeh *et al.*, 2020), the rise in these ARGs in clones of *C. difficile* is concerning for future use of these drugs as resistance to currently used drugs continues to rise. Resistance to vancomycin, which is recommended for use during CDI (Ooijevaar *et al.*, 2018), is increasing, with prevalence currently predicted to be around 1-2% of *C. difficile* strains (Saha *et al.*, 2019; Sholeh *et al.*, 2020). The prevalence of tetracycline resistance genes in *C. difficile* is relatively high, with 20% of strains showing resistance to tetracyclines (Sholeh *et al.*, 2020). Although tetracycline is not used to treat CDI, several *tet*-type genes can mutate to provide high-level resistance against the glycylcycline antibiotic tigecycline (Linkevicius, Sandegren, and Andersson, 2015). Tigecycline is sometimes used to treat severe cases of CDI, often when other treatment has failed (Kechagias *et al.*, 2020), so the high prevalence of resistance against this antibiotic is concerning.

On top of being a reservoir of opportunistic pathogens, the human gut harbours many ARGs, collectively termed the gut resistome (van Schaik, 2015; Anthony *et al.*, 2021; Crits-Christoph *et al.*, 2022). A study in 2013 determined that ARGs make up around 0.03% of all genes in the human gut microbiome (Hu *et al.*, 2013). Many ARGs present in the gut are present in members of the commensal microbiota, although the extent to which these ARGs can transfer to opportunistic pathogens remains understudied (Lamberte and van Schaik, 2022).

1.4 The human gut resistome

As the gut contains large, densely packed populations of bacteria, there are ample opportunities for transfer of ARGs between other members of the microbiota, including opportunistic pathogens (Salyers, Gupta, and Wang, 2004; van Schaik, 2015). Because of this, there is a clear need to characterise the human gut resistome to determine the extent to which commensals serve as a reservoir of ARGs and contribute to the emergence of multidrug-resistance in opportunistic pathogens.

The most common ARGs in the gut microbiota are those that encode for resistance to tetracyclines, macrolides, aminoglycosides, and β-lactams (Singh, Verma, and Taneja, 2019; Lamberte and van Schaik, 2022). A selection of common ARGs found in the human gut microbiota and their host range is shown in Table 1.1. Analysis of 162 human faecal samples from China, Denmark, and Spain detected ARGs putatively conferring resistance to tetracycline (*tet*(32), *tet*(40), *tet*(O), *tet*(Q), and *tet*(W)), aminoglycosides (*ant(6)-Ia*), bacitracin (*bacA*), and vancomycin (*vanRA*, and *vanRG*) in all samples (Hu *et al.*, 2013). Seville *et al.* (2009) determined that *tet*(M) and *tet*(W) are the most prevalent tetracycline resistance genes in the human gut, followed by *tet*(Q) and *tet*(O). These *tet*-type genes are commonly located on mobile elements. The *tet*(Q) gene is highly prevalent among *Bacteroides* spp. due to the widespread conjugative transposon (CTn) CTnDOT (Waters and Salyers, 2013). This CTn also often contains the macrolide resistance gene *erm*(F), and is present in around 80% of *Bacteroides* spp. due to extensive HGT (Shoemaker *et al.*, 2001; Waters and Salyers, 2013). Strains of the gut commensal *Bifidobacterium* has also been found to carry *tet*(W) present in a putative CTn (Duranti *et al.*, 2017). Similarly, *tet*(M) is often present in a Tn*916*-like conjugative transposon that is prevalent amongst members of the gut microbiota (Seville *et al.*, 2009). Tn*916*-like elements are widespread MGEs in the gut, often conferring resistance to tetracycline and macrolides (Roberts and Mullany, 2011).

Aminoglycoside resistance genes are present in both Gram-negative and Gram-positive gut commensals. A recent study found that genes encoding for aminoglycoside resistance genes were prevalent in the gut commensals *Bacteroides* and *Lactobacillus*, with similar ARGs being detected in *Enterococcus* spp. (Alekseeva *et al.*, 2022). The aminoglycoside resistance gene *aph(2")-Ib*, thought to confer high-level gentamicin resistance in *E. faecium* (Kao *et al.*, 2000), has been found to be associated with MGEs in anaerobic gut commensals, including in a plasmid carried by a Firmicutes gut commensal in the genus *Subdoligranulum* (Buelow *et al.*, 2014).

Genes encoding resistance to β-lactam antibiotics are highly prevalent in several genera of the Bacteroidetes phylum (Veloo *et al.*, 2019). The *cfxA* gene is widespread within members of the *Bacteroides* and *Prevotella* genera (García *et al.*, 2008; Tran, Tanaka, and Watanabe, 2013). Other genes such as *cepA* are also common in *Bacteroides fragilis* (Tran, Tanaka, and Watanabe, 2013), an opportunistic pathogen from the gut microbiota that is associated with anaerobic bacteraemia (Cobo *et al.*, 2020). *B. fragilis* is known to be highly resistant to numerous classes of antibiotics due to carriage of various ARGs (Jasemi *et al.*, 2021), and several CTns have been discovered that mediate HGT of ARGs between *B. fragilis* and other related members of the gut (Husain *et al.*, 2014, 2017). The transfer of a plasmid carrying a β-lactam resistance gene from a species of *Bacteroides* to *E. coli* has been observed *in vitro* (Guiney and Davis, 1978), however the prevalent β-lactamases within the *Bacteroides* genus are not detected in members of the Enterobacteriaceae, indicating that this transfer may not happen frequently within the natural gut microbiota. Therefore, although β-lactam resistance is clinically concerning in *B. fragilis*, the β-lactamases detected within *Bacteroides* and related genera are distinct from those found within MDR Enterobacteriaceae (van Schaik, 2015). β-lactamases from *Bacteroides* have been observed *in vitro* to be released in membrane vesicles, and these β-lactamase-containing vesicles were able to protect surrounding bacteria, including *Salmonella*, from β-lactam antibiotics (Stentz *et al.*, 2015), although it is unclear if these protective vesicles majorly contribute to resistance in the gut.

Whilst evidence of β-lactamase gene transfer between *Bacteroides* and Enterobacteriaceae in the gut is lacking, there is capacity for HGT between commensal gut bacteria and opportunistic pathogens. Although rare, transfer between distantly related gut bacteria has been observed (van Schaik, 2015). Identical sequences of the macrolide resistance gene *erm*(B) found in the members of the Gram-negative *Bacteroides* genus can be detected in the Gram-positives *Clostridium perfringens* and *E. faecalis*, indirectly suggesting that the same gene can be transferred to a wide variety of bacteria (Shoemaker *et al.*, 2001). Similarly, there is indirect evidence of ancestral HGT between phyla in the gut as identical plasmid genes have been detected in both Firmicutes and Proteobacteria (Jones, Sun, and Marchesi, 2010). In a recent study, Forster and colleagues (2022) compared published gut microbiota data and showed strong evidence for MGE-mediated transfer between commensal and pathogenic gut bacteria, including, albeit relatively rarely (1.5% of transfer events), inter-phyla transfer of broad host range MGEs. Of the putative transfer events, over 64,000 (~16.5% of total) of the genes transferred between commensals and opportunistic gut pathogens were ARGs (Forster *et al.*, 2022). The authors also demonstrated experimentally that a plasmid carrying a tetracycline resistant gene could be transferred from the Gram-positive gut commensal *Dorea longicatena* to the Gram-negative opportunistic pathogen *Klebsiella oxytoca* (Forster *et al.*, 2022). This work demonstrated the capacity of the gut microbiota to be a reservoir of ARGs and, importantly, that commensal species are capable of disseminating these genes.

Transfer of ARGs from commensals in the gut microbiota could be particularly important for the dissemination of vancomycin resistance to Enterococcaceae or other opportunistic pathogens from the Firmicutes phylum. Although Hu *et al.* (2013) showed that *vanRA* and *vanRG* were highly prevalent in the healthy gut microbiota, these are regulatory genes that may not contribute to vancomycin resistance as they are not directly involved in cell wall synthesis (Ammam *et al.*, 2013). However, *vanB*, one of the most common vancomycin resistance gene found in MDR *Enterococcus* spp. (Ahmed and Baptiste, 2018), has been detected in several anaerobic gut commensals (Stinear *et al.*, 2001; Ballard *et al.*, 2005; Graham *et al.*, 2008; Howden *et al.*, 2013). The findings of these studies suggest intergeneric transfer of vancomycin resistance operons may occur relatively frequently in the gut. Transfer of a *vanB*2 gene between genera has been demonstrated *in vitro* from both *Eggerthella lenta* and *Clostridium symbiosum* to *E. faecium*, as well as *in vivo* from *C. symbiosum* to *Enterococcus* spp. in mice (Launay *et al.*, 2006). *In vivo* intergeneric transfer to *Enterococcus* spp. has also been demonstrated for other ARGs, such as the transfer of plasmids carrying *tet*(M) and *erm*(B) from *Lactobacillus plantarum* to *E. faecalis* in rats (Jacobsen *et al.*, 2007). Treatment with antibiotics likely increases these transfer events within the gut, as is observed in Enterobacteriaceae (Goren *et al.*, 2010; Huddleston, 2014). Jakobsson *et al.* (2010) demonstrated that short-term treatment with the macrolide drug clarithromycin resulted in high-levels of the macrolide resistance gene *erm*(B) in the gut microbiota that still persisted four years later. Another study showed that a conjugative plasmid carrying aminoglycoside (*aac(6')-Ie-aph(2")-Ia*) and macrolide (*erm*(B)) resistance genes transferred between antibiotic-sensitive and MDR strains of *E. faecium* during streptomycin treatment in mice (Lester, Frimodt-Moller, and Hammerum, 2004).

Overall, the findings from these studies suggest that there is considerable potential for clinically-important transfer of ARGs from commensals to opportunistic pathogens, likely contributing to the emergence of MDR strains. The extent of this is understudied, and there is a clear need to fully characterise the human gut resistome to determine which bacteria are carrying and transferring ARGs within this complex ecosystem.

Table 1.1. Host ranges of a selection of common antimicrobial resistance genes in the human gut

Antibiotic class	ARG*	CARD accession	Mechanism of resistance	Representative hosts†
Tetracycline	*tet*(M)	ARO:3000186	antibiotic target protection	*Enterococcus*, *Gemella*, *Megasphaera*, *Peptoniphilus*, *Streptococcus*
	tet(W)	ARO:3000194		*Bifidobacterium*, *Christensenella*, *Eggerthella*, *Eubacterium*, *Megasphaera*
	tet(Q)	ARO:3000191		*Alistipes*, *Bacteroides*, *Enterocloster*, *Phocaeicola*, *Prevotella*
	tet(O)	ARO:3000190		*Enterocloster*, *Eubacterium*, *Roseburia*, *Streptococcus*
Macrolide	*erm*(F)	ARO:3000498	antibiotic target alteration	*Bacteroides*, *Histophilus*, *Phocaeicola*, *Prevotella*, *Riemerella*
	erm(B)	ARO:3000375		*Enterococcus*, *Gemella*, *Megasphaera*, *Staphylococcus*, *Streptococcus*
Aminoglycoside	*ant(6)-Ia*	ARO:3002626	antibiotic inactivation	*Enterococcus*, *Ligilactobacillus*, *Streptococcus*
	aac(6')-Ie-aph(2")-Ia	ARO:3002597		*Enterococcus*, *Eubacterium*, *Megasphaera*, *Staphylococcus*, *Streptococcus*
Beta-lactam	*cfxA3*	ARO:3003003	antibiotic inactivation	*Bacteroides, Capnocytophaga*, *Phocaeicola*, *Prevotella*
	cepA	ARO:3003559		*Bacteroides*
Glycopeptide	*vanRA*	ARO:3002919	antibiotic target alteration	*Enterococcus*, *Bacillus*, *Brevibacillus*
	vanB	ARO:3000013		*Enterococcus*, *Enterocloster*
Other	*bacA*	ARO:3002986	antibiotic target alteration	*Citrobacter*, *Enterobacter,* *Escherichia*, *Salmonella*, *Shigella*

*selection of common antimicrobial resistance genes (ARGs) in the gut microbiota mentioned in the main text or from the top 20 (median) ARGs in the human gut microbiota determined by Hu *et al.* (2013)

†representative selection of (up to 5) genera most commonly associated with ARG according to the Comprehensive Antibiotic Resistance Database (CARD) (Alcock *et al.*, 2020)

1.5 Methodologies to study the gut resistome

The progress in knowledge of the gut microbiome can be attributed to the great advancement of research methodologies and technologies used to study the field in recent decades. Pioneering studies to uncover the role of the gut microbiome in health and disease used gnotobiotic mice, where all organisms in the gut were known, testing the consequence of different treatments (Dubos and Schaedler, 1960; Savage and Dubos, 1968). These initial studies used classical culturing methods to try and uncover the components of the gut microbiota. However, the vast majority of species in the gut are obligate anaerobes, meaning they strictly grow in anaerobic conditions, making culturing of these organisms nontrivial (Sarangi, Goel, and Aggarwal, 2019), with many organisms being deemed "unculturable" (Stewart, 2012). Anaerobic culturing methods were developed (Attebery and Finegold, 1969), allowing culture of some obligate anaerobes from the gut (Moore and Holdeman, 1974). Despite this, isolation of gut microbes remained difficult, and culture-based techniques for describing the composition of the gut microbiota continued to be limited by the laborious nature of the methods, as well as the biases towards high-abundant and aerobic organisms (Sarangi, Goel, and Aggarwal, 2019).

However, more recently, novel culture-based methods have been developed to culture many of the "unculturable" gut microbiota (Lagier *et al.*, 2012, 2016, 2018; Pfleiderer *et al.*, 2013; Browne *et al.*, 2016; Lau *et al.*, 2016; Zou *et al.*, 2019). This improvement of anaerobic culturing has allowed the isolation and characterisation of hundreds of novel bacterial species and genera from the gut microbiota. These studies use numerous different culture conditions to isolate many and diverse organisms from human faecal samples (Lau *et al.*, 2016; Lagier *et al.*, 2018; Zou *et al.*, 2019). The number of culture conditions makes these large-scale projects incredibly laborious and time-consuming, but isolation of these gut microbes allows exploration of novel physiological processes as well as full taxonomic classification and characterisation (Hitch *et al.*, 2021). Broad host range, rich culture media have also been developed to allow culture of many diverse gut species using fewer conditions, such as YCFA (yeast extract, casein hydrolysate, fatty acids) and GAM (Gifu Anaerobic Medium) (Alou *et al.*, 2021).

Culture-based methods continue to improve, and in the coming years more of the gut microbiota will be isolated and characterised. That said, the huge boom in research of the gut microbiota over the last few decades was boosted by advancements in culture-independent methods.

1.5.1 Culture-independent methods to study the gut microbiota

Culture-independent techniques allow the characterisation of the gut microbiota without having to culture and isolate individual species. Techniques such as quantitative polymerase chain reaction (qPCR) have been used to assess the prevalence and abundance of ARGs and other genetic markers in faecal samples by amplifying the genes and quantifying the abundance (Buelow *et al.*, 2014, 2017; Sun *et al.*, 2017). However, these methods are limited to detecting only genes that are complementary to the primers used for amplification, and are thus only able to detect specific, known ARGs. There also exists the risk of nonspecific amplification in PCR-based approaches, or the presence of false negatives when primers do not bind to homologous ARGs that have mutations in the primer-binding sites.

Development of high-throughput, low-cost DNA-sequencing technologies in the last decade has enabled a substantial increase in the knowledge of the composition of the gut microbiota (Rajilić-Stojanović and de Vos, 2014). These modern sequencing technologies, collectively termed next-generation sequencing (NGS), have rapidly evolved since the early 2000s, allowing cheaper and higher-quality sequencing of DNA (Goodwin, McPherson, and McCombie, 2016). This has fuelled a substantial increase in research utilising bacterial genomics, where the bacterial genome is studied. Bacterial whole-genome sequencing (WGS) technology has been transformative for microbiology as it allows a comprehensive insight into the organism, including taxonomic classification and investigation of the genomic context of ARGs, in a quick, affordable way (Uelze *et al.*, 2020). However, WGS requires culturing of the organism for DNA isolation, so has not been effective for studying large, complex ecosystems like the human gut microbiota. To overcome this challenge, instead of sequencing the whole-genome of an organism, specific genes can be amplified, barcoded, and sequenced. The most common example of this is amplifying the 16S ribosomal RNA (16S rRNA) gene (Sarangi, Goel, and Aggarwal, 2019). The 16S rRNA gene is around 1,500 base pairs (bp) in length, and is present in all bacteria, encoding for the RNA component of the 30S ribosomal subunit (Janda and Abbott, 2007). This highly conserved gene contains hypervariable regions that differ between species of bacteria, allowing taxonomic identification and phylogenetic study of bacterial species (Sarangi, Goel, and Aggarwal, 2019). 16S rRNA gene amplicon sequencing has allowed research into the composition of the gut microbiota, and insights into the abundance of individual species, without the need to culture the organisms (Eckburg *et al.*, 2005; Ley *et al.*, 2005; Mizrahi-Man, Davenport, and Gilad, 2013; Clavel *et al.*, 2022).

Amplicon sequencing allows for profiling of the composition of a sample, but does not obtain other genomic information such as presence of ARGs or MGEs. The advent of deep, high-quality, NGS has enabled the rise of metagenomics, whereby all genetic material from a sample can be studied (Venter *et al.*, 2004; Gill *et al.*, 2006; Mandal, Saha, and Das, 2015). Shotgun metagenomic sequencing is a technique whereby all DNA is extracted from a sample, randomly fragmented, and deep-sequenced (Quince *et al.*, 2017). The DNA is typically sequenced using "short-read" technologies such as Illumina-based sequencing, a highly-accurate sequencing platform that outputs short (usually 75-300 bp) sequences of DNA. A paired-end approach is usually employed, whereby each end of the DNA is sequenced to obtain a pair of reads, facilitating analysis of the sequencing data, especially for repetitive regions of a genome (McCombie, McPherson, and Mardis, 2019). Following sequencing, the reads undergo metagenome *de novo* assembly, where the short reads are assembled into longer, contiguous, sequences of DNA called contigs (Quince *et al.*, 2017). Assembly is typically performed using a de Bruijn graph approach (Pevzner, Tang, and Waterman, 2001), where each read is broken into short, overlapping sequences called *k*-mers. The assembly program uses these overlapping *k*-mers to construct a de Bruijn graph and then determines paths through the graph to assemble contigs (Quince *et al.*, 2017). Bioinformatic analysis of the reads and assembled contigs allows a comprehensive investigation into both the composition of a sample and the genes and MGEs present in the sample, making it a powerful tool for analysis of the gut microbiota (Garmendia *et al.*, 2012; Hu *et al.*, 2013).

One difficulty during analysis of sequencing data is the ability to taxonomically classify the assembled contigs. Many tools exist to classify contigs in metagenomic assemblies

(Ye *et al.*, 2019). Many of these are based on DNA alignment, such as the Basic Local Alignment Search Tool (BLAST) which aligns DNA sequences to large databases of published sequences (National Library of Medicine, 2022). One of the most popular tools, Kraken (Wood and Salzberg, 2014), works by breaking up sequences into *k*-mers and mapping them to a database storing *k*-mer information for individual organisms. Kraken bases the taxonomic classification of the lowest common ancestor (LCA) of the genomes that the *k*-mers mapped to (Wood and Salzberg, 2014). Both BLAST and Kraken are limited by need for the sequenced organism to be present and correctly classified in the taxonomy database being used. This is particularly a problem for MGEs present in multiple species or contigs originating from species that are not yet present in the database (Von Meijenfeldt *et al.*, 2019). Binning techniques can increase the accuracy of classification by classifying groups of contigs rather than individual sequences. Binning is the process of grouping contigs that are considered to originate from the same genome, with the goal of constructing high-quality bins representing complete microbial genomes, termed metagenome-assembled genomes (MAGs) (Yang *et al.*, 2021). Several tools have been developed to bin contigs and construct MAGs based on various variables such as tetranucleotide frequencies and read depths (Imelfort *et al.*, 2014; Kang *et al.*, 2019), taxonomic alignments (Wang *et al.*, 2019), information from the assembly graph (Mallawaarachchi, Wickramarachchi, and Lin, 2020), or a combination thereof to obtain a consensus set of bins from multiple approaches (Sieber *et al.*, 2018). Like classifying contigs, taxonomic classification of metagenomic bins can be performed using various tools, with one of the most popular being the Genome Taxonomy Database Toolkit (GTDB-Tk) (Chaumeil *et al.*, 2020). GTDB-Tk uses concatenated marker gene alignment to construct a phylogenetic tree and then assigns a taxonomic classification of the bin based on its placement in the tree as well as its relative evolutionary divergence and average nucleotide identity (ANI) to its most related reference genome from the Genome Taxonomy Database (GTDB) (Chaumeil *et al.*, 2022). The GTDB is a database curated from all genomes published in the RefSeq (O'Leary *et al.*, 2016) and GenBank (Sayers *et al.*, 2022) databases, including unclassified genomes from uncultured organisms that are assigned placeholder names in the GTDB (Parks *et al.*, 2018).

A major limitation of short-read sequencing is the difficulty of assembling highly repetitive regions of the genome, including insertion sequences and rRNA operons, as the short reads do not span many of the repeats that are present in multiple copies in a bacterial genome. ARGs and MGEs are often flanked by repetitive elements, meaning genomic regions containing ARGs or MGEs are often fragmented in the assembled metagenome (Koessler *et al.*, 2010; Berglund *et al.*, 2019). This makes it very difficult to associate ARGs and MGEs with their microbial hosts. "Long-read" sequencing technologies such as PacBio and Oxford Nanopore sequencing platforms are capable of sequencing long (>10 kilobase pairs (kbp)) DNA fragments that can span these repetitive regions, allowing for the assembly of considerably longer contigs (Amarasinghe *et al.*, 2020). Long-read metagenomic sequencing is a relatively new technique that has been used to improve acquisition of high-quality MAGs from gut microbiota sequencing data (Cuscó *et al.*, 2021; Jin *et al.*, 2022). However, even with long-read technologies, associating plasmids and other MGEs with their microbial hosts is still difficult with metagenomic sequencing alone.

1.5.2 Linking ARGs to their microbial hosts

Asides from improved culturing methods discussed above, several experimental and bioinformatic techniques have been recently developed attempting to detect HGT or identify the hosts of ARGs in the human gut microbiota (McInnes and McCallum *et al.*, 2020). A bioinformatic tool, MetaCherchant, extracts information from the de Bruijn assembly graph to link ARGs and MGEs to their hosts (Olekhnovich *et al.*, 2018). By performing this on sequencing data from gut microbiota samples before and after antibiotic treatment, MetaCherchant was able to detect a change of host for aminoglycoside resistance gene *aph(3')* and β-lactamase gene *cfx* following antibiotic treatment. However, this program is limited by the ability to accurately classify the linked hosts using Kraken, and MGEs with high copy-numbers create complexity in the assembly graph which complicates host assignment of these elements (Olekhnovich *et al.*, 2018). Another bioinformatic tool to associate plasmids with their microbial hosts utilises DNA methylation patterns measured during long-read sequencing to bin contigs (Beaulaurier *et al.*, 2017). DNA methyltransferases in bacteria methylate all DNA inside the cell, including both the chromosome and any MGEs, and methylation patterns differ among different strains or species (Blow *et al.*, 2016). Therefore, by scoring these methylation patterns and finding contigs with matching scores, MGEs can be associated with chromosomal DNA (Beaulaurier *et al.*, 2017). The authors implemented this method on a mouse gut microbiota sample and were able to link 8/19 contigs identified as a plasmid or CTn to a host bin, 7/9 of which were classified as members of the Bacteroidales class by Kraken (Beaulaurier *et al.*, 2017). This method is, however, limited to low complexity samples, as the uniqueness of each methylation pattern is reduced as the number of genomes in the community increases.

For tracking HGT experimentally, Munck *et al.* (2020) developed a system that uses the CRISPR (clustered regularly interspaced short palindromic repeats) spacer acquisition process to detect and record HGT events. The CRISPR-Cas9 system is a bacterial defence system that incorporates invading DNA into the CRISPR locus in the genome (Barrangou *et al.*, 2007). An *E. coli* 'recording strain' was engineered to capture and record any DNA that enters the cell, allowing HGT events to be tracked. The authors then incubated this strain in a suspension consisting of human faecal samples and recorded frequent HGT events (Munck *et al.*, 2020). For now, this technique has only been employed in a strain of *E. coli*, however it shows promise for future studies investigating which species are transferring genes in the gut, particularly if it can be adapted to operate in other species.

Two major experimental approaches to link functional and phylogenetic data have also advanced in recent years (McInnes and McCallum *et al.*, 2020). Emulsion, Paired Isolation, and Concatenation PCR (epicPCR) was designed to link functional genes to the 16S rRNA gene (Spencer *et al.*, 2016). The technique works by encapsulating single bacterial cells in polyacrylamide beads, followed by a fusion PCR of the target gene and the 16S rRNA gene, producing one single amplicon that can then be sequenced. In the first study using epicPCR, the authors were able to link the dissimilatory sulphite reductase gene, *dsrB*, with 16S rRNA genes in a freshwater lake sample (Spencer *et al.*, 2016). The same group also used epicPCR to detect the bacterial hosts of four ARGs in wastewater (Hultman *et al.*, 2018). Both *tet*(M) and *bla*$_{\text{OXA-58}}$ were detected in members of the Bacteroidetes, Firmicutes, Fusobacteria, and Proteobacteria phyla, although *tet*(M) had a broader host range and was also linked to members of the phyla Tenericutes and Gracilibacteria (Hultman *et al.*, 2018).

Interestingly, the authors observed putative HGT events in this study, linking *bla*$_{\text{OXA-58}}$ with the 16S rRNA gene from a member of the Leptotrichiaceae family in an effluent wastewater sample that was not detected in the influent sample, and observed similar occurrences for *tet*(M). However, the authors acknowledge that this observation could be due to the detection limits of epicPCR, and further investigation would be needed to confirm that HGT events had occurred (Hultman *et al.*, 2018). A similar technique called One-step Isolation and Lysis PCR (OIL-PCR) has recently been developed and employed on human faecal samples to link plasmid-based β-lactamases to their bacterial hosts (Diebold *et al.*, 2021). The authors linked *bla*$_{\text{TEM}}$, *bla*$_{\text{SHV}}$, and *bla*$_{\text{CTX-M}}$ with the opportunistic pathogen *Klebsiella* as well as the commensal genus *Romboutsia*. However, 16S rRNA gene sequences for both *Klebsiella* and *Romboutsia* were also found to be fused together, which the authors acknowledged indicated that *Klebsiella* and *Romboutsia* were emulsified together rather than showing an HGT event (Diebold *et al.*, 2021). These studies demonstrate that both epicPCR and OIL-PCR have potential to be used to link ARGs to their hosts in human faecal samples. However, the poor detection limit and issues with multiple cells being encapsulated together make these techniques difficult for comprehensive, high-throughput characterisation of the gut resistome. Another set of recently developed methods, termed proximity ligation, may show more promise.

1.5.2.1 Proximity ligation techniques

Proximity ligation techniques, known as chromosome conformation capture (3C), were originally designed to study the frequency of chromatin interactions in eukaryotic cells (Dekker *et al.*, 2002). More recently, these techniques have been developed by several groups to physically link bacterial genes in single cells (Marbouty and Koszul, 2017).

The initial steps of 3C (Figure 1.5a) involve incubating cells in formaldehyde to cross-link the DNA, followed by restriction digestion and re-ligation of the sticky ends of the cross-linked DNA fragments. The ligation step is carried out under dilute conditions to facilitate more intramolecular ligation of the cross-linked DNA rather than intermolecular ligation of different fragments (Sati and Cavalli, 2017). Originally, the generated 3C library was then used as a template for PCR, which was used to detect individual ligation products (Figure 1.5b) (Dekker *et al.*, 2002). Quantification of the 3C library can reveal the proximity of interacting chromosomal regions in the cells, thus allowing the inference of the three-dimensional organisation of the chromosome (Dekker, 2006). This PCR approach to quantify the 3C library is limited to detecting cross-links between known, selected pairs of sequences as it requires locus-specific primers. However, advancements of sequencing technologies have led to the combination of 3C and NGS, and 3C libraries have been pair-end sequenced (referred to as 3C-seq) to allow detection of many cross-link events (Figure 1.5c) (Rodley *et al.*, 2009; Tanizawa *et al.*, 2010). Bioinformatic analysis of the sequenced 3C library can indicate which genes or regions of the chromosome were cross-linked together, which allows a high-throughput, comprehensive approach to unveiling the organisation of the genome using 3C technologies (Tanizawa and Noma, 2012).

Various derivatives of 3C have also been developed, as demonstrated in Figure 1.5. Most of these methods involve enriching the cross-linked sample, either to enrich for a specific gene or to eliminate non-cross-linked DNA from the 3C library. Firstly, Hi-C (Figure 1.5d) was developed by Lieberman-Aiden *et al.* in 2009. This method follows the same cross-linking and digestion steps as 3C, but prior to ligation, the 5'-overhangs left by the restriction digest are blunted and tagged with biotin. Following ligation of the blunt-ends, the Hi-C library is created by shearing the DNA and selecting the biotin-containing fragments with streptavidin beads, resulting in enrichment for ligation junctions only, thus increasing the proportion of cross-linked DNA in the library (Lieberman-Aiden *et al.*, 2009).

Methods to enrich for specific DNA fragments have also been developed. 'Circular 3C' or '3C on chip', both known as 4C, involve an additional digestion and ligation step to create a circularised 3C library, on which inverse PCR can be performed to amplify the DNA fragment that was cross-linked to the target gene (Figure 1.5e). Capture Hi-C (CHi-C) combines the specific enrichment advantage of 4C with the Hi-C method (Figure 1.5f). After a Hi-C library is prepared, custom-designed DNA probes are used to 'capture' Hi-C products containing specific fragments of DNA (Moreau *et al.*, 2018). Various other 3C-based methods exist which, like 4C and CHi-C, aim to enrich for specific genes or cross-linking of specific proteins of interest, such as ChIP-loop (de Wit and de Laat, 2012). Methods like standard 3C, 4C, and CHi-C were designed for investigating genes within proximity of specific target sequences of DNA, whereas 3C-seq and Hi-C are powerful tools for more high-throughput analysis of samples. 3C and its derivatives have so far been predominantly used for studying the structure and organisation of eukaryotic genomes (reviewed by Sati and Cavalli, 2017). Recently, however, several studies have applied 3C-based methods on metagenomic analysis of bacterial communities.

Metagenomic 3C (meta3C) was developed by Marbouty *et al.* (2014) to characterise individual genomes in a complex microbial community by optimising a 3C-seq protocol that can be used on metagenomic samples (Figure 1.5c). By performing meta3C on a mixture of yeast and bacteria, the authors were able to scaffold the genomes of many diverse species in the mixture (Marbouty *et al.*, 2014). Meta3C was also performed on river sediment samples, validating its potential for use on complex environmental samples (Marbouty *et al.*, 2014). In the same year, two other groups described methods for deconvolution of bacterial metagenomes using Hi-C (Beitel *et al.*, 2014; Burton *et al.*, 2014). Due to the additional step in Hi-C to enrich for ligation junctions, a key difference between Hi-C and 3C-seq/meta3C is that the proportion of read pairs originating from cross-linked DNA is increased. This means that a meta3C library must be sequenced more deeply than a Hi-C library to ensure sequencing of cross-linked fragments. However, it also means that, for Hi-C, additional shotgun sequencing of the sample is required for metagenomic assembly, which can then be coupled with the Hi-C data to link assembled contigs (Beitel *et al.*, 2014; Burton *et al.*, 2014). In contrast, the high proportion of non-cross-linked fragments sequenced from a meta3C library allows generation and scaffolding of contigs directly from meta3C sequencing data (Marbouty *et al.*, 2014). This requirement of an additional shotgun sequencing step means that these Hi-C methods are limited by the amount of sample material available compared to meta3C. Additionally, both Hi-C studies did not test the methods on environmental samples, and only showed results from synthetic mixtures of bacteria and yeast (Beitel *et al.*, 2014; Burton *et al.*, 2014), in contrast to the meta3C study, which showed meta3C working on an unknown, semi-complex environmental sample (Marbouty *et al.*, 2014). Notably, all studies were able to link plasmid sequences to chromosomal genes, enabling the plasmid to be linked to a bacterial host (Beitel *et al.*, 2014; Burton *et al.*, 2014; Marbouty *et al.*, 2014). This demonstrates that 3C-based techniques can be used to link plasmids to bacterial species in complex microbial communities.

Since the pioneering 2014 studies, 3C-based approaches have been applied to other complex microbial communities, including gut microbiome samples. Meta3C was performed on a mouse gut microbiome to link phage sequences to their bacterial hosts (Marbouty *et al.*, 2017). Additionally, following similar methods to the Beitel *et al.* and Burton *et al.* (2014) studies, Hi-C has been used to assemble nearly 1,000 microbial genomes from a cow rumen microbiome (Stewart *et al.*, 2018). Using Hi-C paired with shotgun sequencing, the authors were also able to identify the bacterial hosts of plasmids (Stewart *et al.*, 2018), although the authors used ProxiMeta, a proprietary Hi-C procedure and bioinformatic pipeline, and thus did not give details on how the microbial genomes where deconvoluted. Subsequently, this Hi-C method was combined with long-read sequencing to link viruses and ARGs to their microbial hosts (Bickhart *et al.*, 2019). A combination of long-read sequencing and Hi-C was also recently performed on sheep faecal samples to resolve MAGs and link viruses and plasmids to their microbial hosts (Bickhart *et al.*, 2022). ProxiMeta has also been used on a human gut microbiome sample, where the authors claimed to have linked plasmids with host genomes, although the authors only showed that they were able to link a 600 kbp megaplasmid to several species in the order Clostridiales (Press *et al.*, 2017). The same group also attempted to identify the bacterial hosts of ARGs in wastewater samples, also using the ProxiMeta Hi-C method (Stalder *et al.*, 2019). The results indicated that the phylum Bacteroidetes was the most prominent reservoir of ARGs in the sample, with the genus *Bacteroides* being associated with various ARGs such as *tet*(Q), *erm*(G), and *bla*$_{\mathrm{CFX}}$ (Stalder *et al.*, 2019). These data align with previous studies of ARGs carried by the *Bacteroides* species (Shoemaker *et al.*, 2001). Various mobile genetic elements were also linked to multiple taxa. Incompatibility group Q

(IncQ) plasmids, which are known to have a broad host range (Jain and Srivastava, 2013), were reported to have the broadest range of bacterial hosts, whilst most plasmids were linked to Enterobacteriaceae, which the authors suggest was probably due to biases in the plasmid database that they used (Stalder *et al.*, 2019).

More recently, 3C was used to track the transfer of genes in the human gut microbiome in two samples collected ten years apart, although the authors did not focus on linking ARGs to their microbial hosts (Yaffe and Relman, 2020). Another group used Hi-C to link ARGs and mobile genes to microbial hosts and track HGT events in the human gut microbiome in multiple haemopoietic transplant patients over a three week time course (Kent *et al.*, 2020). ARGs and mobile elements were linked to a wide range of taxa, and putative HGT events occurred throughout the course of the study. This is arguably the most comprehensive study of the human gut resistome using 3C-based techniques to date, although the authors did not discuss the specific identity of the microbial hosts of ARGs. Additionally, some of the inferred HGT events are surprising, such as frequent HGT between *Klebsiella* and *Enterococcus* (Kent *et al.*, 2020), while most evidence suggests that HGT between phyla remains extremely rare in the human gut microbiome (Porse *et al.*, 2018; Forster *et al.*, 2022).

Several recent studies have also used proximity ligation techniques to link bacteriophages to their microbial hosts in gut microbiome samples. Firstly, using meta3C and an improved derivative, called metaHiC, on ten healthy human gut samples, 715 MAGs were able to be linked to an average of 6.9 phage contigs each (Marbouty *et al.*, 2021). Another study using ProxiMeta Hi-C was able to assign a host to 78% of identified phages in a human gut microbiome sample, as well as to 77% and 69% of phages in a cow rumen sample and wastewater samples, respectively (Uritskiy *et al.*, 2021). Similarly, and again using ProxiMeta Hi-C combined with long-read metagenomic sequencing, a study was able to identify and link 51 phages to their bacterial hosts in a canine faecal sample (Cuscó *et al.*, 2022). From this dataset, the authors were also able to link 6 plasmids to their bacterial hosts, including one plasmid carrying the lincosamide resistance gene *lnu*(A), which was linked to a species of *Fusobacterium* (Cuscó *et al.*, 2022). Another recent study developed a bioinformatic pipeline, HAM-ART (Hi-C Assisted Metagenomics for Antimicrobial Resistance Tracking), that utilises Hi-C data for the generation of MAGs to link both chromosomal and plasmid-based ARGs to their bacterial host in pig faecal samples (Kalmar *et al.*, 2022). Using HAM-ART, *lnu*(A), carried in a plasmid, was found to be highly prevalent in the samples and was able to be linked to three species of *Lactobacillus* (Kalmar *et al.*, 2022). Hi-C was also recently used on faecal samples of chronically critically ill patients to improve the generation of MAGs from metagenomic data (Ivanova *et al.*, 2022). Using Hi-C data, the authors were able to link phages and MGEs to opportunistic pathogens and linked a plasmid carrying multiple ARGs to *Klebsiella pneumoniae* (Ivanova *et al.*, 2022).

Altogether, these publications using proximity ligation techniques on real, natural, complex microbiomes to link plasmids and ARGs to their host show considerable potential for future studies to further characterise the human gut resistome.

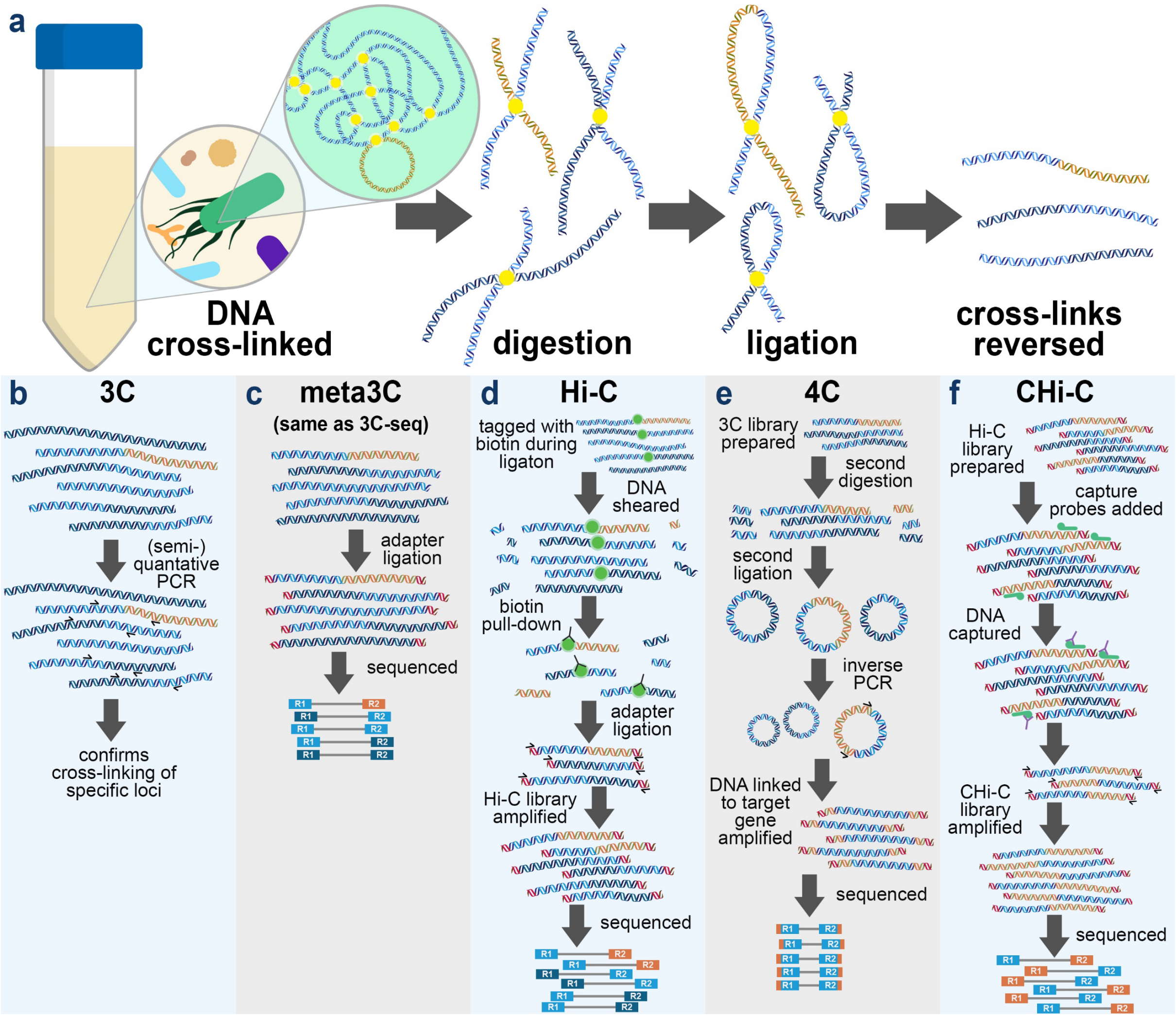


Figure 1.5. Proximity ligation techniques.
In this example, orange strands of DNA represent a plasmid carrying antibiotic resistance genes (ARGs), whilst blue represents chromosomal DNA. **a)** All proximity ligation techniques begin with cross-linking DNA (cross-links shown in yellow) by incubating cells in formaldehyde, followed by restriction digestion, ligation, and reverse cross-linking. **b)** For standard chromosome conformation capture (3C), the cross-linked fragments of DNA then undergo polymerase chain reaction (PCR) using locus-specific primers to confirm cross-linking of known regions of the chromosome. **c)** During 3C-seq or metagenomic 3C (meta3C), adapters are ligated to the cross-linked DNA fragments, followed by sequencing of the library. **d)** Hi-C is similar to meta3C, however, for Hi-C, the ends of the DNA fragment are blunted and tagged with biotin before ligation (biotin shown in green). After removal of the cross-links, DNA is sheared and streptavidin beads are used to pull down biotin-tagged fragments, thus enriching for ligation junctions. The enriched library then undergoes adapter ligation and sequencing. **e)** For circularised 3C (4C), a second digestion and ligation step are performed to circularise the 3C library, followed by inverse PCR to amplify the gene cross-linked with the target gene, and then sequenced. **f)** Capture Hi-C (CHi-C) follows the Hi-C protocol, but after the library is prepared, target genes are enriched for using custom DNA probes that bind to target regions and are pulled down with streptavidin beads, before sequencing of the enriched library.

1.6 Aims

The overarching aim of this project was to use proximity ligation techniques to explore the extent to which the human gut microbiota act as a reservoir of ARGs. With the work presented in this thesis, the following research questions were asked:

- Do commensal bacteria act as a reservoir of ARGs in the human gut microbiome?
- Can 3C-based techniques be used to identify the hosts of ARGs in human faecal samples?
- Can culturing be used to validate 3C-based data?

To answer these questions, the objectives of the study were, therefore, to:

1. Develop a novel bioinformatic workflow to link ARGs to their bacterial host using proximity ligation data.
2. Implement proximity ligation techniques on human faecal samples and link ARGs to their bacterial hosts.
3. Culture and isolate the hosts of ARGs to determine the genomic context of the ARGs and validate the proximity ligation data.

CHAPTER 2

MATERIALS AND METHODS

2.1 General methods

2.1.1 Bacterial strains used in this study

Strains used in this study are shown in Table 2.1. All strains were stored as stocks with 15% (volume/volume (v/v)) glycerol (Fisher Bioreagents) at -80°C. *Escherichia coli* E3090 and *Acinetobacter pittii* OB7 were grown in lysogeny broth (LB) (Sigma-Aldrich), and *Enterococcus faecium* E745 was grown in brain heart infusion (BHI) broth (Sigma-Aldrich). All strains were grown in 30 mL universal containers (STARLAB) at 37°C with shaking at 200 revolutions per minute (rpm). *E. faecium* 64/3 was used as a recipient strain during conjugation assays, described in Section 2.4.7.

Table 2.1. Bacterial strains used in this study

Name	ARGs	Description and BioSample accession number	Reference
E3090	*sul1*, *ant(3")-Ia*, *bla*$_{\text{OXA-1}}$, *floR*, *sul2*, *mdf*(A), *bla*$_{\text{TEM-1B}}$, *tet*(B), *catA1*, *aph(6)-Id*, *aph(3")-Ib*, *bla*$_{\text{CTX-M-1}}$, *mcr-1.1*	An *Escherichia coli* clinical isolate – **SAMEA4699317**	(Janssen *et al.*, 2020)
E745	*aac(6')-Ii*, *msr*(C), *vanHAX*, *dfrG*	An *Enterococcus faecium* clinical isolate – **SAMN04045274**	(Zhang *et al.*, 2017)
OB7	*bla*$_{\text{OXA-500}}$, *mph*(E), *msr*(E), *bla*$_{\text{OXA-58}}$	An *Acinetobacter pittii* clinical isolate (strain WCHAP100020) – **SAMN08031302**	Long, H., Ye, H., Feng, Y., and Wang, X. (2019) Direct GenBank submission.
64/3	*aac(6')-Ii*, *msr*(C) (high-level rifampicin and fusidic acid resistance due to chromosomal mutations)	A plasmid-free *Enterococcus faecium* laboratory strain – **SAMN04009964**	(Bender *et al.*, 2015)

ARG = antimicrobial resistance gene; genes highlighted in blue are present in plasmids; **BioSample accession numbers are shown in bold**

2.1.1.1 Calculating number of colony forming units/mL

To determine viable counts in an overnight broth culture, 10-fold dilutions were made in phosphate-buffered saline (PBS) (AppliChem). Dilutions ranging from 10^{-5} to 10^{-9} were then used to spread 100 µL onto individual agar plates of their respective media and incubated at 37°C for 24 hours (h). Once grown, colonies were counted where possible (>10, <300). An average of the counts for the different dilutions was used to calculate an estimate for the number of colony forming units (cfu)/mL for the given strain.

2.1.2 Stool sample preparation

Stool samples (Table 2.2) were collected and kindly provided by Dr Amanda Rossiter (University of Birmingham). Ethical approval for this study has been obtained from the Bradford Leeds Research Ethics Committee (REC 16/YH/0100).

For meta3C/Hi-C experiments, the stool samples were divided into ~500 mg aliquots and stored at -80°C until use. For future culturing, approximately 2 g of each stool sample was diluted in 10 mL PBS containing 1 mM dithiothreitol (DTT) (Roche Diagnostics) and 20% glycerol and stored as 400 µL aliquots at -80°C until use.

Table 2.2. Stool samples used in this study

Sample name	Description
H1 (dataset G_3C in Chapter 3)	Stool sample obtained from an adult with inflammatory bowel disease
H2	Stool sample obtained from a healthy adult
H3	Stool sample obtained from a healthy adult
H4	Stool sample obtained from a healthy adult

2.1.3 Polymerase chain reaction

PCR was performed using DreamTaq Green PCR Master Mix 2x (Thermo Fisher Scientific) unless otherwise stated. A standard PCR mix consisted of 12.5 µL DreamTaq Green PCR Master Mix 2x, 0.125 µL 100 µM forward primer, 0.125 µL 100 µM reverse primer, 1 µL template DNA (100-500 ng), and 11.25 µL HyClone™ Molecular Biology Grade Water (Cytiva). The PCR was then carried out in a Mastercycler® Pro Thermal Cycler (Eppendorf) using the program described in Table 2.3.

Table 2.3. Standard PCR program

Step	Temperate (°C)	Time	Number of cycles
Initial denaturation	95	3 min	1
Denaturation	95	30 s	32
Annealing	Tm-5	30 s	
Extension	72	1 min per kb	
Final extension	72	15 min	1

PCR = polymerase chain reaction; min = minutes; s = seconds; Tm-5 = melting temperature of primers minus 5°C; kb = kilobase

2.1.4 Agarose gel electrophoresis

Agarose gel electrophoresis was used to quantify and visualise DNA fragment length after DNA extraction, library preparation, or PCR. Agarose (Sigma-Aldrich) was dissolved in 1× Tris/Borate/Ethylenediaminetetraacetic acid (EDTA) (TBE) buffer (National diagnostics) at 1.5% (weight/volume (w/v)) unless otherwise stated. SYBR safe (Invitrogen) was added to the gel at a 1:10,000 dilution prior to pouring. Lanes were loaded with 10 µL DNA mixed with 2 µL 6X Purple Gel Loading Dye (New England BioLabs (NEB)), unless using DreamTaq Green PCR Master mix, where the PCR product was instead loaded directly onto the gel, for each sample. The first and last lane were loaded with 2 µL 100 bp ladder (NEB) and 1 Kb Plus DNA Ladder

(Invitrogen), respectively. Electrophoresis was performed at 100-120 volts for around 50 minutes (min), before visualisation and imaging using blue light epifluorescence (with an automatic exposure time) on an Amersham Imager 680 gel imager (General Electric Healthcare).

2.1.5 Purification of PCR products

For purification of PCR products or DNA, the GeneJET PCR Purification Kit (Thermo Fisher Scientific) was used. A 1:1 volume of Binding Buffer was added to the PCR reaction mix and mixed thoroughly using a pipette. Up to 800 µL of the mixed solution was added to a GeneJET purification column and centrifuged at 14,000 × g for 1 min. The flow-through was then discarded and the process was repeated if any PCR reaction and Binding Buffer mix remained. Once all the mix was added to the GeneJET purification column, 700 µL Wash Buffer was added and centrifuged at 14,000 × g for 1 min. The flow-through was discarded and the column was again centrifuged at 14,000 × g for 1 min. The GeneJET purification column was then transferred to a clean 1.5 mL microcentrifuge tube (Sarstedt). Elution Buffer (50 µL unless stated otherwise) was then added to the centre of the column which was then incubated at room temperature (RT) for 2 min before centrifugation at 14,000 × g for 1 min. The GeneJET purification column was then discarded, and purified DNA was either stored at -20°C or used immediately.

2.1.6 DNA extraction of faecal samples

DNA was extracted from faecal samples using the FastDNA™ Spin Kit for Soil (MP Biomedicals). Up to 500 mg of stool was added to a Lysing Matrix E tube along with 978 µL sodium phosphate buffer and 122 µL MT buffer. The sample was then homogenised using a FastPrep-24™ bead-beater (MP Biomedicals) for

40 seconds (s) at a speed setting of 6.0 m/s before centrifugation at 14,000 × g for 10 min to pellet debris. The supernatant was then transferred to a 2 mL microcentrifuge tube (Sarstedt) and 250 µL Protein Precipitation Solution was added and mixed by inverting the tube 10 times. The tube was then centrifuged at 14,000 × g for 5 min to pellet the precipitate, and the supernatant was transferred to a 15 mL conical tube (Greiner Bio-one) containing 1 mL Binding Matrix suspension and rotated by hand for 2 min. The suspension was then allowed to settle for 3 min before 500 µL of the supernatant was removed and discarded. The Binding Matrix was resuspended, and up to 600 µL of the mixture was added to a SPIN™ filter and centrifuged at 14,000 × g for 1 min. The flow-through was then discarded and the process was repeated until all remaining mixture had been added, then 500 µL SEWS-M solution was used to gently resuspend the pellet with a pipette. The SPIN™ filter was centrifuged at 14,000 × g for 1 min. The flow-through was discarded and the filter was again centrifuged at 14,000 × g for 2 min. The SPIN™ filter was transferred to a clean tube and air-dried for 5 min at RT. Binding Matrix was gently resuspended in the SPIN™ filter with 50 µL DNase/pyrogen-free water and incubated for 5 min at 55°C in a heat block. The DNA was eluted by centrifugation at 14,000 × g for 1 min, and either stored at -20°C or used immediately.

2.1.7 Measurement of DNA concentrations

To measure the concentration of DNA in a sample, the Qubit™ dsDNA HS and BR Assay Kits (Invitrogen) were used for DNA concentrations between 0.2-100 ng/µL and 2-1,000 ng/µL, respectively. A working solution was prepared by diluting the Qubit™ dsDNA HS/BR Reagent 1:200 in Qubit™ dsDNA HS/BR Buffer. Standards were made by mixing 190 µL of the working solution with 10 µL Qubit™ dsDNA HS/BR Standard

#1 and 10 µL Qubit™ dsDNA HS/BR Standard #2 in two Qubit™ Assay Tubes. To measure the DNA samples, 198 µL of the working solution was mixed with 2 µL DNA in a Qubit™ Assay Tube for each sample. The solutions were all mixed thoroughly using a vortex for 3-5 s and incubated at RT for 2 min. DNA concentrations were then measured using a Qubit™ 4 Fluorometer (Invitrogen).

2.1.8 Quantification of DNA using TapeStation analysis

DNA was quantified on a 2200 TapeStation system (Agilent Technologies) to assess the quality and fragment size before sequencing. For meta3C and Hi-C libraries, the High Sensitivity D5000 reagents and ScreenTape (Agilent Technologies) were used by mixing 2 µL High Sensitivity D5000 Sample Buffer with 2 µL DNA. The ladder was prepared by mixing 2 µL High Sensitivity D5000 Sample Buffer with 2 µL High Sensitivity D5000 Ladder. Samples were spun down in a MyFuge™ Mini centrifuge (Benchmark Scientific) and run on a 2200 TapeStation instrument. For metagenomic DNA extracted from stool samples, DNA was sent to Genomics Birmingham and analysed using a Genomic DNA ScreenTape (Agilent Technologies).

2.1.9 Bioinformatic analyses and visualisation

The Cloud Infrastructure for Microbial Bioinformatics (CLIMB) servers (Connor *et al.*, 2016) were used for bioinformatic analyses. The command-line interface of the server was accessed using MobaXterm version (v)21.2 (Mobatek, 2022) and files were transferred to and from my local machine via Secure File Transfer Protocol using WinSCP v5.21.3 (Prikryl, 2022). GNU Parallel v20201122 (Tange, 2020) was often used, where possible, to execute multiple serial command line programs in parallel. Some analyses and visualisation (as stated in detailed descriptions of chapter-specific methods) was performed using the programming language R v4.2.1 (R Core Team,

2022) on RStudio v2022.7.1.554 (RStudio Team, 2022), as well as Microsoft Excel v2209 (Microsoft Corporation, 2022). Data visualisation and statistical analyses were carried out using GraphPad Prism v9.4.1 (GraphPad Software, 2022). Detailed descriptions of specific bioinformatic analysis methods are present in chapter-specific methods.

2.1.9.1 Processing and assembly of metagenomic reads

Duplicate reads were removed using PrinSeq-lite (Schmieder and Edwards, 2011). Reads were then quality filtered (`--nextseq-trim=20` or `-q 20` if the reads were not sequenced on an instrument that uses two-colour chemistry, and min length 60 nucleotides (nt) `-m 60`) and had adapter sequences removed using CutAdapt v2.5 (Martin, 2011). Host sequences were removed by aligning the reads to the GRCh38.p13 human reference genome (or GRCm38.p6 mouse reference genome for the dataset described by Marbouty *et al.* (2017)) from the National Centre for Biotechnology Information (NCBI) (Sayers *et al.*, 2022) with Bowtie2 v2.3.4.1 (Langmead and Salzberg, 2012). Unmapped reads were then extracted using SAMtools v0.1.19 (Li *et al.*, 2009) (`view -b -f 12 -F 256`). The resulting BAM files were then sorted with SAMtools (`sort -n`), and converted back to separate FASTQ files using BEDtools v2.25.0 (Quinlan and Hall, 2010) (`bamtofastq`). The remaining high-quality, non-human (or mouse), unique, paired reads were then assembled using MEGAHIT v1.1.3 (Li *et al.*, 2016) using default parameters and filtering out contigs shorter than 1 kilobase (kb) (`--min-contig-len 1000`). Assembly quality was assessed using QUAST v5.0.2 (Gurevich *et al.*, 2013).

2.1.9.2 Taxonomic profiling of metagenomic reads

Taxonomic profiles of processed metagenomic reads were generated using MetaPhlAn3 v3.0 (Beghini *et al.*, 2021) (`--unknown-estimation --add-viruses`). To create a figure of the class abundance in all samples, `merge_metaphlan_tables.py` was used to create a merged table of all results. The second column of the merged table was removed (`cut -f 1,3-200`) and the column titled `clade_name` was renamed to `ID`. A Comma Separated Values file of the class abundance for all samples was then generated using a python script published by *flannsmith* on GitHub (Smith, 2020) and the data were imported into GraphPad Prism for visualisation.

2.1.9.3 Identification of ARGs, plasmids, and IS elements in assemblies

ARGs were identified using ABRicate v1.0.1 (Seemann, 2021) with the ResFinder database (2020) (Bortolaia *et al.*, 2020) (≥75% coverage, ≥95% identity). Plasmids contigs were identified in the same way using the PlasmidFinder database (Carattoli *et al.*, 2014) (≥80% coverage, ≥90% identity). ABRicate was also used to find IS element contigs using the ISfinder database (Siguier *et al.*, 2006) (≥60% coverage, ≥99% identity). To use the ISfinder database with ABRicate, the ISfinder sequences were downloaded from https://github.com/thanhleviet/ISfinder-sequences, and converted to an ABRicate database using `abricate --setupdb`.

2.1.9.4 Calculating relative abundance of ARGs in metagenomic assemblies

To calculate the abundance of the ARGs, they were first extracted from their contigs using a custom Bash script (Appendix Section A.1). CoverM v0.4.0 (Woodcroft, 2022) was then used to calculate the number of reads mapping to each ARG. The number of mapped reads was then used to calculate the reads per kilobase per million mapped reads (RPKM) using the following formula:

$$RPKM = \frac{(reads\ mapped/(total\ number\ of\ reads/1000000))}{(gene\ length/1000)}$$

2.2 Chapter 3 methods

2.2.1 Measuring abundance of ARGs in a stool sample

Quantitative PCR was used to determine the abundance of ARGs in a stool sample. First, DNA was extracted from 400 mg of stool as described in Section 2.1.6. The extracted DNA was diluted to 100 ng/µL, and 10 µL added to each well of the first row of a 96-well qPCR plate (STARLAB). Next, 9 µL PCR grade water (Sigma-Aldrich) was added to rows B-H, and a serial 1:10 dilution was carried out down the plate from rows A-H, so that each column contained a DNA concentration gradient from 100 ng/µL to 1×10^{-5} ng/µL. A separate master mix was created for each column (1-12), containing a different set of primers (Table 2.4) (Sigma-Aldrich) diluted to 454 nM in 2× Luna® Universal qPCR Master Mix (NEB). Each well then had 11 µL of corresponding master mix added, so that the PCR mix had a final volume of 20 µL and a final primer concentration of 250 nM each for forward and reverse primers. A no template control containing no DNA was used in column 11, as well as a no primer control in column 12. The plates were sealed with MicroAmp™ Optical Adhesive Films (Applied Biosystems), and the qPCR was then run on a CFX Connect™ Real-Time PCR Detection System (Bio-Rad) following Luna® Universal qPCR Master Mix Protocol (M3003) (New England BioLabs, 2022): initial denaturation at 95°C for 60 s followed by 45 cycles of 15 s at 95°C (denaturation) and 30 s at 60°C (extension) with a plate read.

Table 2.4. Primers used in Chapter 3

Use	Name	DNA sequence (5' to 3')	Description	Reference
qPCR	*qacA*	F GACCCTTCTGGTACCCAACA R TCCCCATTTATCAGCAAAGG	To amplify *qacA* for qPCR	(Buelow *et al.*, 2017)
	tet(W)	F GGTGCAGTTGGAGGTTGTTT R AAATGACGGAGGGTTCCTTT	To amplify *tet*(W) for qPCR	(Buelow *et al.*, 2017)
	tolC	F CTGAAAGAAGCCGAAAAACG R CGTCGGTAAGTGACCATCCT	To amplify *tolC* for qPCR	(Buelow *et al.*, 2017)
	vanB	F CCTGCCTGGTTTTACATCGT R GCTGTCAATCAGTGCAGGAA	To amplify *vanB* for qPCR	(Buelow *et al.*, 2017)
	tet(Q)	F GCAAAGGAAGGCATACAAGC R AAACGCTCCAAATTCACACC	To amplify *tet*(Q) for qPCR	(Buelow *et al.*, 2017)
	bla$_{\text{TEM}}$	F AAGCCATACCAAACGACGAG R TTGCCGGGAAGCTAGAGTAA	To amplify *bla*$_{\text{TEM}}$ for qPCR	(Buelow *et al.*, 2017)
	erm(B)	F GGTTGCTCTTGCACACTCAA R CTGTGGTATGGCGGGTAAGT	To amplify *erm*(B) for qPCR	(Buelow *et al.*, 2017)
	aph(3”)-III	F CCGGTATAAAGGGACCACCT R CTTTGGAACAGGCAGCTTTC	To amplify *aph(3")-III* for qPCR	(Buelow *et al.*, 2017)
	mecA	F TCCAGGAATGCAGAAAGACC R GGCCAATTCCACATTGTTTC	To amplify *mecA* for qPCR	(Buelow *et al.*, 2017)
	16S rRNA gene	F CAACGCGARGAACCTTACC R ACAACACGAGCTGACGAC	To amplify the 16S rRNA gene for qPCR	(Gloor *et al.*, 2010)
NEBNext Library Prep	NEBNext Adapter	/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCdUACACTCTTTCCCTACACGACGCTCTTCCGATC-s-T	Illumina adapters to be used with the NEBNext Library Prep Kit to prepare and barcode DNA for Illumina sequencing	NEBNext #E7335
	NEBNext Universal PCR Primer	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC-s-T		
	NEBNext Index Primer	CAAGCAGAAGACGGCATACGAGATNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC-s-T		

F = forward primer; R = reverse primer; qPCR = quantitative polymerise chain reaction

2.2.1.1 qPCR data analysis: ARG relative abundance

Data from the qPCR were exported to Microsoft Excel. To calculate the efficiency of the qPCR reaction, a standard curve was generated for each gene following the instructions of the Thermo Fisher qPCR efficiency calculator (Thermo Fisher Scientific, 2017): cycle threshold (Ct) values for each dilution was plotted against the log[DNA amount], and the formula of the standard curve and R-squared (R^2) value was measured. Genes were excluded from further data analysis if the R^2 value was less than 0.95. The qPCR efficiency was calculated using the efficiency calculator (Thermo Fisher Scientific, 2017), and genes were excluded if the efficiency was lower than 80% or greater than 130%. The relative quantity was then calculated for the remaining genes using the following formulae:

$$Ct_{ARG} - Ct_{16S\,rRNA} = \Delta Ct$$

$$Relative\ quantity = 2^{-\Delta Ct}$$

2.2.1.2 qPCR data analysis: estimating stool sample cell count

To estimate the number of bacterial cells per gram of stool, the copy number of the 16S rRNA gene in the stool sample was estimated using a method described by (Alexander *et al.*, 2015). First, amplicons (111 nt) generated with qPCR primers targeting the V6 region of 16S rRNA gene (Table 2.4, Gloor *et al.*, 2010) of *E. coli* MG1655 were cloned into the pJET1.2 cloning vector (Thermo Fisher Scientific), and qPCR was performed for a concentration range of the pJET1.2-16S construct by Siu Fung Stanley Ho (University of Birmingham). Then, the cycle threshold (Ct) values from the pJET1.2-16S construct were plotted against the log[DNA amount] in Microsoft Excel and the formula of the trendline was generated. The Ct values from the 16S rRNA gene qPCR on the stool sample for each amount of template DNA were then inputted into the formula and exponentiated to determine the equivalent amount of pJET1.2-16S construct DNA. This value was then used to calculate the number of 16S rRNA gene copies using the following formula where [length of DNA] = 3085 bp:

$$number\ of\ copies = \frac{[amount\ of\ pJET1.2\text{-}16S\ DNA] \times (6.022 \times 10^{23})}{[length\ of\ DNA] \times (1 \times 10^{9}) \times 650}$$

These values were then divided by the average 16S rRNA gene copy number in bacteria, estimated to be 3.82 by Sun *et al.* (2013), to result in an approximation of the number of cells making up the template DNA for the respective qPCR reaction. Finally, these values were multiplied by the proportion of DNA for each amount of the extracted stool DNA added to the qPCR. An average of these values was calculated, resulting in the estimated number of bacterial cells in 1 g of stool.

2.2.2 Meta3C

Meta3C was carried out on stool sample H1 (Table 2.2) following the protocol from (Foutel-Rodier *et al.*, 2018), described below. Before cross-linking was performed on the stool sample, a spike-in of *E. coli* E3090 and *E. faecium* E745 (Table 2.1) was added to a final concentration of 1% (0.5% each), calculated using the estimated number of cells/g of stool from Section 2.2.1.2 and number of cfu/mL of the bacterial strains calculated as described in Section 2.1.1.1.

2.2.2.1 Cross-linking

Approximately 250 mg of stool was added to 25 mL PBS with a final concentration of 5% methanol-free formaldehyde (Sigma-Aldrich), and vortexed at high speed for 30 s to resuspend. The resuspended stool was then incubated for 30 min at RT with rapid shaking (200-300 rpm), followed by 30 min at 4°C under gentle agitation (33 rpm using a roller mixer (Stuart)). Glycine (Fisher Scientific) was then added to a final concentration of 420 mM to quench remaining formaldehyde and incubated for 5 min at RT with moderate shaking (120 rpm), followed by 15 min at 4°C under gentle agitation. The fixed sample was centrifuged at 4,800 × g for 10 min at 4°C. The supernatant was carefully removed, and the pellet was resuspended in 3 mL sterile distilled water by briefly vortexing. The cells were again centrifuged at 16,000 × g for 5 min at 4°C. As much of the supernatant as possible was carefully removed before flash freezing the pellet using dry ice and 100% ethanol (Fisher Bioreagents). Pellets were stored at -80°C until use.

2.2.2.2 Meta3C library construction

The frozen pellet of cross-linked cells was first thawed on ice for 30 min before being resuspended in 4 mL 1× TE (Tris/EDTA) buffer pH 8.3 (Sigma-Aldrich) supplemented with cOmplete™ mini EDTA-free protease inhibitors (Roche Diagnostics). The suspended pellet was then transferred to four Lysing Matrix E tubes (MP Biomedicals), and run on the FastPrep-24 bead-beater (MP Biomedicals) for three cycles of 8.0 m/s for 20 s, off for 30 s. This run of three cycles was repeated three times, with cooling of the tubes on ice for 5 min between each run. After the last run, the tubes were left on ice for 5 min to allow the beads to settle, and the lysate was then transferred to a new 15 mL conical tube. The remaining beads were washed with 200 µL TE buffer supplemented with protease inhibitors, and remaining liquid was added to the lysate. Sodium dodecyl sulphate (SDS) (National Diagnostics) was then added to the samples to a final concentration of 0.5%, mixed by inversion, incubated for 20 min at 65°C, and then cooled on ice. A digestion mix was prepared in four 15 mL conical tubes containing sterile distilled water with a final concentration of 1× NEB1 digestion buffer (NEB) and 1% Triton X-100 (Sigma-Aldrich). Lysate was then added to each digestion mix tube at a 1:5 dilution and mixed by inversion. A control for nondigested DNA (ND control) was kept by transferring 250 µL from each digestion reaction tube into a separate tube (termed ND). The restriction enzymes MluCI and HpaII were then added at 1,000 units (U) to two digestion reaction tubes each, followed by incubation of all tubes at 37°C for 3 h with moderate shaking. Following incubation, a control for digested DNA (D control) was prepared by transferring 500 µL from both digestion reaction tubes into a separate tube (termed MluCI-D and HpaII-D) for each restriction enzyme. The rest of the digestion reaction mixes were centrifuged at 16,000 × g for 20 min at 4°C, and each pellet was resuspended in 500 µL cold sterile distilled water. Two ligation mixes were prepared in 50 mL conical tubes (Greiner Bio-one) containing cold sterile distilled water with a final concentration of 1× ligation buffer (50 mM Tris-HCl pH 7.4 (Jena Bioscience), 10 mM $MgCl_2$ (Sigma-Aldrich), and 10 mM DTT (Roche Diagnostics)) and 0.1 mg/mL bovine serum albumin (Sigma-Aldrich). The resuspended pellets were added to the ligation mixes so that one mix contained solutions from the HpaII digestion reactions, and the other contained solutions from the MluCI digestion reactions. Adenosine triphosphate (ATP) (Roche Diagnostics) to a final concentration of 1 mM and 250 U of T4 DNA ligase (NEB) were both added to each ligation reaction tube and mixed by inversion. Each ligation reaction tube had a final volume of 16 mL. The ligation reaction tubes were then incubated at 16°C for 4 h. Proteolytic reversal of the cross-links was then carried out by adding 20 µL 0.5 M EDTA (PanReac AppliChem), 20 µL 10% SDS, and 10 µL 20 mg/mL proteinase K (Qiagen) to each control tube (MluCI-D, HpaII-D, ND), and 200 µL 0.5 M EDTA, 200 µL 10% SDS, and 100 µL 20 mg/mL proteinase K to the ligation tubes, followed by overnight incubation of all tubes at 60°C.

The following day, the decrosslinked ligated samples were cooled on ice. A solution of 3 M sodium acetate pH 5.2 (Fisher Scientific) was added to each tube at a final concentration of 300 mM and mixed by inversion before the solutions were transferred to new Corning™ Falcon™ 50 mL Conical Centrifuge Tubes (Fisher Scientific) containing 16 mL isopropanol (Fisher Bioreagents). The solutions were again mixed by inversion and incubated for at least 1 h at -80°C to precipitate the DNA. After incubation, the tubes were centrifuged at 10,000 × g for 20 min at 4°C. The supernatant was carefully removed, and the pellet was resuspended in 900 µL 1× TE buffer and incubated at 37°C for 15 min with shaking. The DNA was then transferred to 2 mL tubes and an equal volume (900 µL) of phenol-chloroform-isoamyl alcohol (10:9:1) mix (Sigma-Aldrich) was added to each tube. Solutions from each control tube were also transferred to new 2 mL tubes containing 1 mL phenol-chloroform-isoamyl alcohol. All tubes were then vortexed at maximum speed for 30 s. Tubes were then centrifuged at 16,000 × g for 5 min at RT before transferring 400 µL of the top phase of each condition to two new tubes each containing 40 µL 3 M sodium acetate. Tubes were mixed by inverting and 1.1 mL ice-cold 100% ethanol was added followed by mixing again and incubating for at least 30 min at -80°C. After incubation, the samples were centrifuged at 16,000 × g for 20 min at 4°C to pellet the DNA. The supernatants were then carefully removed, and the DNA pellets were dried by incubating the open tubes for 10 min at 37°C on a Thermo Mix HCM (SciQuip). DNA was then resuspended in 60 µL Tris-HCl pH 7.4 with RNAse A (Qiagen) at a final concentration of 0.8 mg/mL and incubated at 37°C for 30 min. Samples from the same conditions were then pooled into 1.5 mL DNA LoBind tubes (Eppendorf), resulting in five final samples: an ND control as well as a meta3C library and D control for each restriction enzyme.

The quality and quantity of DNA was assessed by performing gel electrophoresis on the samples and controls on a 1% gel (described in Section 2.1.4) and using the Qubit dsDNA BR Assay Kit (described in Section 2.1.7). The meta3C libraries were purified using the GeneJET PCR purification kit as described in Section 2.1.5. Meta3C libraries were then stored at -20°C until further processing for sequencing.

2.2.2.3 Preparation of the meta3C sequencing libraries

Meta3C sequencing libraries were generated with the NEBNext® Ultra™ II FS DNA Library Prep Kit for Illumina (NEB) following the manufacturer’s protocol. Briefly, 75 ng DNA of each meta3C library in a volume of 26 µL was added to 7 µL NEBNext Ultra II FS Reaction Buffer and 2 µL NEBNext Ultra II FS Enzyme Mix in a 0.2 mL thin wall PCR tube (Thermo Scientific). The reaction was vortexed for 5 s and briefly spun in a MyFuge™ Mini centrifuge (Benchmark Scientific). The tubes were then incubated in a Mastercycler® Pro Thermal Cycler at 37°C to fragment the DNA. At this point, separate aliquots of the meta3C libraries were fragmented for different incubation times: 2.5 min, 5 min, 7.5 min, and 10 min. Fragmented DNA was quantified on a 2200 TapeStation system, as described in Section 2.1.8, to find the optimal time to obtain fragmentation ranging between 300-1,000 bp, which was determined to be 2.5 min. After incubation at 37°C for 2.5 min, the samples were incubated at 65°C for 30 min. NEBNext Adaptor for Illumina (Table 2.4) was diluted 10-fold (to 1.5 µM) in 10 mM Tris-HCl pH 7.4 (Jena Bioscience) with 10 mM NaCl (Sigma-Aldrich). The fragmented samples (total volume 35 µL) were added to 30 µL NEBNext Ultra II Ligation Master Mix, 1 µL NEBNext Ligation Enhancer, and 2.5 µL diluted NEBNext Adaptor for Illumina. The reaction mix was mixed thoroughly using a pipette, and incubated for 15 min at 20°C. Following adapter ligation, 3 µL USER® Enzyme was added, and the samples were incubated for 15 min at 37°C. For clean-up of the adaptor-ligated DNA, 57 µL of thoroughly resuspended AMPure XP Beads (Beckman Coulter) was added to the samples, mixed thoroughly, and incubated at RT for 5 min. The tubes were then placed on a MagnaRack™ Magnetic Separation Rack (Invitrogen). Once the solution cleared, the supernatant was discarded and 200 µL 80% ethanol (Fisher Bioreagents) was added to the beads for 30 s before being discarded. This washing step was repeated for a total of two washes before air drying the beads for 5 min. The tubes were then removed from the magnetic rack and DNA was eluted by adding 17 µL Elution Buffer (Thermo Fisher Scientific) and incubated for 2 min at RT. The tubes were then placed back onto the magnetic rack and once the solution cleared, 15 µL eluted DNA was transferred to clean tubes. For PCR enrichment of the DNA, 25 µL NEBNext Ultra II Q5 Master Mix, 5 µL Universal PCR Primer (Table 2.4), and 5 µL Index Primer from NEBNext® Multiplex Oligos for Illumina® (NEB) (Table 2.4, unique index for each meta3C library) was added to the eluted DNA. The reaction mixes were mixed thoroughly with a pipette before undergoing PCR amplification using the program shown in Table 2.5.

Table 2.5. PCR program for amplification during meta3C sequencing library preparation

Step	Temperate (°C)	Time	Number of cycles
Initial denaturation	98	30 s	1
Denaturation	98	10 s	4
Annealing/Extension	65	75 s	
Final extension	72	5 min	1

PCR = polymerase chain reaction; meta3C = metagenomic chromosome conformation capture; s = seconds; min = minutes

Following amplification, the libraries were cleaned-up with AMP XP Beads as described previously, with a final elution in 30 µL Elution Buffer. Prepared libraries were quantified on a 2200 TapeStation system as described in Section 2.1.8. Prepared sequencing libraries were then sequenced by Genomics Birmingham on an Illumina NextSeq 550 2×150 paired-end platform using a Mid Output Kit v2.5 (300 cycles) (Illumina) with a 1% PhiX spike-in.

2.2.3 Downloading published 3C/Hi-C datasets

Reads from several published 3C/Hi-C gut microbiome studies were downloaded from the short read archive (SRA) database (Sayers *et al.*, 2022) (Table 2.6) using the `fastq-dump` command of the SRA Toolkit v2.10.7 (SRA Toolkit Development Team, 2022) with the `--split-files` option. WGS reads for the spike-in strains were also downloaded in the same way (NCBI BioSample accession numbers listed in Table 2.1).

Table 2.6. Published 3C datasets downloaded in this study

Name	Reference	Accession number
M_3C	(Marbouty *et al.*, 2017)	PRJNA302158
P_HiC	(Press *et al.*, 2017)	PRJNA413092
Y_3C	(Yaffe and Relman, 2020)	PRJNA505354
D_HiC	(DeMaere *et al.*, 2020)	PRJNA377403
K_HiC	(Kent *et al.*, 2020)	PRJNA649316

3C = chromosome conformation capture

2.2.4 The H-LARGe workflow

The H-LARGe workflow (**H**ost-**L**inkage to **A**ntimicrobial **R**esistance **Ge**nes, https://github.com/gregmcc97/H-LARGe) was developed to link ARGs to their bacterial host using meta3C/Hi-C reads. The workflow included the following stages:

2.2.4.1 Generation of metagenomic assemblies

Reads were first processed and assembled as described in Section 2.1.9.1. Taxonomic profiles were generated as described in Section 2.1.9.2. ARGs were identified in the assemblies as described in Section 2.1.9.3 and 2.1.9.4.

2.2.4.2 Identification of cross-linked reads

The first 50 bp of the reads were mapped to their respective assemblies using the Burrows-Wheeler Alignment Tool v0.7.12 (Li and Durbin, 2009) using the `aln` and `sampe` sub-commands. The aligned reads were then filtered to remove those with a mapping quality (MAPQ) <20 using SAMtools. Read pairs where each mate of the pair mapped to a different contig (intercontig reads) were then identified by using SAMtools to filter out reads in the SAM file that mapped in a proper pair, were unmapped, or had an unmapped mate (`view -F 14`), followed by the Unix "grep" command to remove any remaining reads in the SAM file that mapped to the same contig as their mate (`grep -v "="`).

2.2.4.3 Filtering of spurious intercontig reads

A Bash script was written to remove intercontig reads that mapped within the first or last 500 nt of a contig by using the mapping coordinates in the SAM mapping file and the contig lengths in the assembly (Appendix Section A.2). Further filtering was carried out after intercontig reads linking contigs to ARG-contigs were identified (below).

2.2.4.4 Linking ARGs to their microbial hosts

3C/Hi-C intercontig reads where one mate mapped to an ARG-contig were then identified to generate a list of linked contigs for each ARG-contig. Continuing the filtering of spurious intercontig reads, these lists were then filtered so that only contigs that linked at least five times to an ARG-contig were kept. Additionally, to remove potential false cross-links from contigs that contain IS elements, contigs containing IS elements in the assembly were identified as described in Section 2.1.9.3. Identified IS element contigs were then removed from the lists of contigs linked to ARGs.

2.2.4.5 Taxonomic classification of cross-linked contigs

Remaining contigs for each ARG were then taxonomically classified using Kraken2 v2.0.8 (Wood, Lu, and Langmead, 2019) using the prebuilt (September 2018) kraken2-microbial database (available to download from: https://lomanlab.github.io/mockcommunity/mc_databases.html). The contigs were also mapped to NCBI's nt database (Sayers *et al.*, 2022) using BLASTN v2.2.31 (Camacho *et al.*, 2009) in accordance with the BLAST user manual (Camacho *et al.*, 2008). Contigs that aligned with 99% identity to known plasmid sequences using BLASTN were labelled as “Plasmid DNA” and subsequently removed. The Bash script to link ARGs to their hosts is shown in Appendix Section A.3. The pheatmap R package (Kolde, 2019) was used to create a heatmap of the ARG-host associations.

2.2.5 Identification of the genome region mapped to in spike-in genomes

The genomes of the spike-in strains were downloaded from NCBI (accession numbers in Table 2.1) and annotated using Prokka (Seemann, 2014). WGS reads and 3C reads were then mapped to the genomes and an R script (Appendix Section A.4) previously written by Dr Steven Dunn (University of Birmingham) was used to find the annotated region of the genome being mapped to by each read. From the output file, products labelled as “NA” were assigned as intergenic regions. Products labelled as “*IS*” or “*transposase*”, where * denotes a wildcard (i.e. can match any string), were assigned as IS elements. Products labelled as “*hypothetical_protein*” or “*product=putative protein*” were assigned as genes with unknown functions. Remaining products labelled as “*gene*”, “*locus_tag*”, “*db_xref*”, “*protein*”, “*note*”, or “*product*” were assigned as genes with predicted functions. These assignments were performed using a custom Bash script (Appendix Section A.5).

2.2.6 Calculation of proportion of reads mapping near ends of a contig

To calculate the proportion of reads that map within the first or last 500 nt of a contig, a short Bash script (Appendix Section A.6) was written that checked this by using the mapping coordinates in the SAM mapping file and the contig lengths in the assembly. For statistical analysis, the proportion of reads mapping within the first or last 500 nt of a contig was plotted against the proportion of intercontig reads for each sample in GraphPad Prism (for K_HiC, an average of all samples was used). A simple linear regression analysis was performed using GraphPad Prism to calculate the R^2 value, and the significance of the correlation was calculated using Spearman correlation, with a significance threshold of $P \leq 0.05$.

2.3 Chapter 4 methods

2.3.1 Preparation of Hi-C libraries

Hi-C libraries for four human faecal samples (Table 2.2) were prepared using the ProxiMeta™ Hi-C Kit v4.0 (Phase Genomics) as described below. Prior to using the kit, a spike-in of *A. pittii* OB7 (Table 2.1) was added to 100 mg of each sample to a final concentration of 0.5%, calculated using the estimated number of cells/g of stool by Sender, Fuchs, and Milo (2016) (1×10^{11} cells/g), and number of cfu/mL of *A. pittii* calculated as described in Section 2.1.1.1.

2.3.1.1 ProxiMeta Hi-C Kit

For cross-linking, 100 mg of each stool sample (with 0.5% OB7 spike-in added) was added to 1 mL Crosslinking Solution in a 2 mL microcentrifuge tube and incubated at RT with rotation (60 rpm) using a HulaMixer™ Sample Mixer (Invitrogen) for 15 min. Then, 100 µL Quenching Solution was added and the samples were incubated for 20 min at RT with rotation. The samples were then centrifuged at 17,000 × g for 5 min and the supernatants were discarded. The pellets were washed with 1 mL 1× CRB, and the centrifugation process was repeated.

For cell lysis, the cross-linked cells were resuspended in 700 µL Lysis Buffer 1, transferred to a Lysis Tube, and vortexed at RT for 20 min using a the 24-tube adapter on a Vortex-Genie™ 2 (Scientific Industries). The tubes were briefly pulsed for 10 s in a centrifuge to crush any bubbles and the supernatants were transferred to clean 1.5 mL microcentrifuge tubes. The tubes were centrifuged at 17,000 × g for 5 min and the supernatants were discarded. The pellets were resuspended in 500 µL 1× CRB before being centrifuged at 17,000 × g for 5 min. The supernatants were discarded, and each pellet was resuspended in 100 µL Lysis Buffer 2 and incubated at 65°C for

15 min in a Thermo Mix HCM (SciQuip). The samples were then transferred to 0.2 mL PCR tubes and 100 µL Recovery Beads were added before incubation at RT for 10 min. The beads were then washed by placing the tubes on a MagnaRack™ Magnetic Separation Rack (Invitrogen), and once the solution had cleared after around 2 min, the supernatants were removed and discarded. For samples H1 and H3, the beads were first diluted with 1× CRB, vortexed, and mixed using a pipette to break-up viscous glycoprotein impairing the adherence of the beads to the magnet. Once the supernatants were removed, the tubes were removed from the magnetic rack and the beads were gently resuspended in 200 µL 1× CRB and transferred to 1.5 mL microcentrifuge tubes.

DNA digestion and biotinylation was then performed by adding 139 µL Fragmentation Buffer and 11 µL Fragmentation Enzyme to each sample before 1 h incubation at 37°C. The beads were then transferred to 0.2 mL PCR tubes and washed using a magnetic rack as described previously for a total of two washes with 1× CRB. After the second wash, the tubes were removed from the rack and the beads were resuspended in 85 µL HyClone™ Molecular Biology Grade Water. For proximity ligation, 10 µL 10× Ligation Buffer and 5 µL Ligation Enzyme was added to each tube before 4 h incubation at 20°C followed by 10 min at 65°C in a Mastercycler® Pro Thermal Cycler. The cross-links were then reversed by adding 5 µL RX Enzyme and incubating at 65°C overnight.

DNA was then purified by adding 100 µL Recovery Beads to the sample tubes and incubating at RT for 10 min. The tubes were then placed on a magnetic rack, and once the solution had cleared after around 2 min, the supernatants were removed and discarded. Keeping the tubes on the magnetic rack, the beads were gently rinsed with 200 µL Recovery Wash Buffer, which was left on the beads for 30-60 s before being removed and discarded. These rinse steps were repeated for a total of two washes. After the final removal of the Recovery Wash Buffer, the beads were left on the magnet with the caps open to airdry at RT for 15 min. The tubes were then removed from the magnet and the beads were resuspended with 100 µL Elution Buffer and incubated at RT for 5 min. The tubes were then placed back onto the magnet rack, and once the solution had cleared after around 2 min, the DNA-containing supernatant was recovered and transferred to a fresh 0.2 mL PCR tube.

Streptavidin beads were used to enrich for ligated fragments, which were tagged with biotin during the fragmentation and proximity ligation steps. First, 20 µL Streptavidin Beads were transferred to new 0.2 mL PCR tubes (one for each sample). The tubes were then placed on a magnetic rack for 2 min until the solution had cleared. The supernatants were then removed and discarded, and the beads were washed by removing them from the magnet and adding 200 µL Wash Buffer 1. The tubes were added back to the magnet and the washing step was repeated for a total of two washes, then the beads were removed from the magnet and resuspended in 100 µL Bead Binding Buffer. The purified DNA (100 µL) was then added to the washed Streptavidin Beads. The beads were again washed as described previously with Wash Buffer 2 for a total of two washes, then washed once more with Wash Buffer 1 and finally resuspended in HyClone™ Molecular Biology Grade Water. At this point, the concentration of the DNA was measured using the Qubit™ dsDNA HS Assay Kit as described in Section 2.1.7, with vortex mixing immediately prior to measurement.

For sequencing library preparation, up to 500 ng of DNA-containing Streptavidin Beads was transferred to fresh a 0.2 mL PCR tube for each sample and placed on a magnetic rack. Once the solution had cleared, the supernatant was removed and discarded. The tubes were then placed on a pre-cooled 4°C Mastercycler® Pro Thermal Cycler and 40 µL HyClone™ Molecular Biology Grade Water was added. The tubes were cooled to 4°C for 3 min before 4 µL Frag, Repair, & A-Tail Buffer and 6 µL Frag, Repair, & A-Tail Enzyme was added. The tubes were then vortexed for 10 s and 30 µL of solution was pipetted up and down 20 times to ensure the solutions were properly mixed. Fragmentation, end-repair, and A-tailing of the samples was then carried out by incubating at 30°C for 5 min followed by 65°C for 30 min in a Mastercycler® Pro Thermal Cycler with the lid heated to 105°C.

After incubation, 5 µL Universal Adapter was added to the samples and mixed by pipetting. If the DNA amount measured in the previous stage was <10 ng, then the Universal Adapter was diluted 1 in 5 in HyClone™ Molecular Biology Grade Water to 3 µM prior to addition to the sample. Otherwise, the provided Universal Adapter concentration (15 µM) was used. Then, 20 µL Adapter Ligation Mix was added and mixed by pipetting before incubating at 20°C for 15 min in a Mastercycler® Pro Thermal Cycler with the heated lid turned off. The beads were then washed with 200 µL Wash Buffer 2 for a total of two times as described previously, followed by one wash with 200 µL Wash Buffer 1, one wash with 200 µL HyClone™ Molecular Biology Grade Water, and finally the beads were resuspended in 20 µL HyClone™ Molecular Biology Grade Water.

For on-bead library amplification, 5 µL PCR Primer Mix (different primer for each sample) was added to the samples and mixed by pipetting thoroughly. Then, 25 µL Hot Start PCR Mix was added, and the libraries were amplified using 12 PCR cycles of the program described in Table 2.7. If the DNA carried into the sequencing library preparation stage was <20 ng, then the number of cycles was increased to 16.

Table 2.7. PCR program for amplification of the Hi-C libraries

Step	Temperate (°C)	Time (s)	Number of cycles
Initial denaturation	98	45	1
Denaturation	98	15	12*
Annealing	60	30	
Extension	72	30	
Final extension	72	60	1

PCR = polymerase chain reaction; s = seconds; *if less than 20 ng DNA was carried into the sequencing library preparation stage, the number of cycles was increased to 16

The amplified libraries were then size selected to remove unwanted high molecular weight DNA and very small fragments. The tubes were placed on a magnetic rack, and once the solution had cleared, the library-containing supernatants were transferred to new 0.2 mL PCR tubes. Recovery Beads were thoroughly resuspended, and 55 µL was added to each tube to bind unwanted high molecular weight fragments. The tubes were incubated at RT for 10 min and then placed on a magnetic rack. Once the solution had cleared after 2 min, the supernatants (105 µL) were each transferred to a new 0.2 mL PCR tube containing 17.5 µL Recovery Beads to bind the library and leave unwanted small fragments in the supernatant. The tubes were incubated at RT for 10 min and then placed on a magnetic rack. Once the solution had cleared, the supernatants were removed, and the beads were gently rinsed with 200 µL Recovery Wash Buffer for 30-60 s. The rinsing steps were then repeated for a total of two rinses, and the beads were air dried on the magnet with the cap open for 15 min at RT. Finally, the tubes were removed from the magnet, thoroughly resuspended in 30 µL Elution Buffer, and incubated at RT for 5 min. The tubes were then placed on a magnetic rack, and once the solution had cleared, the Hi-C library-containing supernatants were transferred to clean 1.5 mL DNA LoBind tubes.

The Hi-C libraries were then quantified using a Qubit™ dsDNA HS Assay Kit as described in Section 2.1.7 and 2200 TapeStation as described in Section 2.1.8.

2.3.1.2 Sequencing of the ProxiMeta Hi-C libraries

Before full-depth sequencing, the generated Hi-C libraries were first sent to Novogene (Cambridge, United Kingdom) for shallow sequencing (<10 million reads per sample) on an Illumina NovaSeq 6000 2×150 bp paired-end platform (Illumina). The shallow reads were analysed to confirm the libraries were high-quality (described below), and then the libraries were deep sequenced by Novogene on an Illumina NovaSeq 6000 2×150 bp paired-end platform.

2.3.2 Preparation and sequencing of shotgun metagenomic libraries

Before DNA extraction of the stool samples for shotgun metagenomic sequencing, a spike-in of *A. pittii* OB7 (Table 2.1) was added to 250 mg of each sample to a final concentration of 0.5%, calculated using the estimated number of cells/g of stool by Sender, Fuchs, and Milo, (2016) (1×10^{11} cells/g), and number of cfu/mL of *A. pittii* calculated as described in Section 2.1.1.1. DNA was then extracted (as described in Section 2.1.6), followed by quantification using a Qubit™ dsDNA BR Assay Kit (described in Section 2.1.7) and 2200 TapeStation instrument (described in Section 2.1.8). The extracted DNA samples were then sent to Novogene for shotgun metagenomic library preparation and sequencing on an Illumina NovaSeq 6000 2×150 bp paired-end platform (Illumina).

2.3.3 Quality assessment of shallow Hi-C reads

The preliminary shallow sequenced Hi-C reads were analysed to confirm they were high-quality before full-depth sequencing of the libraries. To assess the quality, the shallow reads were processed as described in Section 2.1.9.1. The first 50 bp of the reads for each sample were then mapped, as described in Section 2.2.4.2, to the G_3C metagenomic assembly generated in Chapter 3, and the intercontig reads were identified. To estimate the proportion of intercontig reads, the number of intercontig reads was divided by the total number of reads that mapped to G_3C with a MAPQ score >20. This estimation method was first tested on the P_HiC and D_HiC datasets (Table 2.6) mapped to the G_3C assembly, so that the estimated values could be compared to the true proportion of intercontig reads found in Chapter 3. The taxonomic profiles of the shallow reads were also analysed as described in Section 2.1.9.2.

2.3.4 H-LARGe version 2 workflow

The H-LARGe workflow described in Section 2.2.4 was further developed and optimised for Chapter 4 to implement metagenomic binning (described below and at https://github.com/gregmcc97/H-LARGe). First, the Hi-C and shotgun metagenomic reads were processed, and the shotgun metagenomic reads were assembled as described in Section 2.1.9.1. The taxonomic profiles were analysed (Section 2.1.9.2), and ARG-contigs identified (Section 2.1.9.3) as previously described. Then, the intercontig reads (reads originating from cross-linked DNA) were identified and filtered as described in Sections 2.2.4.2 and 2.2.4.3, respectively. For comparing the proportion of intercontig reads mapping within the first or last 500 nt of a contig in each dataset, a one-way analysis of variance (ANOVA) with Tukey's multiple comparisons test and Geisser-Greenhouse correction was performed using GraphPad Prism.

2.3.4.1 Clustering of contigs into bins

To improve the taxonomic classifications of hosts linked to ARGs by Hi-C, the contigs were clustered into bins using the cloud-based ProxiMeta™ Hi-C Metagenomic Deconvolution Platform (Phase Genomics Inc., 2022). The bins were then downloaded and assessed for quality (completeness and contamination) using CheckM v0.6.0 (Parks *et al.*, 2015). Quality scores were calculated using the following equation, where completeness is the proportion of marker genes present, and contamination is the percentage of marker gene overrepresentation according to the CheckM output:

$$quality\ score = [completeness] - 5 \times [contamination]$$

Bins were then filtered to discard those that had a completeness <50, contamination >10, or quality score <50.

2.3.4.2 Taxonomic classification of bins

Bins were taxonomically classified using GTDB-Tk v 2.1.0 (Chaumeil *et al.*, 2022) classify workflow with the R207_v2 package of the GTDB (Parks *et al.*, 2022). The classifications were manually checked and edited to remove placeholder names (given for genomes that that do not cluster into groups with taxonomically valid named representatives in the GTDB (Parks *et al.*, 2018)), and assign the bins with the lowest ranking validly or effectively published name given by GTDB-Tk.

2.3.4.3 Identification of plasmids in the assemblies

Plasmid bins were also generated by the ProxiMeta™ Hi-C Metagenomic Deconvolution Platform along with quality assessments based on their comparisons to the closest matching reference plasmid in the NCBI reference sequence (RefSeq) database (O'Leary *et al.*, 2016) (completeness = percentage of reference covered by plasmid bin; contamination = percentage of plasmid bin that has mutually-conflicting alignments to reference). Quality scores were calculated as described in Section 2.3.4.1, and the plasmid bins were then filtered to discard bins that had a completeness <50, contamination >10, or quality score <50. As well as plasmid bins, the ProxiMeta™ Hi-C Metagenomic Deconvolution Platform also annotated contigs in the metagenomic assemblies as plasmids if they aligned to plasmid sequences in the RefSeq database. From these data, contigs were considered plasmid sequences if they mapped to the reference plasmids with >80% coverage and >90% identity. Plasmid contigs and bins were named according to the accession number and sequence name of their closest reference match. Plasmid contigs were also found using ABRicate with the PlasmidFinder database (Carattoli *et al.*, 2014) as described in Section 2.1.9.3. The relative abundance of all plasmid sequences was calculated using the same method described in Section 2.1.9.4, but calculating RPKM based on reads mapping to the whole contig rather than a single extracted gene sequence.

2.3.4.4 Linking ARGs to hosts

Bins were first screened for ARGs using ABRicate as described in Section 2.1.9.3. Then, ARGs were linked to their microbial hosts using the Hi-C data as described in Section 2.2.4.4, with an additional filtering step for H-LARGe v2 of also removing links to other ARG-contigs. Once ARGs were linked to contigs, the linked host was classified based on the taxonomic classification of the bins the linked contigs were present in. Links to contigs that were in discarded bins were labelled as “Discarded bin”, and links to contigs not clustered into a bin were labelled as “Unbinned”. The Bash script to link ARGs to their host for H-LARGe v2 is shown in Appendix Section A.7. The pheatmap R package (Kolde, 2019) was used to create a heatmap of the ARG-host associations.

2.3.4.5 Linking plasmids to hosts

Plasmids were linked to their host in the same way as ARGs described previously but using the plasmid contigs identified in Section 2.3.4.3. For plasmid bins, links to all contigs in the plasmid bin were combined.

2.3.5 Analysis of pOXA58_100020 plasmid contigs

The *A. pittii* OB7 pOXA58_100020 plasmid contigs present in metagenomic assemblies were compared to the complete pOXA58_100020 sequence (GenBank accession CP027253) using the BLASTn web interface (National Library of Medicine, 2022). The search function in Gene Construction Kit v4.5 (Gross, 1990; Textco BioSoftware, 2022) was used, with assistance from Dr Robert Moran (University of Birmingham), to find and annotate p*dif* sites, using sequences published by Blackwell and Hall (2017) as references. The annotated plasmid sequences were labelled to scale using Photoshop CC v20.0.0 (Adobe) (Adobe, 2018).

2.4 Chapter 5 methods

2.4.1 Enrichment of ARG hosts

The hosts of ARGs in the stool samples were enriched during culturing in a MACS MG-500 Anaerobic Chamber Workstation (Don Whitley) set to 37°C and connected to a 5% Carbon Dioxide, 5% Hydrogen/Nitrogen (Anaerobic) Cylinder (BOC). Modified Gifu Anaerobic Medium (mGAM) broth (Nissui Pharmaceutical) and PBS were reduced in the anaerobic chamber for at least 24 h prior to use. A stool sample aliquot for each sample (Table 2.2) stored at -80°C in PBS containing 1 mM DTT and 20% glycerol was transferred to the anaerobic chamber. After thawing, 10 µL stool was added to a 30 mL universal container containing 2 mL mGAM broth supplemented with the respective antibiotic for the target ARG (antibiotic concentrations listed in Table 2.8). In parallel, 10 µL stool was also added to mGAM with no antibiotic as a control. After 24 h, DNA was extracted from 1 mL of each overnight culture using the FastDNA™ Spin Kit for Soil, as described in Section 2.1.6, and subsequently used for qPCR to check for enrichment of ARGs (described below). The overnight cultures supplemented with antibiotic were also serially diluted in reduced PBS, and 100 µL of a range of neat to 10^{-7} dilutions were spread onto mGAM agar (Nissui Pharmaceutical) plates containing antibiotic, which had been reduced in the anaerobic chamber for at least 24 h beforehand. The plates were kept in the cabinet at 37°C for at least 24 h to grow single colonies. Additionally, 10 µL of each overnight culture was added to 2 mL fresh mGAM broth supplemented with antibiotic (or without antibiotic for the control) in a clean 30 mL universal container and the whole process was repeated for a further 24 h.

For enrichment for a vancomycin resistance gene in sample H3, the sample was also treated with an equal volume of 70% (v/v) ethanol for 4 h at RT in aerobic conditions as described by Browne *et al.* (2016) to kill vegetative cells. The sample was then centrifuged at 10,000 × g for 3 min. The supernatant was discarded, and the pellet was resuspended in PBS. This wash process was repeated three times before the sample was transferred to the anaerobic cabinet and added to vancomycin-supplemented mGAM as described above.

Table 2.8. Antibiotics used for enrichment of antibiotic resistant hosts in stool

Target ARG antibiotic class	**Antibiotic (and supplier)**	**Antibiotic stock diluted in:**	**Final conc. in mGAM (µg/mL)**
Tetracycline	Tetracycline (Sigma-Aldrich)	ddH_2O*	8
Macrolide	Erythromycin (Serva)	Ethanol (100%)	8
Aminoglycoside	Kanamycin (Carl Roth)	ddH_2O*	8
	Streptomycin (Agilent Technologies)	ddH_2O*	8
Beta-lactam	Ampicillin (Fisher Bioreagents)	ddH_2O*	8
Chloramphenicol	Chloramphenicol (Alfa Aesar)	Ethanol (100%)	16
Glycopeptide	Vancomycin (Alfa Aesar)	ddH_2O*	8
	Colistin (Alfa Aesar)	ddH_2O*	8

*filter sterilised using a 0.22 µm sterile syringe filter after antibiotic dissolved; ARG = antimicrobial resistance gene; ddH_2O = double distilled water; conc. = concentration mGAM = modified Gifu Anaerobic Medium

2.4.1.1 qPCR to measure enrichment of ARG hosts

Enrichment of ARG hosts in the stools samples was measured using qPCR. Primers for qPCR (Table 2.9) were designed using Primer3 (web v4.1.0 https://primer3.ut.ee/) (Untergasser *et al.*, 2012) using the following settings: product size ranges 80-200 bp, primer size 18-27 bp (optimal size 20 bp), primer melting temperature (Tm) 63-66°C (optimal Tm 65°C, max difference 5°C), and primer GC-content (guanine-cytosine content) 20-65% (optimal GC-content 50%). Primers were manufactured by Integrated DNA Technologies.

Quantitative PCR was used to measure the relative quantities of ARGs in extracted DNA from Section 2.4.1 following enrichment. Whilst on ice, 10 µL 2× Luna® Universal qPCR Master Mix, 8 µL PCR grade water, 1 µL extracted DNA (<100 ng), 0.5 µL 10 µM forward primer, and 0.5 µL 10 µM reverse primer (Table 2.9) was added to single wells of a 96-well qPCR plate. Each target ARG was run in duplicate or triplicate for each stool sample. The plate was sealed with a MicroAmp™ Optical Adhesive Film (Applied Biosystems) and the qPCR program was run on a CFX Connect™ Real-Time PCR Detection System in accordance with the Luna® Universal qPCR Master Mix Protocol (M3003) (New England BioLabs, 2022), as described in Section 2.2.1. The abundance of the 16S rRNA gene was also quantified in triplicate for each DNA extract using the same method with 16S rRNA qPCR primers (same as in Table 2.4) and an annealing temperature of 51°C. Following qPCR, an average of the replicate Ct values was calculated, and the relative quantity of the ARGs was then calculated in Microsoft Excel using the following formulae:

$$Ct_{ARG} - Ct_{16S\,rRNA} = \Delta Ct$$

$$Relative\ quantity = 2^{-\Delta Ct}$$

Table 2.9. qPCR primers used in Chapter 5

Name		DNA sequence (5' to 3')	Annealing temperature used (°C)	Description
T01	F	ACGCACACCCTCCATGACGG	59	To amplify *tet*(32)_2 for qPCR
	R	TGAAGTGCCGCCGAATCCGT		
T02	F	GGCACTAAGAACAGCGCGGG	59	To amplify *tet*(40)_1 for qPCR
	R	GCGCGCCACAGTCTTTTCCA		
T03	F	GCACGCGCCAAAGAGATCCG	59	To amplify *tetA*(46)_1 for qPCR
	R	GCGGATGATCTCGAGCGGATGT		
T04	F	GGGTCTACTTTTATCAGTGGCTCGCT	59	To amplify *tetA*(P)_2 for qPCR
	R	TGCTCCTATCTGCCCTGCTTGT		
T05	F	CTGCTGGAGTGCTGGCTGGA	59	To amplify *tetB*(P)_1 for qPCR
	R	GCAGGGAGAGGCGAAGGTCT		
T06	F	AAGCGGGTCACTGTCGGAGA	59	To amplify *tet*(M)_12 for qPCR
	R	CCCTCCCTCTGCTGCAAACG		
T07	F	TGGAGCGTCAAAGGGGAATCACT	59	To amplify *tet*(O)_2 for qPCR
	R	CGGGTCTGTGCCTGTATGCC		
T08	F	TTGCGCTTGTATGCCTTCCTTTGC	59	To amplify *tet*(Q)_1 for qPCR
	R	**same as T09 F**		
T09	F	TTGCGGAAGTGGAGCGGACA	59	To amplify *tet*(Q)_2 for qPCR
	R	ACGCTCCAAATTCACACCGGC		
T10	F	CAGCCCTGTCACCGCATCCA	59	To amplify *tet*(W)_1 for qPCR
	R	GGGAGGAACAGCAGCGGGTT		
T11	F	CGGCAGCGCAAAGAGAACGG	59	To amplify *tet*(W)_3 for qPCR
	R	ACGACCTCCAACTGCACCCG		
T12	F	GCAAGCGCCCATTACCCATAACA	59	To amplify *tet*(X)_2 for qPCR
	R	TCAAGGCATCCACCAACCCACT		
T13	F	**same as T01 R**	59	To amplify *tet*(32)_2 for qPCR (H2 only)
	R	TCCCCAGCCATATAAGCCCTGC		
T14	F	CCACCTCAGCTTCTCAACGCG	59	To amplify *tet*(B)_2 for qPCR
	R	ACCACCAATAATAGGCCCCGCT		
T15	F	ATTTGGACGACGGGGCTGGC	59	To amplify *tet*(M)_1 for qPCR
	R	TGGAGAAATCCCTGCTCGGTGT		
T16	F	ATCGGCGGGCCTATTGTGGG	59	To amplify *tet*(40)_1 for qPCR (H3 only)
	R	TCCCGCTTCTTTCCCGCCAC		
T17	F	TCCGAGGGAGTACTAACAGGGCT	59	To amplify *tet*(44)_1 for qPCR
	R	GGAGCAAGGTTTGTATGGCTGGGA		
T18	F	TCGAAGTGCCACCCAACCCT	59	To amplify *tet*(Q)_3 for qPCR
	R	TCGATCTGCAACCCTGTCCCT		
T19	F	TCCTGGGGCGCATAGAGGGT	59	To amplify *tet*(W)_2 for qPCR
	R	CTTCCGCTCATTGGCCCCGA		
T20	F	ACCGACTCCTTCTCCCATGGC	59	To amplify *tet*(L)_2 for qPCR
	R	CGTAATGGTTGTAGTTGCGCGCT		

M01	F	GGAACATCTGTGGTATGGCGGGT	59	To amplify *erm*(B)_12 for qPCR
	R	TGCACACTCAAGTCTCGATTCAGCA		
M02	F	CGACAACTTGAACATTTCGGGCATCA	58	To amplify *erm*(F)_3 for qPCR
	R	GGGGCAGGCAAGGGGTTTCT		
M03	F	TGCTTATGTTGTAAGGTATGCTGCCA	58	To amplify *erm*(G)_1 for qPCR
	R	AAAGAGATGTAATTTTGTTACGGCGAT		
M04	F	ACGCAATCTACACTAGGCATGGGA	58	To amplify *erm*(Q)_1 for qPCR
	R	TGTGGCTAGTTATGGAGAAAGGTTC		
M05	F	TCCACTAGCACACCAATGGCAGG	58	To amplify mef(A)_2 for qPCR
	R	TGGGCAGGGCAAGCAGTATCA		
M06	F	TGGAAAGGGCTGCGGAGGAA	58	To amplify msr(D)_2 for qPCR
	R	AGCTTCTACTTTTCCTAAGGCCGCA		
M07	F	TTGAAAGCCGTGCGTCTGACA	58	To amplify *erm*(B)_6 for qPCR
	R	AGAGCAACCCTAGTGTTCGGTGA		
M08	F	ACTTTGCACTCGATCACCCAGCT	58	To amplify mef(A)_2 for qPCR (H2 only)
	R	TGCCATCGACGTATTGGGTGCT		
M09	F	CCTTCGCGCTTACCTCGGCT	58	To amplify mph(A)_2 for qPCR
	R	CTCAAGCTCCATGGCCCGCT		
M10	F	CGCCCTCCGCCTTCAGTACT	58	To amplify msr(D)_3 for qPCR
	R	**same as M06 F**		
M11	F	TCACAGCACCCAATACGTCGATGG	58	To amplify mef(A)_2 for qPCR (H3 only)
	R	TGGAACAGCTTTTCATACCCCAGC		
M12	F	TGGAACGGCTTTTCACACCCC	58	To amplify mef(A)_3 for qPCR
	R	GCCTGCACATTTCGTCAGCTGT		
M13	F	AAATTCACCAACTGATATGTGGCTAGT	58	To amplify *erm*(Q)_1 for qPCR (H4 only)
	R	**same as M04 F**		
A01	F	TCTTTTCCCACGGCGACCTGG	59	To amplify *aph(3')-III*_2 for qPCR
	R	TACCACTTGTCCGCCCTGCC		
A02	F	GCCAGAACATGAATTACACGAGGGC	59	To amplify *aac(6')-aph(2")*_1 for qPCR
	R	TGCCACACTATCATAACCACTACCGA		
A03	F	GCTCATGTTCGGGCAGCTCC	59	To amplify *aac(6')-Im*_1 for qPCR
	R	TGCGGTGATTCTTGACCCACGA		
A04	F	GGGCATCCTTTTCCTTTTCACTCAACT	59	To amplify *aph(2")-Ib*_3 for qPCR
	R	CCCCTGCGGTAGTATATCAAAGTGACC		
A05	F	CCCTGCGGTAGTGTATCAAAGTGACC	59	To amplify *aph(2")-Ib*_2 for qPCR
	R	**same as A04 F**		
A06	F	**same as A03 R**	59	To amplify *aac(6')-Im*_1 for qPCR (H3 only)
	R	ACACTCTCCATTCCATCAGCACACA		
A07	F	TGCCATGGAACGGATCGGAAGT	59	To amplify *aph(2")-Ig*_1 for qPCR
	R	CGCGCACAGATGGCCGTTTG		
A08	F	ACAAACACCACATTAGATAGCCCTCCT	59	To amplify *npmA*_1 for qPCR
	R	TCGTGTGCATATAGATTTGGGTACTGGA		
A09	F	CGAAACCTCCCGGAACAATGCC	59	To amplify *ant(6)-Ia*_3 for qPCR
	R	GCTTCGCATGATTTCCTGGCTGA		

A10	F	TCCAACTGATCTGCGCGCGA	59	To amplify *ant(3")-Ia*_1 for qPCR
	R	GGTACAGCGCAGTAACCGGC		
A11	F	GGGGTTCCTCAGATGCGGCA	59	To amplify *aph(6)-Id*_1 for qPCR
	R	ATGGCGCAATGGGAGGGGAG		
A12	F	TCGGTGCGCTCTTGGTCGAG	59	To amplify *aph(3")-Ib*_5 for qPCR
	R	CCGCAATGCCGTCAATCCCG		
B01	F	ACCGCCACACCAATTTCGCC	59	To amplify *cfxA3* or *cfxA5* for qPCR
	R	CGACAAAAGATAGCGCAAATCCTCCT		
B02	F	CTCTTGCCCGGCGTCAACAC	59	To amplify *bla*$_{\text{TEM-1B}}$_1 for qPCR
	R	AGTTGGGTGCACGAGTGGGT		
B03	F	CCGCCACACCAATTTCACCAGG	59	To amplify *cfxA6*_1 for qPCR
	R	ACAAAAGGTAGCGCGAATCCTCCA		
B04	F	CCGCCTGATTGCTCCACTGC	59	To amplify *cepA*_6 for qPCR
	R	AACCCGATGTCATTGCAACCGA		
C01	F	AGTTCACCCTCCTGATTAATTGCCGT	59	To amplify *catS*_1 for qPCR
	R	ACCCTGCGATGCTGTATTATCTTGC		
C02	F	CGGTTGCGGCTTCGGTTGTA	59	To amplify *cfr*(C)_2 for qPCR
	R	CGGGGCGGTCAGATCATGCA		
C03	F	GCTTCCATACCGTTGCATATCATTCGC	59	To amplify *catS*_1 for qPCR (H3 only)
	R	ACGGCAATCAATCAGGAGGGTGA		
C04	F	CGCCCTGCACATATAGTATGACGGT	59	To amplify *cat*_2 for qPCR
	R	TGCCGTCCTGAACTCTTCGTGT		
V01	F	TGCCGCCGAGAGAATGGGAA	59	To amplify *vanB* for qPCR
	R	TGCACCCCGTATGGCCATCA		

qPCR = quantitative polymerase chain reaction; F = forward primer; R = reverse primer

2.4.2 Colony PCR to find ARG-positive isolates

Following enrichment of ARG-hosts in stool samples and growth on mGAM agar plates supplemented with antibiotics described in Section 2.4.1, colony PCR was performed to check for presence of an ARG. Primers for colony PCR (Table 2.10) were designed using Primer3 web v4.1.0 (Untergasser *et al.*, 2012) using the following settings: product size ranges 501-1,000 bp, primer size 18-23 bp (optimal size 20 bp), primer Tm 57-62°C (optimal Tm 59°C, max difference 5°C), and primer GC-content 40-60% (optimal GC-content 50%). Primers were manufactured by Integrated DNA Technologies.

Single colonies were randomly selected and patch-plated onto a fresh antibiotic-supplemented mGAM agar plate using sterile wooden toothpicks in an anaerobic chamber. After 24-48 h growth at 37°C in the anaerobic chamber, colonies were picked with a sterile toothpick and transferred to 10 µL molecular biology grade water in a PCR tube. The colonies were boiled for 10 min at 95°C to lyse the cells. The lysates were used for colony PCR, which was performed using DreamTaq Green PCR Master Mix as described in Section 2.1.3, using the annealing temperatures shown in Table 2.10. Extracted DNA from Section 2.4.1.1 was used as a positive control for the enriched ARGs, and a colony of *A. pittii* OB7 (Table 2.1) was used as negative control. A no template control containing no DNA was also present for all colony PCRs.

Table 2.10. Colony PCR primers used in Chapter 5

Name		DNA sequence (5' to 3')	Annealing temperature used (°C)	Description
T21	F	CAATGAGTTGTTGGACGCCA	58	To amplify *tet*(Q)_2 for colony PCR
	R	TCGGTTCGAGAATGTCCACA		
T22	F	CATCCGAAAATCTGCTGGGG	58	To amplify *tet*(M)_12 for colony PCR
	R	ACTGTTGAACCGAGCAAACC		
T23	F	CGAGATCATCCGCAAAAGCA	58	To amplify *tetA*(46)_1 for colony PCR
	R	TGTTGAGCAAAACCGGGAAG		
T24	F	ACTCCGGGACACATGGATTT	58	To amplify *tet*(Q)_1 for colony PCR
	R	CTGTTTGAGACTGATGCCGG		
T25	F	GGTTAGGGGCAAGTTTTGGG	58	To amplify *tet*(B)_2 for colony PCR
	R	ATCCCACCACCAGCCAATAA		
T26	F	CGGCATCCCCAATCATTGTT	58	To amplify *tet*(X)_2 for colony PCR
	R	CGAAAGAGACAACGACCGAG		
M14	F	TCCGAAATTGACCTGACCTGA	58.6	To amplify erm(F)_3 for colony PCR
	R	GCAGGCAAGGGGTTTCTTAC		
M15	F	TCCGAAATTGTCCTGACCTGA	58	To amplify *erm*(F)_4 for colony PCR
	R	**same as M14 R**		
M16	F	ATACGTGAGGAGGAGCTTCG	58	To amplify *mph*(A)_2 for colony PCR
	R	GCCGATACCTCCCAACTGTA		
M17	F	TGGCGTGTTTCATTGCTTGA	58	To amplify *erm*(B)_18 for colony PCR
	R	GCATTTAACGACGAAACTGGC		
A23	F	CGCGACTGGAGAACATGATG	57	To amplify *aph(6)-Id*_1 for colony PCR
	R	ATGTTCATGCCGCCTGTTTT		
B05	F	CCTGAACCTGTCTTATGCGC	57.8	To amplify *cfxA3* and *cfxA5* for colony PCR
	R	TTGTCCTGGCGAAATTGGTG		
B06	F	CTGCAACTTTATCCGCCTCC	57.8	To amplify *bla*$_{\text{TEM-1B}}$_1 for colony PCR
	R	GTGCACGAGTGGGTTACATC		
B08	F	CTGGTAGTTGCGCAGAACAG	57.8	To amplify *cepA*_6 and *cepA-49*_1 for colony PCR
	R	CCGATGTCATTGCAACCGAT		
16S 27F		AGAGTTTGATCCTGGCTCAG	50	To amplify the 16S rRNA gene for colony PCR (Lane, 1991)
16S 1492R		GGTTACCTTGTTACGACTT		

PCR = polymerase chain reaction; F = forward primer; R = reverse primer

2.4.2.1 16S rRNA gene sequencing

Colonies that were PCR-positive for ARGs were selected for taxonomic identification based on ARG-profile and colony morphology. For taxonomic identification, colony PCR, as described previously, was used to amplify the 16S rRNA gene (primers and annealing temperature in Table 2.10). Following amplification, the amplicons were purified as described in Section 2.1.5, and sent for Sanger sequencing using the Eurofins Genomics TubeSeq service. The resulting sequences were classified using the BLASTn web interface (National Library of Medicine, 2022) with the 16S ribosomal RNA sequences (Bacteria and Archaea) database (Sayers *et al.*, 2022).

2.4.3 Whole-genome sequencing of isolates

Isolates were selected for sequencing based on their ARG-profiles and 16S rRNA gene identification result. Selected isolates were streaked on mGAM agar and grown at 37°C in an anaerobic chamber for 24 h. Single colonies were then transferred to 3 mL mGAM broth and grown for 24 h. After growth, 1 mL of each culture was transferred to a cryotube with 15% (v/v) glycerol and stored as a stock at -80°C. The remainder of the overnight culture was used for DNA extraction, described below.

2.4.3.1 Genomic DNA extraction

Genomic DNA was extracted using the Wizard® Genomic DNA Purification Kit (Promega). First, 1 mL of overnight culture (described above) was added to a 1.5 mL microcentrifuge tube centrifuged at 16,000 × g for 2 min and the supernatant was discarded. At this point, for Gram-positive bacteria (as identified by 16S rRNA gene classification), the pellets were resuspended in 480 µL 50 mM EDTA (PanReac AppliChem) and 120 µL 50 mg/mL Lysozyme from chicken egg white (Sigma-Aldrich) followed by 45 min incubation at 37°C, centrifugation at 16,000 × g for 2 min, and removal of the supernatant. For Gram-negative bacteria, the lysozyme steps were skipped. For all cells, the pellets were then resuspended in 600 µL Nuclei Lysis Solution using a pipette and incubated for 5 min at 80°C. The tubes were cooled to RT before the addition of 3 µL RNase Solution and incubation at 37°C for up to 60 min. The tubes were again cooled to RT, and then 200 µL Protein Precipitation Solution was added and mixed by inversion followed by incubation on ice for 5 min and centrifugation at 16,000 × g for 3 min. The supernatants were transferred to clean 1.5 mL microcentrifuge tubes containing 600 µL 100% isopropanol, mixed by inversion, and centrifugated at 16,000 × g for 2 min. The supernatants were discarded and 600 µL 70% (v/v) ethanol was added and mixed by inversion. The tubes were again centrifugated at 16,000 × g for 2 min. The supernatants were removed, and the DNA was air dried for 15 min. The DNA pellet was resuspended in 100 µL Elution Buffer (Thermo Fisher Scientific) and incubated at 65°C for 1 h. The DNA was quantified using a Qubit™ dsDNA BR Assay Kit (Section 2.1.7), checked on a 1% (w/v) agarose gel (Section 2.1.4), and stored at -20°C in 1.5 mL DNA LoBind tubes (Eppendorf).

2.4.3.2 Short-read whole-genome sequencing

For short-read sequencing, the extracted DNA samples were diluted to similar amounts in Elution Buffer (Thermo Fisher Scientific) and sent to MicrobesNG (Birmingham, UK) for sequencing on an Illumina HiSeq 2500 2×250 bp paired-end platform (Illumina).

2.4.3.3 Long-read whole-genome sequencing

Following preliminary assembly of the short-reads (described below), some isolates were selected for long-read sequencing using the same extracted DNA. Extracted DNA was prepared for long-read sequencing using the Ligation Sequencing Kit

(SQK-LSK109) (Oxford Nanopore Technologies). Approximately 1 µg DNA in 48 µL HyClone™ Molecular Biology Grade Water was added to 3.5 µL NEBNext FFPE DNA Repair Buffer, 3.5 µL Ultra II End-prep reaction buffer, 3 µL Ultra II End-prep enzyme mix, and 2 µL NEBNext FFPE DNA Repair Mix from the NEBNext® Companion Module for Oxford Nanopore Technologies® Ligation Sequencing Kit (NEB). The reaction mixes were mixed using a pipette and incubated at 20°C for 5 min followed by 65°C for 5 min in a Mastercycler® Pro Thermal Cycler and then transferred to a clean 1.5 mL DNA LoBind tubes. AMPure XP Beads (Beckman Coulter) were resuspended by vortexing, and 60 µL were added to each tube and mixed by flicking, then incubated for 5 min at RT on a HulaMixer™ Sample Mixer set to 60 rpm. The tubes were then placed on a MagnaRack™ Magnetic Separation Rack (Invitrogen) and, once cleared, the supernatants were discarded. Whilst still on the magnetic rack, 200 µL 70% (v/v) ethanol was added to the pellet and then removed and discarded. This wash step was repeated for a total of two washes, followed by 30 s of air-drying. The tubes were then removed from the rack and the pellets were resuspended in 25 µL molecular biology grade water, placed back on the magnetic rack, and once cleared, 25 µL of elute was transferred to a clean 1.5 mL DNA LoBind tube. DNA was quantified using a Qubit™ dsDNA BR Assay Kit (Section 2.1.7). The end-prepped DNA was then barcoded using the Native Barcoding Expansion 1-12 (EXP-NBD104) and 13-24 (EXP-NBD114) Kits (Oxford Nanopore Technologies). Unique barcodes for each sample were thawed on ice, and up to 500 ng of each end-prepped sample in 22.5 µL molecular biology grade water was added to 2.5 µL Native Barcode (unique for each sample) and 25 µL Blunt/TA Ligase Master Mix (NEB) and incubated for 10 min at RT. The DNA was then purified with 50 µL AMPure XP Beads as previously described, eluted in 26 µL molecular biology grade water, and quantified using a Qubit™ dsDNA HS Assay Kit (Section 2.1.7). Equimolar amounts of each barcoded sample were pooled into a 1.5 mL DNA LoBind tube to a total of 700 ng DNA in 66 µL molecular biology grade water, followed by quantification of 1 µL pooled DNA using a Qubit™ dsDNA HS Assay Kit. Adapters were then ligated onto the pooled DNA by adding 65 µL pooled DNA to 5 µL Adapter Mix II, 20 µL NEBNext Quick Ligation Reaction Buffer (5X) (NEB), 10 µL Quick T4 DNA Ligase (NEB), and incubating at RT for 10 min. AMPure XP Beads were resuspended and 50 µL was added to the reaction and incubating on a Hula mixer for 5 min at RT. The tube was then placed on a magnetic rack and, once cleared, the supernatant was discarded. The beads were then washed by adding 250 µL Short Fragment Buffer, mixed by flicking, and then placed back on the magnetic rack, after which the supernatant was removed. This process was repeated with 250 µL Long Fragment Buffer. The beads were then air-dried for 30 s, resuspended in 15 µL Elution Buffer, and incubated at 37°C for 10 min. The tube was then placed onto a magnetic rack and, once cleared, the elute was transferred to a clean 1.5 mL DNA LoBind tube. The DNA was quantified using a Qubit™ dsDNA HS Assay Kit (Section 2.1.7) and stored on ice until sequencing.

For sequencing of the DNA, an R9.4.1 Flow Cell (Oxford Nanopore Technologies) was inserted into a MinION Mk1B (Oxford Nanopore Technologies) and primed by removing any air bubbles at the priming port using a pipette and loading 800 µL priming mix (30 µL Flush Tether added to the Flush Buffer tube) into the flow cell via the priming port. The DNA library was then prepared for sequencing by adding 37.5 µL Sequencing Buffer and 25.5 µL Loading Beads to 12 µL DNA library. A further 200 µL of the priming mix was loaded into the flow cell before 75 µL of the sequencing library was added to the flow cell via the SpotON sample port in a dropwise fashion. Sequencing was performed for around 48 h until no pores remained actively sequencing.

2.4.4 Assembly of isolate genomes

Preliminary short-read only assemblies using the quality-filtered, processed short-reads from MicrobesNG were generated using SPAdes Genome Assembler v3.15.4 (Prjibelski *et al.*, 2020). For isolates that underwent long-read sequencing, basecalling was performed using Guppy Basecalling Software v6.0.1 (Oxford Nanopore Technologies) with super accuracy (`--recursive --chunk_size 3000 --chunks_per_runner 200 --min_qscore 7 --trim_barcodes --config dna_r9.4.1_450bps_sup.cfg`). Long-reads were then filtered using Filtlong v0.2.1 (Wick, 2022) (`--min_length 1000 --keep_percent 95`). The filtered long-reads were then used, along with the short-reads, for hybrid assembly using Unicycler v0.5.0 (Wick *et al.*, 2017). Assemblies were quality checked using CheckM v0.6.0 (Parks *et al.*, 2015), and ARGs were identified using ABRicate as described in Section 2.1.9.3. Assembly graphs generated by Unicycler/SPAdes were visualised using Bandage v0.8.1 (Wick *et al.*, 2015).

2.4.5 Taxonomic classification of sequenced isolates

Isolates were named numerically based on which stool sample they originated from. To taxonomically classify the isolates, the whole-genome sequences were classified with GTDB-Tk using the `classify` workflow. The *E. coli* isolate (H2-04, described in Chapter 5) was further classified using mlst v2.19.0 (Seemann, 2022) which utilises the PubMLST database (Jolley and Maiden, 2010) for multilocus sequence typing (MLST).

2.4.5.1 Phylogenetic analysis of isolates

For isolates H1-02 (*Bacteroides*) and H2-03 (*Collinsella*), further phylogenetic analysis was used to taxonomically place them within a species. All *Bacteroides* and *Collinsella* genomes from the NCBI GenBank database (Sayers *et al.*, 2022) were downloaded using NCBI Datasets (National Center for Biotechnology Information, 2022) and NCBI Genome Downloading Scripts v0.3.1 (Blin, 2022). Panaroo v1.3.0 (Tonkin-Hill *et al.*, 2020) was used to generate a core-genome alignment of H1-02 or H2-03 and all reference genomes of *Bacteroides* and *Collinsella* species, respectively. For the *Bacteroides* core-genome alignment, other sequenced *Bacteroides* isolates from Chapter 5 were also included. IQ-Tree v2.2.0-beta (Nguyen *et al.*, 2015) was used to construct phylogenetic trees using the core-genome alignments using a general time reversible-gamma model (`-m GTR+G`). All species in the GenBank database of the closest aligned reference species were also used for subsequent core-genome alignment and tree construction using the same method. Trees were visualised and annotated using iTOL v6.6 (Letunic and Bork, 2021), using the increasing nodes and mid-point rooted settings. ANIs were determined using FastANI v1.33 (C. Jain *et al.*, 2018), and a heatmap was generated using the pheatmap R package (Kolde, 2019).

2.4.6 Identification of plasmids and mobile elements in isolate genomes

Circular, complete contigs (outside of the larger chromosomal contig(s)) in the whole-genome sequence assemblies were identified as plasmids and named after characters from the Star Wars film series. The closest reference sequences of the plasmids were found using the BLASTn web interface (National Library of Medicine, 2022). Plasmid types were identified, where possible, by running ABRicate (Seemann, 2021) with the PlasmidFinder database as described in Section 2.1.9.3.

Other mobile genetic elements were found by using the ImmeDB (Intestinal microbiome mobile element database) (Jiang *et al.*, 2019) and the ICEberg 2.0 database (Liu *et al.*, 2019) with ABRicate. The databases were downloaded and converted to ABRicate databases in the same way as the ISfinder database described in Section 2.1.9.3. MGEs associated with ARGs were found by identifying overlapping regions of the genome that contained both according to the ABRicate results.

2.4.7 Conjugation assays between isolates and *Enterococcus faecium*

Conjugation assays were performed to try and transfer ARGs from two isolates to a plasmid-free strain of *E. faecium* (strain 64/3 (Bender *et al.*, 2015), Table 2.1). Five replicates of the two donor strains, *Streptococcus parasanguinis* H1-01 and *Enterocloster aldenensis* H3-04, and the recipient strain, *E. faecium* 64/3, were grown for 24 h in 2 mL mGAM broth at 37°C in an anaerobic chamber. The optical density at a wavelength of 600 nm (OD_{600}) was measured, and the cultures were diluted to equal OD_{600} values. Subsequently, in triplicate for each donor, 1 mL aliquots of donor and recipient cultures were mixed. The mixed cells were then centrifuged at 14,000 × g for 2 min and washed by discarding the supernatant, resuspending the cells in reduced PBS, and again centrifuging at 14,000 × g for 2 min. The supernatants were discarded, and cells were resuspended in 30 µL mGAM broth. The donor-recipient mixed cell suspensions were then spotted onto three sterile 0.45 µm Whatman nitrocellulose membrane filters (GE Healthcare) placed on mGAM agar plates. Controls containing donor-donor and recipient-recipient mixed cell suspensions were also spotted onto membrane filters. These plates were then incubated at 37°C for 24 h in an anaerobic chamber. After incubation, the filters were added to 30 mL universal containers containing 1 mL mGAM broth and vortexed to resuspend the cells. The resuspended cells were then serially diluted from neat to 10^{-9}, and 10 µL of each dilution was spotted onto mGAM agar plates containing different antibiotics to select for donors, recipients, and transconjugants: 8 µg/mL tetracycline or erythromycin was used to select for H1-01 and H3-04 donor cells, respectively; 25 µg/mL rifampicin (Fisher Chemical) and 25 µg/mL fusidic acid (Acros Organics) to select for recipient cells; and a combination of 25 µg/mL rifampicin and fusidic acid and the respective donor antibiotic to select for transconjugants. These plates were then incubated at 37°C for 24 h in an anaerobic chamber, after which the colonies were counted.

2.4.8 Comparison of sequenced isolates with Hi-C data

ARG-host associations by Hi-C in Chapter 3 were compared to the results from the sequenced isolates. Firstly, for each ARG in each sequenced isolate, the Hi-C ARG-host associations to the ARG in the metagenomic assembly of the faecal sample that the isolate originated from were checked to calculate the proportion of ARG-host associations to a bin with the same classification of the cultured isolate.

For further investigation, the contigs in the metagenomic assembly of the respective faecal sample that were linked to the ARG were aligned to the whole-genome sequence of the isolate using BLASTN v2.2.31 (Camacho *et al.*, 2009). ARG-linked contigs that aligned with both coverage and identity >85% were filtered from the BLAST results, and the proportion of aligned contigs was calculated, taking into account how many times the contigs were linked to an ARG using the following formula:

$$\begin{array}{c} proportion\ of\ ARG\text{-}linked \\ contigs\ aligning\ to\ WGS \end{array} = \frac{[number\ of\ Hi\text{-}C\ links\ linking\ ARG\ to\ aligned\ contigs]}{[total\ number\ of\ Hi\text{-}C\ links\ to\ ARG]}$$

Finally, for each ARG-linked contig that also aligned to the isolate genome, the classification of the bin that the contig was present in was compared to the classification of the sequenced isolate. The proportions of the contigs present in bins with identical classifications to the isolate, present in bins with classifications of the same family or order, present in a discarded bin, or contigs that were unbinned were calculated, again taking into account how many times the contigs were linked to the ARG.

2.4.8.1 Calculation of isolate genome coverage by metagenomic reads

Coverage of the isolate whole-genome sequence by Hi-C reads and shotgun metagenomic reads were calculated using CoverM v0.4.0 (Woodcroft, 2022) (`-m covered_fraction`). The coverage of the isolate genome by contigs in the metagenomic assemblies was calculated using MetaQUAST v5.0.2 (Mikheenko, Saveliev, and Gurevich, 2016). To assess correlation between coverage and Hi-C ARG-host links, R^2 values and significance of the correlations were calculated using GraphPad Prism in the same way as described in Section 2.2.6.

CHAPTER 3

METAGENOMIC CHROMOSOME CONFORMATION CAPTURE AS A TOOL TO UNCOVER LINKAGE BETWEEN ANTIBIOTIC RESISTANCE GENES AND THEIR BACTERIAL HOSTS

3.1 Introduction

Research into the gut microbiome has been greatly impacted by the technological advancements of high-throughput, low-cost sequencing in the last 20 years. This has caused a remarkable increase in studies aiming to uncover the gut microbiota's composition and function, and its roles in health and disease (Durack and Lynch, 2019). Whilst it is apparent that the gut microbiome plays a key role in human health by aiding digestion and supporting the immune system (Jandhyala *et al.*, 2015), this complex microbial community also serves as a potential reservoir for ARGs (van Schaik, 2015).

The last decade has seen a considerable rise in infections caused by MDR opportunistic pathogens originating in the human gut microbiota such as *E. coli* (Mehrad *et al.*, 2015) and *E. faecium* (Guzman Prieto *et al.*, 2016). The global spread of these MDR opportunistic pathogens and ARGs in the gut microbiota indicate the need to characterise the human gut resistome.

So far, most of the research performed on the resistome have used methods to investigate the presence and abundance of ARGs in the gut microbiota, such as qPCR (Buelow *et al.*, 2014, 2017; Sun *et al.*, 2017) and shotgun metagenomic sequencing (Hu *et al.*, 2013; Tyagi *et al.*, 2019). The taxonomic composition of a sample can also be examined using techniques such as 16S rRNA gene sequencing or shotgun metagenomic sequencing (Andersen *et al.*, 2016; Tyagi *et al.*, 2019). However, linking these data to establish which bacteria are carrying ARGs is nontrivial. Various techniques have been developed in recent years with the aim of identifying the bacterial hosts of ARGs (McInnes and McCallum *et al.*, 2020), including proximity ligation. This chapter focuses on the use of proximity ligation techniques to link ARGs to their bacterial hosts.

Several studies have performed proximity ligation techniques on gut microbiota samples from mice (Marbouty *et al.*, 2017), humans (Press *et al.*, 2017; DeMaere *et al.*, 2020; Kent *et al.*, 2020; Yaffe and Relman, 2020; Ivanova *et al.*, 2022), sheep (Bickhart *et al.*, 2022), dogs (Cuscó *et al.*, 2022), and pigs (Kalmar *et al.*, 2022). Some of these used meta3C (Marbouty *et al.*, 2017; Yaffe and Relman, 2020), whilst others performed Hi-C (Press *et al.*, 2017; DeMaere *et al.*, 2020; Kent *et al.*, 2020; Bickhart *et al.*, 2022; Cuscó *et al.*, 2022; Ivanova *et al.*, 2022; Kalmar *et al.*, 2022). The majority of these studies did not specifically attempt to use the 3C/Hi-C data to link ARGs to their hosts. Kalmar *et al.* (2022) developed a binning pipeline using Hi-C data to link ARGs to their hosts in pig faeces and were able to link several ARGs, including ARG-plasmids, to their host. Another recent study also demonstrated a novel binning pipeline with Hi-C data from human faecal samples, and were able to link ARG-plasmids to opportunistic pathogens (Ivanova *et al.*, 2022). The study by Kent *et al.* (2020) demonstrated ARG-host linkage in human faecal samples using Hi-C, albeit with a low taxonomic resolution of the linked hosts.

These studies show the potential for the use of 3C-based techniques to study the human gut resistome, however have either not focussed on ARG-host associations in the human gut or reported the hosts of ARGs at a low taxonomic resolution. Therefore, by using meta3C and a novel bioinformatic pipeline to identify reads cross-linking ARG-containing contigs to genomic contigs, the work in this chapter aimed to investigate the microbial hosts of ARGs to the genus- and species-level in a human stool sample.

3.2 Results

3.2.1 ARGs in the stool sample

To quantify the abundance of 9 ARGs in the stool sample, qPCR was performed and the relative quantity of ARGs compared to the 16S rRNA gene were calculated. To exclude targets that may have undergone nonspecific amplification, *qacA* and *bla*$_{\text{TEM}}$ were removed from further analysis for having an R^2 value less than 0.95, and *vanB*, *aph*(3'')-III, and *mecA* were removed for having an efficiency greater than 130% (see Methods Section 2.2.1.1). The relative quantity of the remaining ARGs (Figure 3.1) indicated that *tet*(W) and *tet*(Q) were highly abundant in the stool sample (an estimated 0.017 and 0.010 copies/copy 16S rRNA, respectively), whilst *erm*(B) was present at lower levels (3.6×10^{-4} copies/copy 16S rRNA). The *tolC* gene had a very low relative quantity (1.0×10^{-6} copies/copy 16S rRNA).

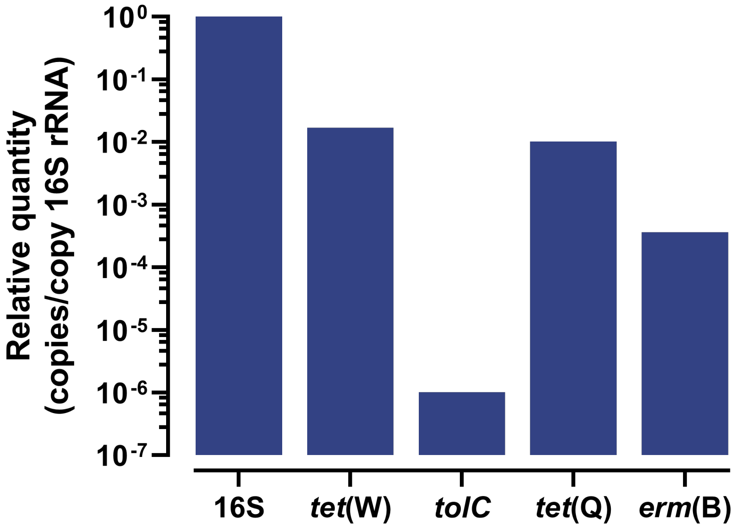


Figure 3.1. Relative quantity of antibiotic resistance genes (ARGs) in a stool sample.
Quantitative polymerase chain reaction (qPCR) was performed on DNA extracted from a stool sample to quantify the abundance of various ARGs. The 16S rRNA gene was used as a reference gene to quantify the relative abundance against. Relative quantities of the ARGs were calculated using the formula $2^{-\Delta Ct}$, where $\Delta Ct = Ct_{\text{ARG}} - Ct_{\text{16S rRNA}}$.

The 16S rRNA gene qPCR data were also used to approximate the number of cells in 1 g of the stool sample (see Methods Section 2.2.1.2), which was calculated to be approximately 5.1 × 10^{11} bacterial cells/g of stool, in line with a previous estimation that the human colon typically contains 10^{11} bacterial cells/mL content (Sender, Fuchs, and Milo, 2016).

3.2.2 Implementation of meta3C on a human stool sample

3.2.2.1 Preparation of meta3C library

The first attempts at meta3C on the human stool sample resulted in low DNA yield (<200 ng). However, after switching to methanol-free formaldehyde, the DNA yields were significantly higher (>1,500 ng) in repeated meta3C experiments. Meta3C was performed on the same stool sample using two different restriction enzymes, HpaII and MluCI, which have CCGG and AATT as their recognition sites, respectively. Using the approximate number of cells/g of stool calculated from the qPCR data, ARG-carrying *E. coli* (E3090) and *E. faecium* (E745) strains were spiked in at 0.5% each (6.4 × 10^{8} cells added each to 250 mg stool) as controls during the experimental steps in meta3C and the subsequent bioinformatic analysis of the sequence data.

Gel electrophoresis of the DNA obtained from the meta3C experiment (Figure 3.2a) showed a strong band between 5-20 kbp for the nondigested control and DNA smears were present for both digested controls (one for each restriction enzyme), with most of the DNA being between 300 and 700 bp, as expected. For both meta3C libraries, the DNA smears were larger, ranging from 20 kbp to around 400 bp, with the majority of DNA present at the higher region. Slight tapering was seen on the smears, and both the nondigested and digested controls contained a faint arrow shaped band at around 1,200 bp. These abnormalities were likely due to the presence of salts or other impurities as they were removed after further purification of the DNA was carried out on the meta3C libraries using a PCR purification kit, which resulted in DNA smears without any tapering (Figure 3.2b).

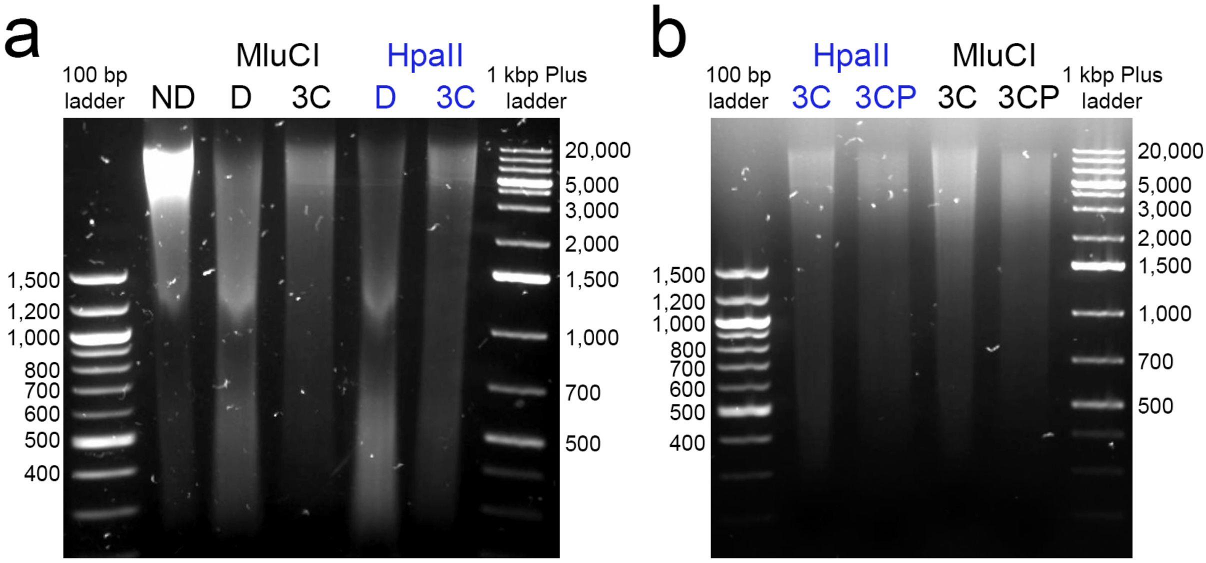


Figure 3.2. Quantification of prepared meta3C library.
DNA was quantified with gel electrophoresis. Numbers on either side of gel images represent the band size in base pairs (bp) of the DNA ladders. Meta3C was performed using two separate restriction enzymes: MluCI (black) and HpaII (blue). Lanes labelled 3C indicate final meta3C libraries. Two control samples for each library were used: digested control (D), where ligation was not performed; and nondigested control (ND), where neither digestion nor ligation were performed. **a)** shows gel electrophoresis of all samples at the end of the meta3C protocol. **b)** shows gel electrophoresis of the 3C samples before (3C) and after (3CP) DNA purification using the GeneJET PCR Purification Kit.

3.2.2.2 Preparation of sequencing library

Following the preparation of the meta3C library, the DNA was prepared for sequencing using the NEBNext® Ultra™ II FS DNA Library Prep kit. The kit involves an enzymatic fragmentation step, which needed to be optimised for the meta3C library to get DNA fragments ranging between a preferred distribution range of around 300 to 1,000 bp.

Incubation times between 2.5 and 10 min were used during the fragmentation step, and DNA fragment sizes were quantified (Figure 3.3). DNA was over-fragmented during 10 min of incubation, with sizes ranging from 150-600 bp. Fragment sizes increased with lower incubation times, and the optimal incubation time was 2.5 min, where over 70% of the DNA was within the optimal distribution range of 300-1,000 bp for both 3C libraries. The prepared libraries were then sequenced on an Illumina NextSeq 550 2x150 bp paired-end platform.

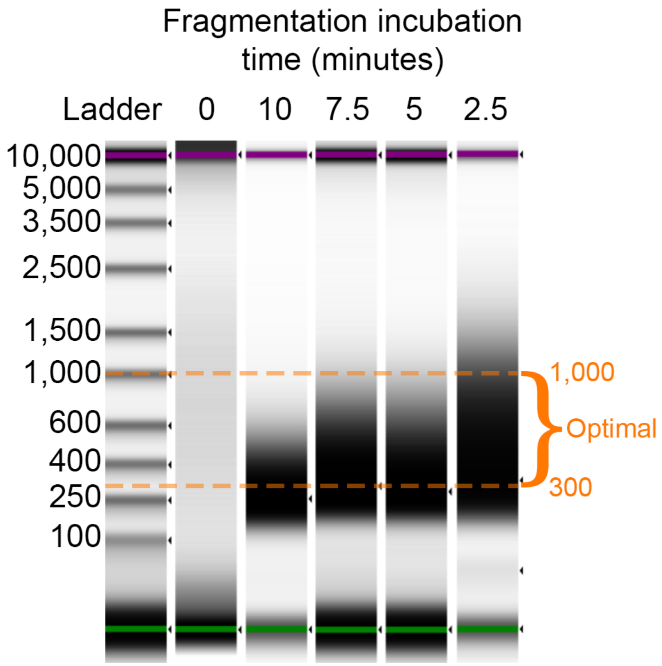


Figure 3.3. Optimising fragmentation incubation step during sequencing library preparation.
During sequencing library preparation with the NEBNext library prep kit, DNA was enzymatically fragmented for 2.5-10 minutes to find the optimal time to prepare a library with DNA fragments ranging between 300-1,000 base pairs (bp) (indicated by the orange dashed lines). DNA was quantified on an Agilent TapeStation. Numbers on the left represent the band size in bp. The lane labelled 0 was loaded with DNA that did not undergo any fragmentation, and is all >10,000 bp.

3.2.3 Analysis of meta3C sequencing data

3.2.3.1 Processing reads and metagenomic assembly

Over 223 million raw paired-end 150 nt reads were generated from the sequencing library. The HpaII meta3C library had a slightly higher proportion of reads, making up 53% of the raw reads. After processing of the reads, 101 million and 97 million high-quality reads remained for the HpaII and MluCI meta3C library, respectively. These reads were then combined and used for the metagenomic assembly. The reads were assembled into 89,005 contigs, with an N50 of 10,778 and total length of 404,824,063 bp (Table 3.2).

3.2.3.2 Published 3C data

To benchmark the meta3C data generated here against previously published 3C gut microbiota data, and to determine whether the same host-ARG links were observed in other studies, several published datasets were included in the analysis (Table 3.1).

Table 3.1. Published 3C datasets used in this study

Reference	**Method**	**Shotgun metagenomic sequence data?**	**Number of samples**	**Name in this study**
This study	meta3C	No	1	G_3C
(Marbouty *et al.*, 2017)	meta3C (mouse)	No	1	M_3C
(Press *et al.*, 2017)	Hi-C (PhaseGenomics)	Yes	1	P_HiC
(Yaffe and Relman, 2020)	meta3C	Yes	2	Y_3C
(DeMaere *et al.*, 2020)	Hi-C (PhaseGenomics)	Yes	1	D_HiC
(Kent *et al.*, 2020)	Hi-C	Yes	43	K_HiC

3C = chromosome conformation capture; meta3C = metagenomic 3C

The raw reads from these published datasets were downloaded from NCBI and processed in the same way as the reads generated in this study. Where more than one enzyme was used and sequenced as separate libraries, or if meta3C/Hi-C datasets were made up of technical repeats, the reads were combined before metagenomic assembly. Some datasets also contained shotgun metagenomic reads from the stool samples (Press *et al.*, 2017; DeMaere *et al.*, 2020; Kent *et al.*, 2020; Yaffe and Relman, 2020). These metagenomic reads were processed and assembled separately to the meta3C/Hi-C reads of the same sample (Table 3.2). The assembly qualities varied across the datasets, including varying qualities between the 3C/Hi-C assemblies and the assemblies generated from the accompanying shotgun metagenomic sequence data for the same sample. This was most apparent in the P_HiC dataset, where the N50 of the shotgun assembly (P_SG) was over four times that of the assembly generated from the Hi-C data and the total length of the P_SG assembly was considerably higher. The opposite was true for the D_HiC dataset, likely due to the differences in the depth of the sequencing (Table 3.2). For Hi-C datasets, due to the additional enrichment step causing artifacts in the sequence data, the assemblies from the accompanying shotgun metagenomic library should be used for further analysis of ARG-host links. For Y_3C_A, although shotgun metagenomic sequencing was performed on the sample, the assembly generated from the meta3C reads was considerably larger with a total length of 1,040,533,919 bp compared to 658,813,968 bp from the shotgun read assembly (Y_SG_A) (Table 3.2). This was again due to the differences in the depth of sequencing, as the meta3C library for Y_3C_A contained over 3 billion raw reads compared to 416 million raw reads from the Y_SG_A library, and the same was true for Y_3C_B vs Y_SG_B. As Y_3C used meta3C rather than Hi-C, the larger assemblies generated from the meta3C library were used during downstream analysis for both samples.

Table 3.2. Read counts and assembly statistics

Dataset	Assembly	Read length (bp)	Raw reads	Processed reads	Total length of assembly (bp)	No. of contigs	N50
G_3C	G_3C	2 × 150	223,169,682	198,493,086	404,824,063	89,005	10,778
M_3C	M_3C	2 × 75	375,815,400	366,961,002	480,933,195	116,057	7,562
P_HiC	P_HiC	2 × 150	171,853,886	157,755,162	201,471,765	75,274	3,279
	P_SG	2 × 150	250,884,672	237,293,522	528,999,126	104,368	14,455
Y_3C	Y_3C_A	2 × 160	3,019,738,680	2,921,579,828	1,040,533,919	177,689	17,376
	Y_SG_A	2 × 160	416,571,650	410,280,634	658,813,968	101,575	22,551
	Y_3C_B	2 × 160	682,773,219	1,239,950,680	866,666,497	146,989	18,851
	Y_SG_B	2 × 160	202,617,904	198,775,312	484,955,068	83,017	19,100
D_HiC	D_HiC	2 × 80	143,286,468	133,509,800	169,028,509	30,519	13,667
	D_SG	2 × 150	20,088,550	18,925,950	131,298,239	37,723	5,924
K_HiC (average*)	K_HiC	2 × 150	41,021,508	37,984,239	69,211,602	16,645	16,582
	K_SG	2 × 150	90,510,991	83,397,427	156,853,335	32,197	18,026

*for K_HiC, an average of 43 samples is presented in this table; x_SG = accompanying shotgun metagenomic reads; assemblies highlighted in yellow were used during analysis of 3C/Hi-C data; bp = base pairs

3.2.3.3 Taxonomic profiles and ARGs present

The compositions of the processed reads from all datasets were profiled using MetaPhlAn3. On average, 57.4±15.1% (± standard deviation) and 67.6±15.7% of the taxonomic profiles from shotgun metagenomic and 3C/Hi-C data, respectively, were not identified by the MetaPhlAn database (classified as UNKNOWN). Of the reads that were classified, most samples showed results that can be expected for a human faecal sample, with the majority of identified reads being assigned to the Clostridia and Bacteroidia classes (Figure 3.4). Some samples differed greatly from the others, such as K_HiC_N1-4, where 88.55% of the identified reads were assigned to 'Viruses_unclassified' (Figure 3.4).

For the dataset generated for this PhD thesis (G_3C), 53.05% of reads were unclassified, and of the classified reads, 54.30% were assigned to the Firmicutes phylum, 28.27% to Actinobacteria, 9.94% to Bacteroidetes, and 7.36% to Proteobacteria. *Bifidobacterium* and *Collinsella* were the most abundant known genera with a relative abundance of 7.90% and 4.59%, respectively (Figure 3.5). The *Enterococcus* and *Escherichia* genera had similar abundances to each other (3.79% and 3.43%, respectively), which suggests that the *E. coli* and *E. faecium* strains had been spiked in at a higher level than the 0.5% target.

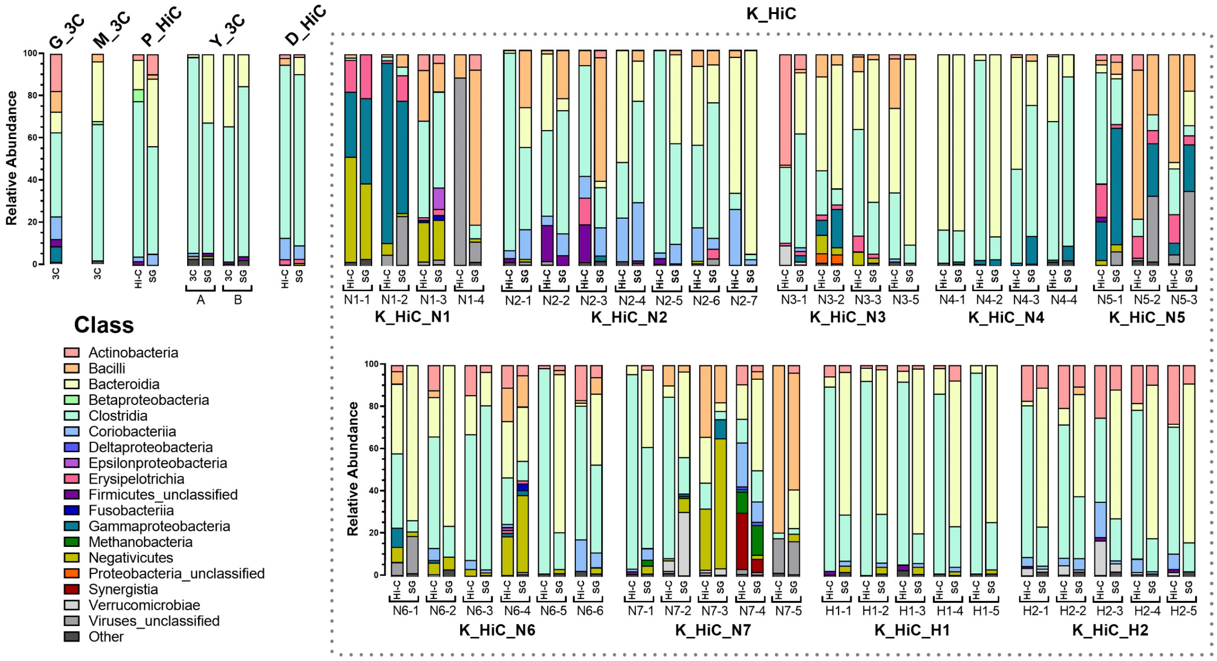


Figure 3.4. Class-level compositions of all datasets.
The reads from all datasets were taxonomically profiled using MetaPhlAn3. The stacked bars show the relative abundance (%) of each class for the classified reads. Reads that could not be classified by MetaPhlAn3 (~60% of reads for each dataset) are excluded here.

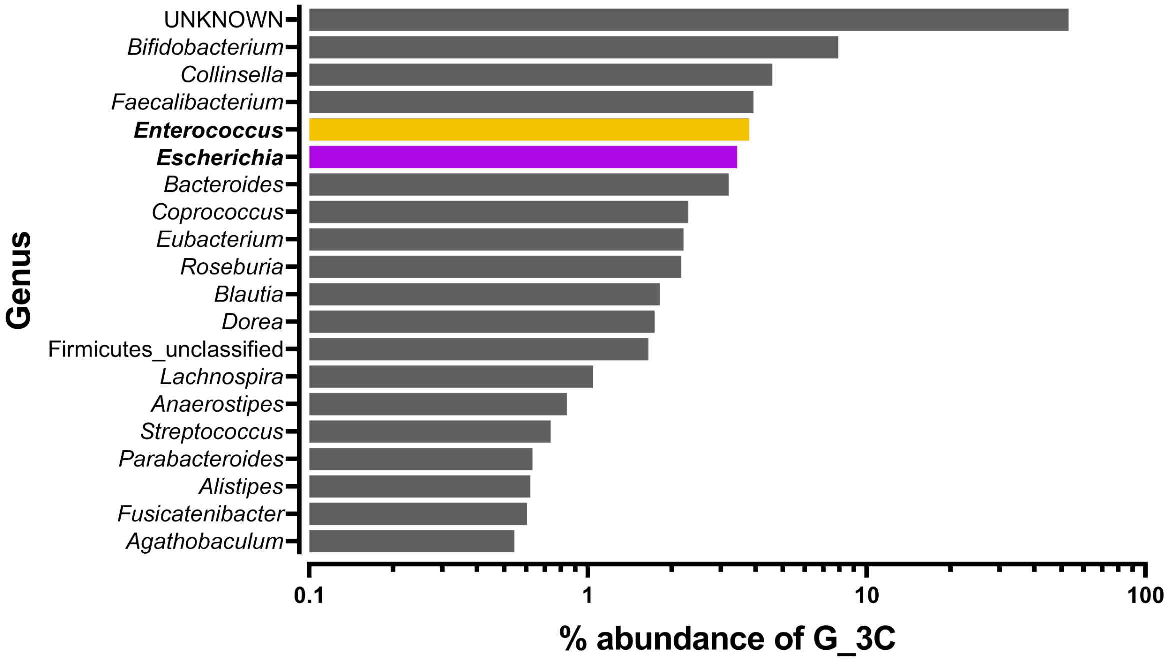


Figure 3.5. Top 20 most abundant genera in meta3C sequence data G_3C. Following sequencing of the G_3C meta3C library, the reads were taxonomically profiled using MetaPhlAn3. Bars show the relative abundance (%) of each genus (log(10) scale). 53.05% of the reads were not mapped to any known taxon (UNKNOWN). *Enterococcus faecium* and *Escherichia coli* were spiked into the sample prior to meta3C library preparation, and are highlighted in yellow and purple, respectively.

After profiling the composition of the reads, ABRicate was used to identify contigs containing ARGs in the metagenomic assemblies. All samples contained ARGs, however, for the datasets that contained shotgun metagenomic data, there were discrepancies between the ARGs present in the assemblies generated from the Hi-C reads and shotgun reads of the same sample. The abundance (RPKM) of the ARGs in all the datasets were calculated (Figure 3.6, Figure A3.1 for K_HiC dataset).

In the G_3C assembly, 37 contigs containing ARGs were identified. In line with the qPCR results (Figure 3.1), *tet*(W) had a high relative abundance of 53.5 RPKM, 10-fold higher than that of *erm*(B) (4.0 RPKM). The *tolC* gene was not present in the metagenomic assembly, reflecting its low abundance as determined by qPCR.

Although the qPCR data suggested that *tet*(Q) was present with a high relative abundance, the *tet*(Q) gene in the metagenomic assembly did not meet the ARG inclusion criteria (identity was <95% to the reference database) so was excluded from further analysis.

The known ARGs from the *E. coli* E3090 and *E. faecium* E745 spike-ins were all present (Figure 3.6, highlighted). For E745, the two chromosomal ARGs (*aac(6')-Ii* and *msr*(C)) had similar abundances of 15.0 and 14.3 RPKM, respectively. The other ARGs from E745, *vanHAX* and *dfrG*, are carried in plasmids, and had higher abundance (43.4 and 34.9 RPKM) than the chromosomal ARGs, likely due to being carried in a plasmid that has a higher copy number than the chromosome. For the E3090 ARGs, six chromosomal ARGs (*sul1*, *sul2*, *ant(3")-Ia*, *bla*$_{\text{OXA-1}}$, *floR*, and *mdf*(A)) had relatively similar abundances, ranging from 9.6-27.4 RPKM. The ARGs carried in plasmids in E3090 had higher relative abundances, such as *mcr-1.1* which had an abundance of 98.1 RPKM, nearly 4 times higher than the chromosomal ARGs. Similarly, *bla*$_{\text{TEM}}$ was present at a high abundance of 104.8 RPKM. The *bla*$_{\text{TEM}}$ gene present in the metagenomic assembly was *bla*$_{\text{TEM-116}}$, as opposed to *bla*$_{\text{TEM-1B}}$ that the E3090 genome contains, however these genes differ by only 5 single nucleotide polymorphisms (SNPs), so this is most likely due to a misassembly.

The rest of the datasets contained many and diverse ARGs, with 71 unique ARGs in total across the datasets, excluding the K_HiC samples (Figure 3.6). The 86 samples in the K_HiC dataset (43 Hi-C, and 43 corresponding shotgun metagenomic samples) alone contained 141 unique ARGs and have been shown separately in Figure A3.1.

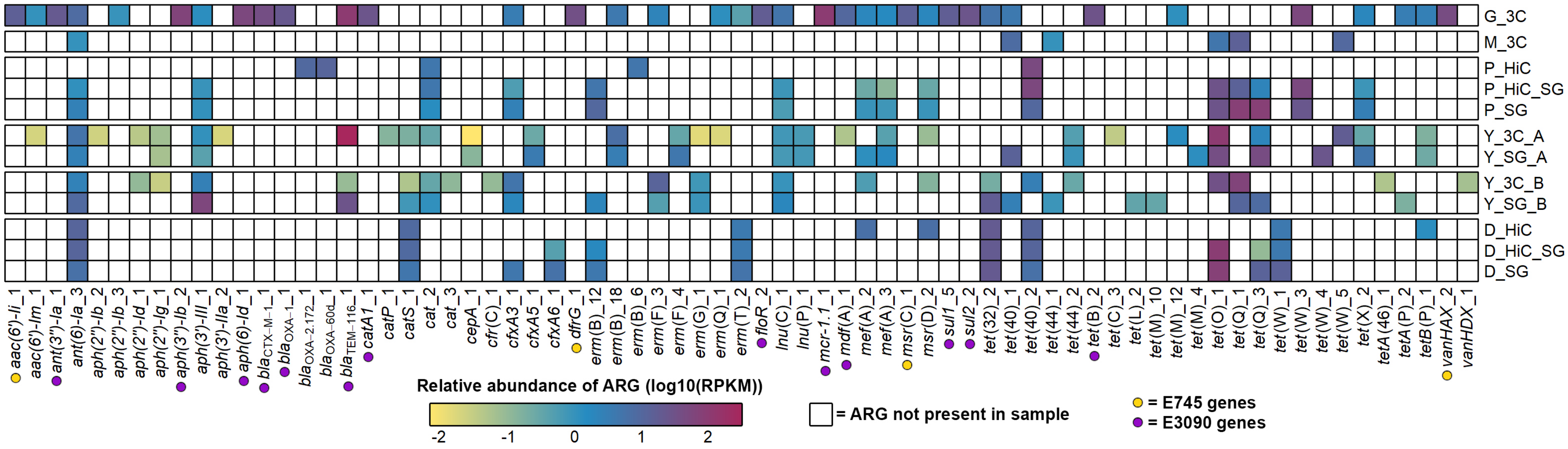


Figure 3.6. Relative abundance of antimicrobial resistance genes (ARGs) in 3C/Hi-C datasets.
The ARG sequences from the assemblies of each dataset were isolated, and the reads from that dataset were mapped to the ARGs (columns). The relative abundance was calculated as reads per kilobase per million mapped reads (RPKM). White cells mean the ARG was not present, and coloured cells show that the ARG was present, with the colour relating to the relative abundance of the ARG within that set of reads (log(10) transformed RPKM values). Each row represents a different 3C/Hi-C set of reads (*_HiC/*_3C), or the shotgun metagenomic reads from the datasets that contained shotgun data (*_SG). Rows labelled *_HiC_SG show the RPKM of the Hi-C reads mapping to the ARGs identified in the shotgun assembly. The ARGs highlighted with a coloured dot are ARGs from the spike-ins in the G_3C dataset (purple = *E. coli* E3090, yellow = *E. faecium* E745). ARGs in the K_HiC datasets are shown as a separate heatmap (Figure A3.1).

3.2.3.4 Identifying cross-linked reads

To identify reads originating from cross-linked fragments of DNA, the first 50 bp of the 3C/Hi-C reads from each dataset were first mapped against their respective metagenomic assemblies. For Hi-C datasets (P_HiC, D_HiC, K_HiC), the Hi-C reads were mapped to assemblies generated from the accompanying shotgun metagenomic library, whereas 3C reads from the 3C datasets (G_3C, M_3C, Y_3C) were mapped to assemblies generated directly from the 3C library. From the reads that mapped with a MAPQ >20, "intercontig" read pairs were identified where both reads of the pair mapped to different contigs (Table 3.3), meaning the read pair potentially originated from a cross-linked fragment of DNA.

The proportion of intercontig reads greatly varied across the datasets, with the highest being 13.74% for P_HiC, and the lowest being 0.2% for K_HiC_N1-1 (0.64% average across all K_HiC samples). For the meta3C datasets, M_3C had the highest proportion at 9.73%. The G_3C dataset had the lowest of the meta3C datasets at 1.65% (Table 3.3). For G_3C, the MluCI reads had a higher proportion of intercontig reads compared to the HpaII reads (2.12% compared to 1.20%, Table 3.4).

Table 3.3. Read counts during meta3C/Hi-C analysis

Dataset	G_3C	M_3C	Y_3C		D_HiC	P_HiC	K_HiC*
			Y_3C_A	Y_3C_B			
Processed reads	198,493,086	366,961,002	2,921,579,828	1,239,950,680	133,509,800	157,755,162	37,984,239
Reads mapped (MAPQ>20)	181,467,148	278,726,053	2,868,601,794	1,155,939,113	108,556,752	124,877,330	32,412,518
Percentage mapped	91.42%	75.96%	98.19%	93.22%	81.31%	79.16%	85.66%
Intercontig reads	3,271,007	35,717,451	188,322,547	94,104,831	9,488,683	21,679,019	192,510
Percentage intercontig	1.65%	9.73%	6.45%	7.59%	7.11%	13.74%	0.64%

*for K_HiC, an average of 43 samples is presented in this table; For datasets that used multiple restriction enzymes, numbers presented are a combined total; MAPQ = mapping quality

Table 3.4. Read counts for each read set in G_3C

Meta3C restriction enzyme used	HpaII	MluCI
Processed reads	101,286,272	97,206,814
Reads mapped (MAPQ>20)	92,405,460	89,061,688
Percentage mapped	91.23%	91.62%
Intercontig reads	1,210,559	2,060,448
Percentage intercontig	1.20%	2.12%

MAPQ = mapping quality

3.2.3.5 Investigation of intercontig reads

Due to the large differences in the proportion of intercontig reads across the datasets, these reads were investigated further to examine whether intercontig reads were truly a result of physical cross-linking. First, shotgun metagenomic reads, which, by definition, cannot have been physically cross-linked, were mapped back to the assemblies in the same way as the 3C/Hi-C reads were in the previous step for the datasets that contained shotgun reads (Y_3C_A/B, D_HiC, P_HiC, K_HiC). They were then analysed in the same way as the 3C/Hi-C reads to isolate the intercontig read pairs and calculate the proportion of intercontig reads. The shotgun metagenomic reads showed a background level of 0.16-4.20% intercontig reads (Figure 3.7), which shall be referred to as 'spurious intercontig reads', i.e. intercontig reads that are not originating from physically cross-linked fragments of DNA. In the K_HiC datasets, the average proportion of intercontig reads from the shotgun metagenomic reads was 0.74%, even higher than the 0.64% average from the Hi-C reads. This suggested that there may be no, or very few, reads resulting from the physical cross-linking of DNA in the K_HiC dataset. As the G_3C dataset had an intercontig read proportion of 1.65%, within the range of the spurious intercontig reads from the shotgun metagenomic data, it also may not have been sufficiently cross-linked during the experimental procedure. However, because the G_3C dataset contained the E3090 and E745 spike-ins, for which whole-genome sequences are available, the intercontig reads mapping to those could be examined further. G_3C reads were mapped to the E3090 and E745 genomes to identify the isolate spike-in 3C reads for each genome. These reads were then compared to WGS reads downloaded from NCBI for each genome by mapping the first 50 bp of all reads to the G_3C assembly (Table 3.5).

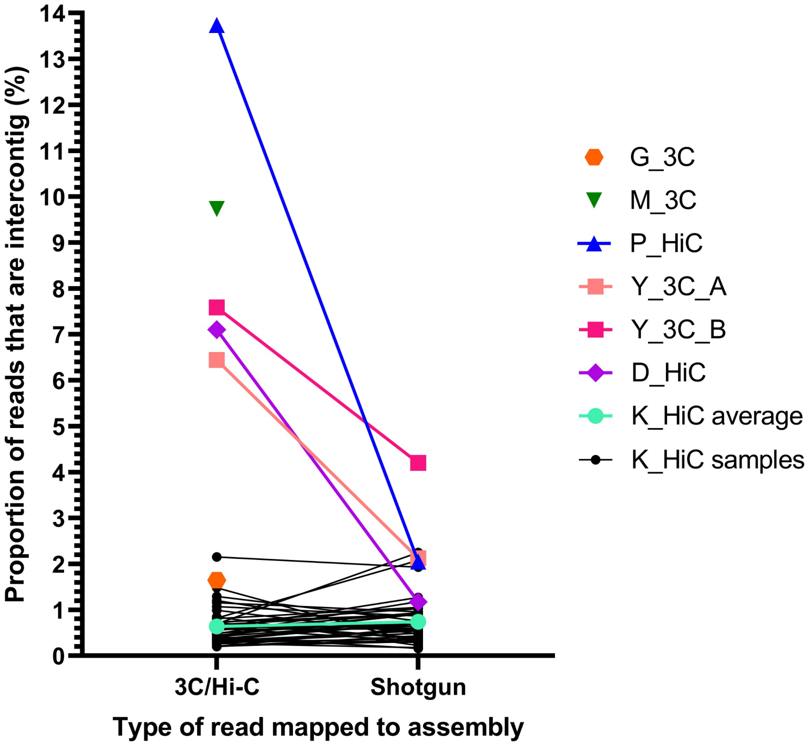


Figure 3.7. Proportion of intercontig reads in 3C/Hi-C and shotgun reads of the same sample.

The first 50 bp of each read was mapped against the corresponding assembly, and pairs where each read of the pair mapped to different contigs were labelled as intercontig reads. Y-axis shows the percentage of reads that were intercontig. K_HiC average (cyan) is the average for all 43 K_HiC samples (black). G_3C (orange) and M_3C (green) did not have accompanying shotgun reads, so only the intercontig proportion for the 3C reads are shown.

Table 3.5. Comparison of spike-in WGS reads and G_3C that map to spike-in genomes

Spike-in	**E3090**		**E745**	
Dataset	**G_3C**	**WGS**	**G_3C**	**WGS**
Total reads	14,497,782	1,284,538	8,170,430	3,333,334
Reads mapped to G_3C assembly (MAPQ>20)	14,237,338	1,256,272	7,764,542	3,175,581
Percentage mapped to G_3C assembly	98.20%	97.80%	95.03%	95.27%
Intercontig reads	141,473	10,080	81,191	78,106
Percentage intercontig	0.98%	0.78%	0.99%	2.34%
Average insert size of intercontig reads (nt)	989,321	116,193	147,107	221,937
Average insert size of non-intercontig reads (nt)	291	407	259	461

WGS = whole genome sequencing; MAPQ = mapping quality; nt = nucleotides

The proportion of intercontig reads from the 3C reads were comparable to the WGS reads, confirming again that short read sequencing produces a considerable background level of spurious intercontig reads. Aligning both the intercontig and non-intercontig reads from the G_3C spike-in and the WGS reads back to their respective genomes revealed which regions they were mapping to in the genome, as well as the insert size between the paired reads. As expected, the insert sizes of the G_3C spike-in intercontig reads were much greater than the non-intercontig reads, with an average of 989,321 nt and 147,107 nt for the for the G_3C E3090 and G_3C E745 intercontig reads, respectively, compared to 291 nt and 259 nt for the respective non-intercontig reads. However, the same was seen for the spurious intercontig reads from the WGS data of both spike-ins (Table 3.5).

Examining where the reads map to in the spike-in genomes, both the intercontig and non-intercontig reads spanned the whole genome for both spike-ins and aligned to a wide range of regions (Figure 3.8). A greater proportion of the intercontig reads mapped to IS elements in the genome compared to the non-intercontig reads for all sets of reads except for the G_3C E3090 reads. This was most clear in the E745 reads, where over 20% of the intercontig reads for both the G_3C and WGS reads aligned to IS elements, compared to less than 1% of the non-intercontig reads (Figure 3.8), suggesting that the presence of IS elements in the assembly is responsible for some spurious intercontig reads. Using ABRicate with the IS finder database, 93 IS elements (18 unique) were found across the E745 genome (3,168,411 bp), compared to 79 IS elements (25 unique) in the E3090 genome (5,270,976 bp). When controlling for genome size, E745 has almost double the amount of IS elements per megabase pair (Mbp), with 29.3 and 15.0 IS elements/Mbp for E745 and E3090, respectively.

Therefore, the E745 reads have a higher chance of mapping to an IS element, causing more spurious intercontig reads, which may explain the dissimilarity between the E745 and E3090 spike-ins here.

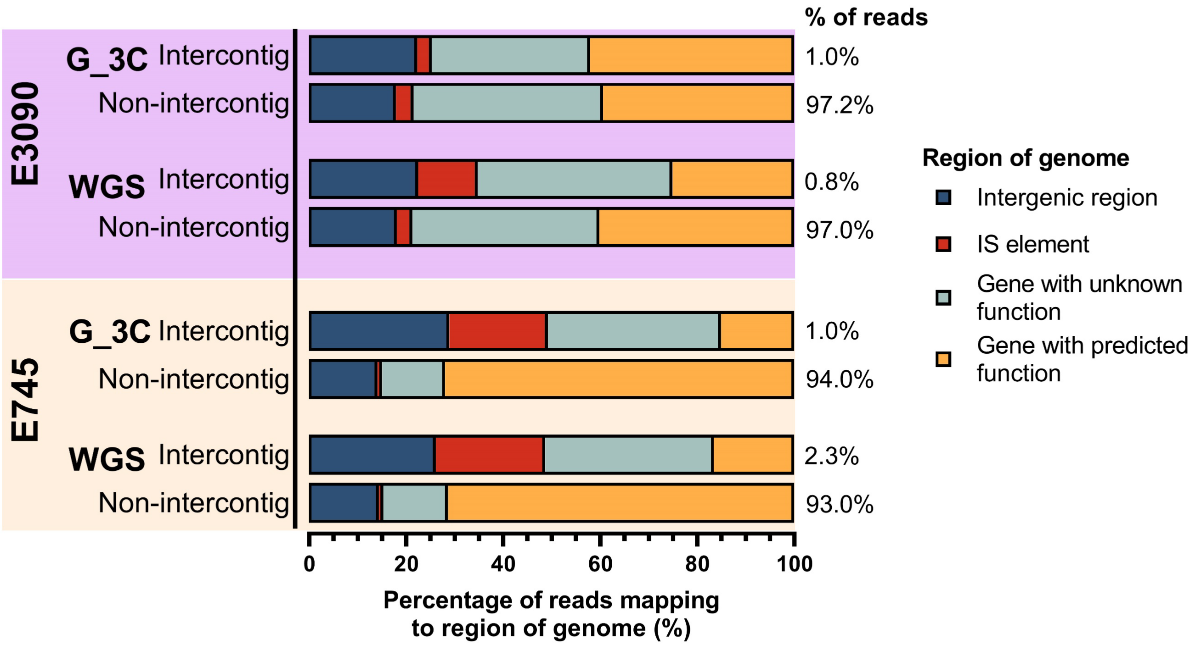


Figure 3.8. Proportions of where G_3C and whole genome sequencing (WGS) reads map to in their respective spike-in genomes.
Both the intercontig and non-intercontig reads for G_3C spike-in reads and WGS reads of the spike-ins were mapped to their respective genomes. The genomes were annotated using Prokka and the regions in which the reads mapped to were grouped into four categories (see legend). IS element = insertion sequence element. Percentages at the end of the stacked charts show the proportion of total reads that were assigned as intercontig/non-intercontig.

The wide range of mapping seen in both the G_3C intercontig reads and the spurious WGS intercontig reads, as well as the large insert sizes, suggested that there are no clear ways of identifying spurious intercontig reads by examining what and where they map to. However, the considerably greater proportion of intercontig reads aligning to IS element regions in the spike-in genomes compared to the non-intercontig reads indicated that intercontig reads should be filtered to remove reads that align to a contig containing an IS element or other repetitive elements.

Next, the position in the contigs from the G_3C assembly that the spike-in reads mapped to was checked to determine whether spurious intercontig reads were more likely to map near to the beginning or end of a contig, meaning they were potentially caused by fragmentation in the assembly. Indeed, a greater proportion of the intercontig reads for both the G_3C and WGS spike-in reads mapped within 500 nt of the ends of a contig compared to the non-intercontig reads (Figure 3.9). For G_3C E3090 reads, 37.7% of the intercontig reads mapped within the first or last 500 nt of a contig, compared to only 5.0% of non-intercontig G_3C E3090 reads. This observation was even clearer for the E3090 WGS and G_3C/WGS E745 reads, where over 80% of the intercontig reads were mapping near the ends of a contig, compared to less than 10% of the non-intercontig reads (Figure 3.9). These results may suggest that meta3C worked better for the E3090 spike-in than the E745 spike-in.

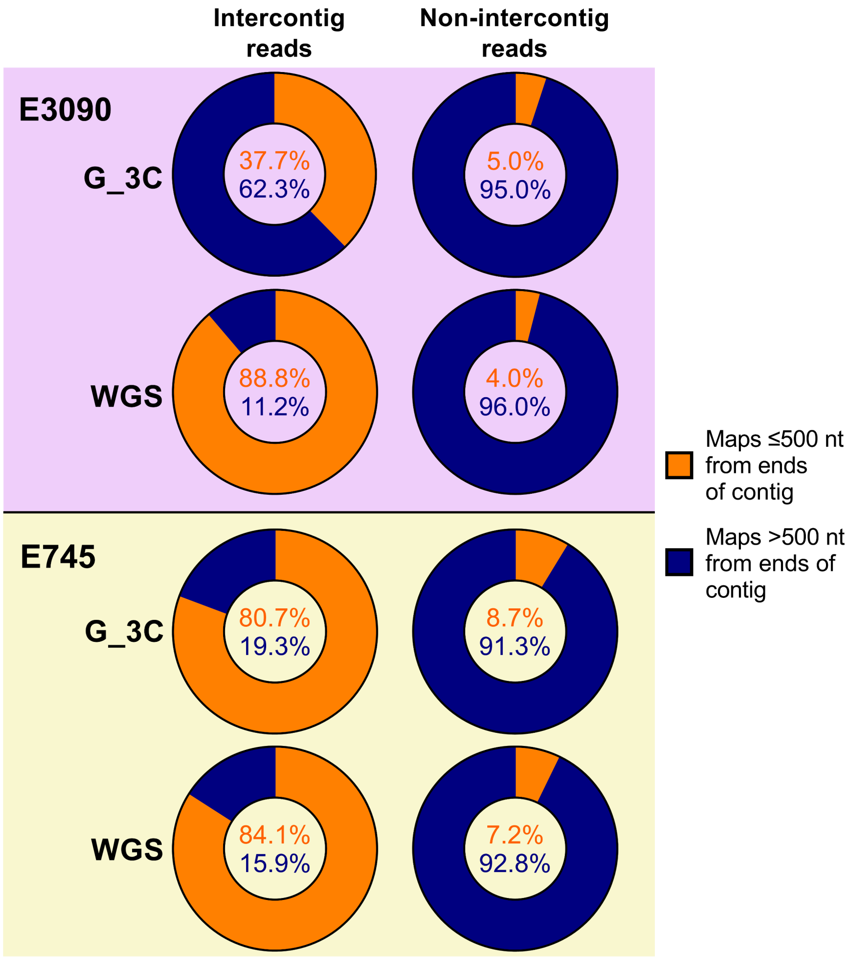


Figure 3.9. Proportions of reads mapping within the first or last 500 nucleotides (nt) of a contig in the G_3C assembly for spike-in G_3C and whole genome sequencing (WGS) reads.

The position of the alignment to contigs in the G_3C assembly was checked for both intercontig and non-intercontig read pairs from WGS reads and reads from G_3C that mapped to each spike-in genome. Orange shows the proportion of reads mapping within 500 nt of the ends of a contig. Blue shows the proportion of reads mapping more than 500 nt away from the ends of a contig.

To determine whether intercontig reads mapped to the ends of contigs for all 3C/Hi-C reads, the positions in the metagenomic assembly that the reads mapped to were checked for all datasets. The majority of spurious intercontig reads from the shotgun metagenomic data mapped within the first or last 500 nt of a contig for all datasets that had shotgun data (Figure 3.10). For the 3C/Hi-C intercontig reads, the proportion varied, but was lower for P_HiC, D_HiC, Y_3C_B, and M_3C (12.7%, 27.5%, 32.1%, and 19.4%, respectively), compared to around 52% for G_3C and Y_3C_A. For G_3C, 51.9% of all intercontig reads mapped near the ends of a contig. However only 41.7% of the MluCI G_3C reads alone mapped near the ends of a contig, compared to 69.2% of the HpaII G_3C reads, again indicating that meta3C may have worked better in the MluCI library for G_3C.

Interestingly, the proportion of intercontig 3C/Hi-C reads mapping near ends of a contig correlated (R^2 = 0.86; P = 0.0028) with the proportion of intercontig reads in the dataset, whereby datasets with a higher proportion of intercontig reads in the sequence data had a lower proportion of intercontig reads mapping near the ends of a contig (Figure 3.11). This, along with the high proportion of shotgun intercontig reads mapping near the ends of a contig, suggested that many spurious intercontig reads could be filtered out by removing those which map within the first or last 500 nt of a contig.

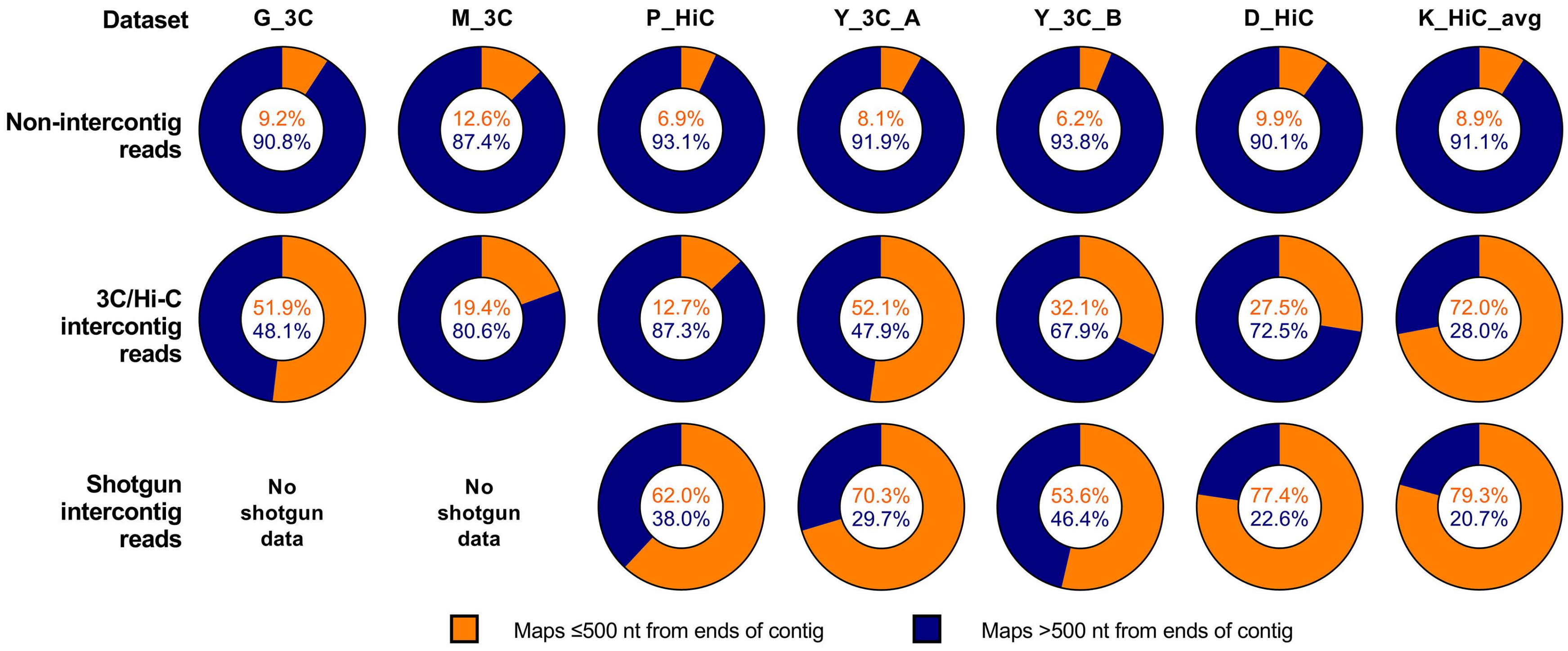


Figure 3.10. Proportions of intercontig reads mapping within the first or last 500 nucleotides (nt) of a contig in their respective assemblies for all datasets.
The position of the alignment to contigs was checked for the intercontig reads in all datasets. Orange shows the proportion of reads mapping within 500 nt of the ends of a contig. Blue shows the proportion of reads mapping greater than 500 nt away from the ends of a contig.

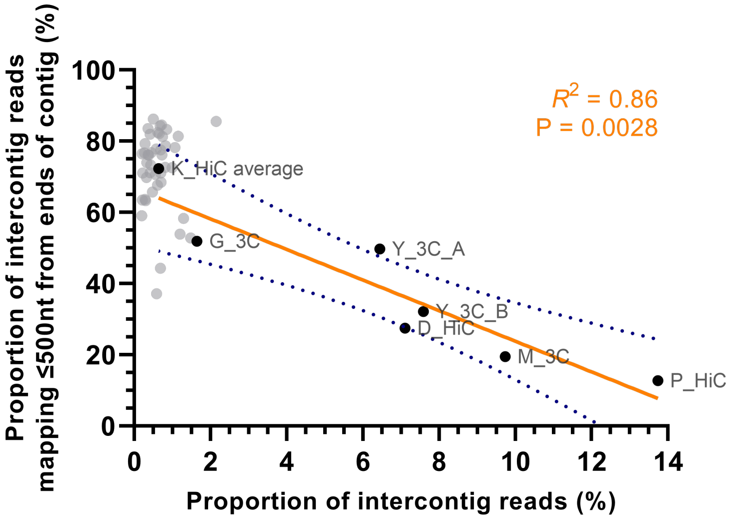


Figure 3.11. The proportion of identified intercontig reads vs the proportion of intercontig reads mapping within the first or last 500 nucleotides (nt) of a contig for all 3C/Hi C datasets.
Each point represents a different 3C/Hi-C sample (labelled). Unlabelled grey points are the individual samples in the K_HiC dataset. Slope calculated via linear regression analysis showing a statistically significant correlation (P = 0.0028, Spearman correlation). Blue dotted lines indicate the 95% confidence interval.

3.2.3.6 Filtering of spurious intercontig reads

To reduce the number of spurious intercontig reads in the data, intercontig reads that mapped within the first 500 nt of a contig were removed in all datasets. This reduced the proportion of intercontig reads across all the datasets (Figure 3.12). The proportion of intercontig 3C/Hi-C reads decreased, on average, by 32.6% when filtered. However, filtering had a greater effect on the spurious intercontig reads identified in the shotgun data, which decreased by 63.8%, suggesting that this step is essential to reduce the number of spurious intercontig reads in the data.

After removing the reads mapping near the ends of contigs, the proportion of intercontig reads from the Hi-C data in the K_HiC dataset was as low as 0.18% on average (Table 3.6), hardly different from the average of the K_SG intercontig reads (0.16%). Therefore, this dataset was removed from further analysis.

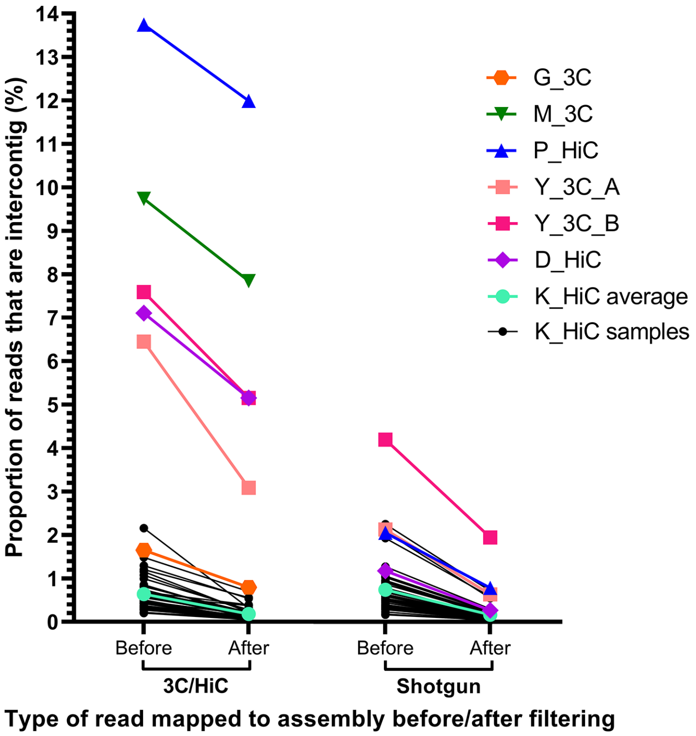


Figure 3.12. Proportion of intercontig reads in 3C/Hi-C and shotgun reads before and after filtering.
The first 50 bp of each read was mapped against the corresponding assembly, and pairs where each read of the pair mapped to different contigs were labelled as intercontig reads ('Before' on X-axis). These were then filtered to remove intercontig reads that mapped within the first or last 500 nt of a contig ('After' on X-axis). Y-axis shows the percentage of reads that were intercontig. K_HiC average (cyan) is the average for all 43 K_HiC samples (black). G_3C (orange) and M_3C (green) did not have accompanying shotgun reads, so only the intercontig proportion for the 3C reads are shown.

3.2.3.7 Linking ARGs to their microbial hosts

After filtering the intercontig reads, pairs where one read mapped to an ARG-contig in its respective assembly were identified. To reduce the impact of the noise from any remaining spurious intercontig reads, contigs were only considered "linked" to ARG-contigs if there was at least five unique intercontig read pairs linking them. On top of this, ARG-linked contigs identified as IS elements were also filtered out.

For G_3C, this resulted 26,607 intercontig reads that linked a total of 466 contigs to 27 out of 37 of the ARG-contigs at least five times (Table 3.6). Of the 466 contigs linked to ARGs, 48 of these came from the HpaII meta3C library, and 418 came from the MluCI meta3C library. Linked contigs that mapped with >99% identity to known plasmid sequences in the NCBI nt database, which were all linking to ARGs from the spike-ins, were removed (Table 3.6), and the remaining contigs were then taxonomically classified using Kraken2. This revealed that the ARGs were linked to a wide range of taxa (Figure 3.13). Genes from the E745 spike-in were correctly linked to *Enterococcus faecium*, although *vanHAX* was excluded as it only linked to plasmid contigs. The same was true for *catA1* and *bla*$_{\text{TEM}}$ in the *E. coli* E3090 spike-in, however the remaining E3090 ARGs were all linked to *Escherichia coli*. A small proportion (1.7-3.7%) of the contigs that linked to several of the E3090 ARGs (*bla*$_{\text{CTX-M-1}}$, *mcr-1.1*, *aph(3")-Ib*, *aph(6)-Id*, *mdf*(A), *sul1*, *ant(3")-Ia*, *bla*$_{\text{OXA-1}}$, and *sul2*) were only classified to the family-level as Enterobacteriaceae, with the remaining contigs linked to these genes being successfully classified to species-level as *E. coli*.

These results indicated that the analysis pipeline used here could successfully link the spike-in ARGs to their correct host. As Kraken2 can sometimes misclassify a relatively large proportion of reads incorrectly at the species-level (Wood, Lu, and Langmead, 2019), hosts were assigned only to genus-level (Figure 3.13). The non-spike-in ARGs linked to a wide range of hosts. Some ARGs such as *cfxA3* and *tet*(X) linked to single hosts, whereas others like *tet*(40) and *tet*(W) were widespread and linked to various gut commensals. Where ARGs were associated with multiple taxa, the potential microbial hosts were usually related at phylum-level, such as *tet*(40) which linked to the genera *Streptococcus*, *Flavonifractor*, and *Lachnoclostridium*, which are all in the Firmicutes phylum.

Table 3.6. Number of contigs linking to ARG-contigs in 3C/Hi-C datasets

Dataset	G_3C	M_3C	Y_3C		D_HiC	P_HiC	K_HiC*
			Y_3C_A	Y_3C_B			
Intercontig reads	3,271,007	35,717,451	188,322,547	94,104,831	9,488,683	21,679,019	192,510
Percentage intercontig	1.65%	9.73%	6.45%	7.59%	7.11%	13.74%	0.64%
Filtered intercontig reads	1,574,468	28,773,234	90,197,910	63,855,164	6,880,255	18,917,493	53,157
Percentage intercontig after filtering	0.79%	7.84%	3.09%	5.15%	5.15%	11.99%	0.18%
Reads linking contigs to an ARG-contig	26,607	17,321	186,177	128,774	19,475	28,200	-
Unique contigs linked to ARG-contig	4,767	6,763	9,172	14,757	3,007	4,517	-
Linked ≥5 times & no IS element	466	264	606	1,655	392	443	-
Removed plasmid contigs	342	259	594	1,627	387	439	-
Number of ARGs linked to host(s) (/ number of ARGs in sample)	27 / 37	6 / 7	23 / 30	16 / 23	6 / 11	9 / 15	-

*for K_HiC, an average of 43 samples is presented in this table; For datasets that used multiple restriction enzymes, numbers presented are a combined total; ARG = antimicrobial resistance gene; IS element = insertion sequence element

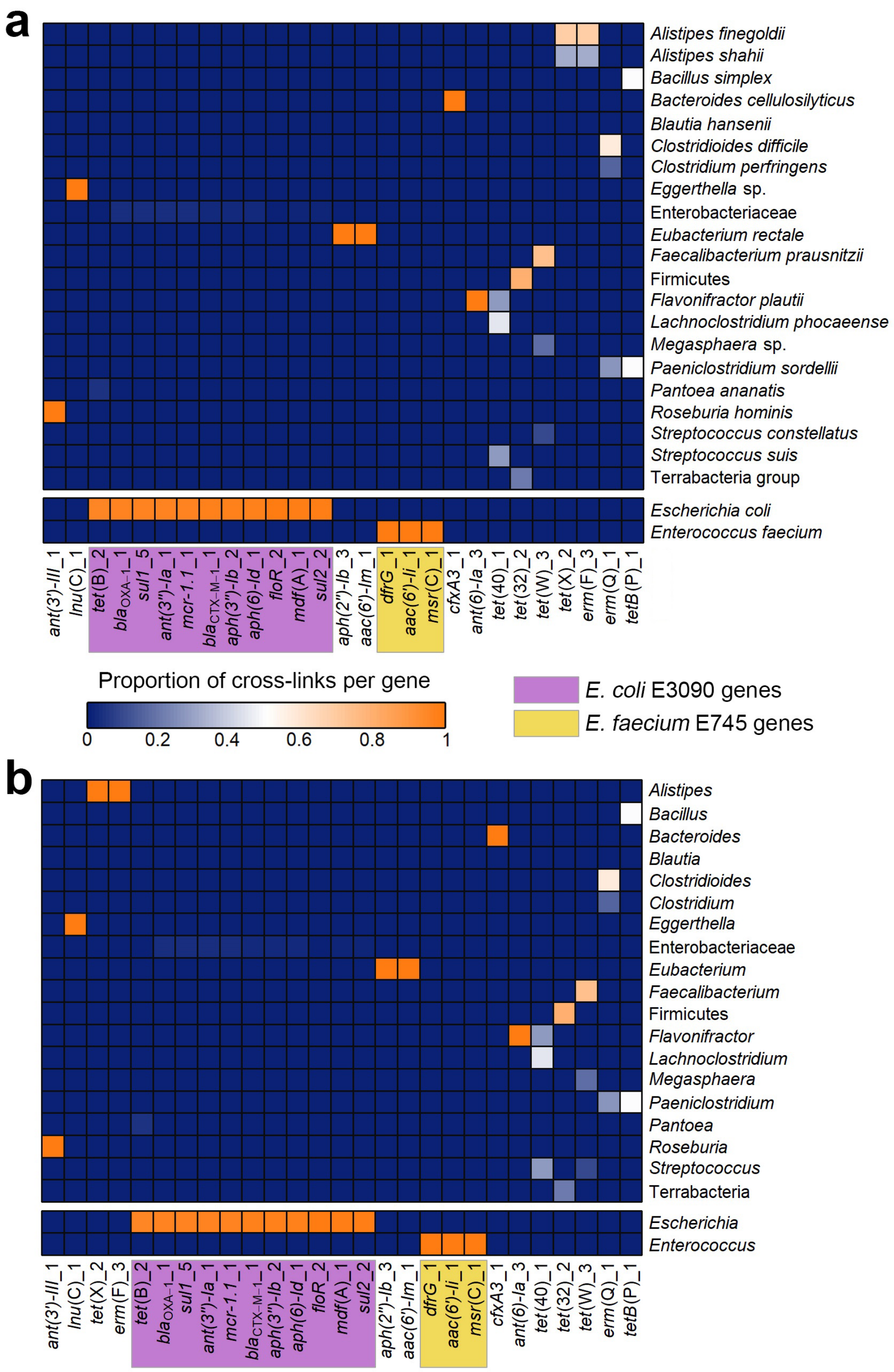


Figure 3.13. Heatmap showing antimicrobial resistance genes (ARGs) linked with their microbial hosts for G_3C.
Contigs linked to ARG-containing contigs were taxonomically classified using Kraken2. The heatmap shows the proportion of contigs linked to each ARG that was classified as the taxon on the right. Two spike-in strains were added to the sample: *Enterococcus faecium* (E745) and *Escherichia coli* (E3090) strain were spiked-in to the stool sample, and the ARGs that these strains carried are highlighted in yellow and purple, respectively. **a)** shows hosts classified to species-level. **b)** shows hosts classified to genus-level.

ARGs were then linked to their microbial hosts for the other datasets. As with G_3C, some ARGs were linked to few microbial hosts, whereas others were linked to a wide range of hosts (Figure 3.14), and the proportions of ARGs successfully linked to their hosts were high, with 6/11, 9/15, 23/30, 16/23, and 6/7 for D_HiC, P_HiC, Y_3C_A, Y_3C_B, and M_3C, respectively (Table 3.6).

The same ARGs that were successfully linked in three or more datasets were compared. Some of the shared ARGs linked to the same hosts across datasets (Figure 3.15), whereas others linked to multiple diverse hosts (Figure 3.16). The *tet*(X) (Figure 3.15a), *tet*(Q) (Figure 3.15b), and *erm*(F) (Figure 3.15c) genes were linked predominantly to *Alistipes*, and *Bacteroides*, both from the order Bacteroidales, in all datasets that they were present in. The β-lactamase gene *cfxA3* was also linked to *Bacteroides* in all datasets that it was present in. Conversely, *tet*(O)*, tet*(40), *lnu*(C), *cat*, *ant(6)-Ia*, and *tet*(W) (Figure 3.16a-f) showed a wide range of hosts across the datasets, with *tet*(W) linking to over 20 taxonomic classifications in total across five datasets.

Overall, these results indicate that the ARGs identified in the assemblies were able to be linked to their microbial hosts using meta3C/Hi-C data, with stringent filtering to minimise the impact of spurious links, revealing some genes to be promiscuous and linking to a wide range of gut bacteria.

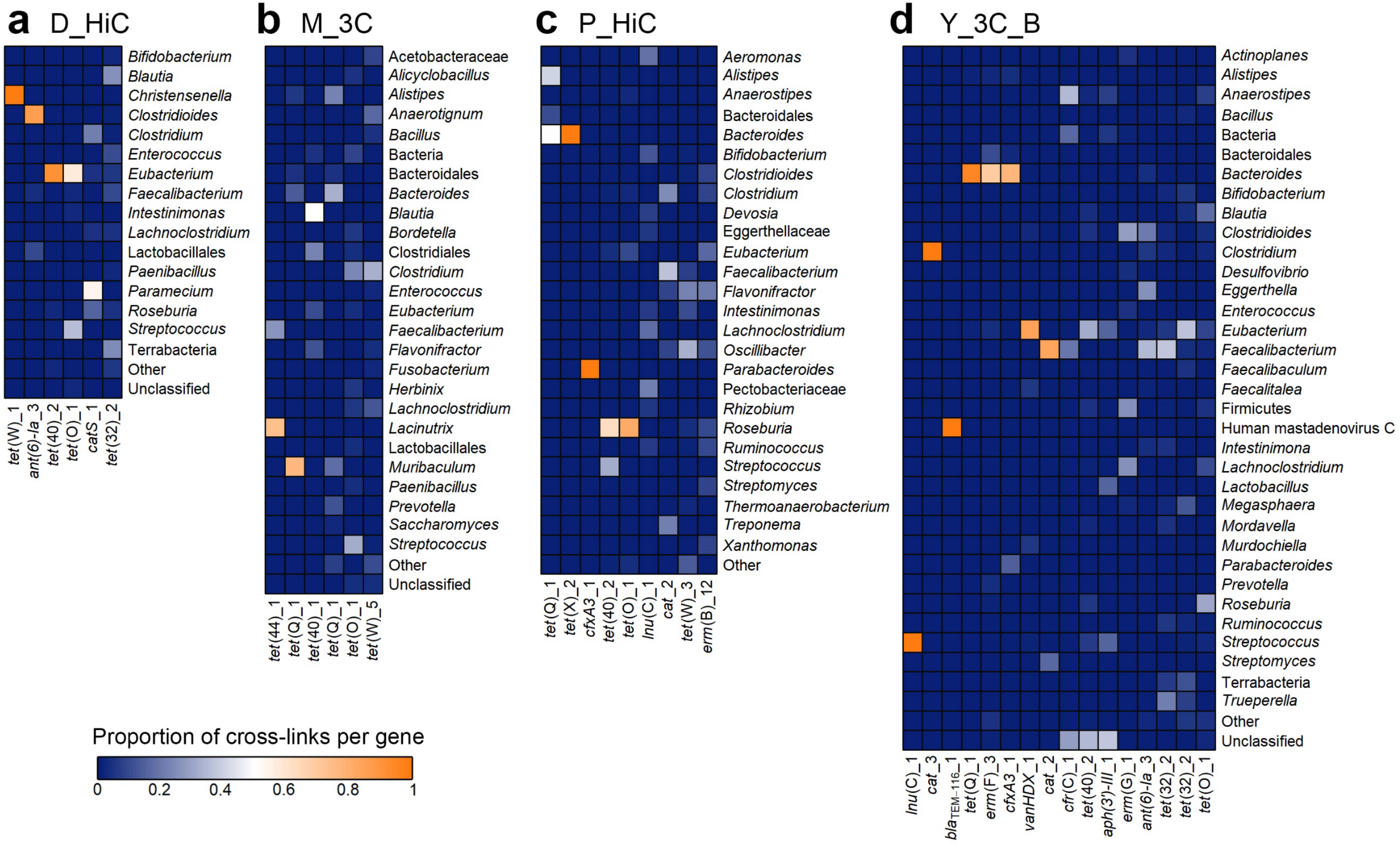


Figure 3.14 continued on next page.

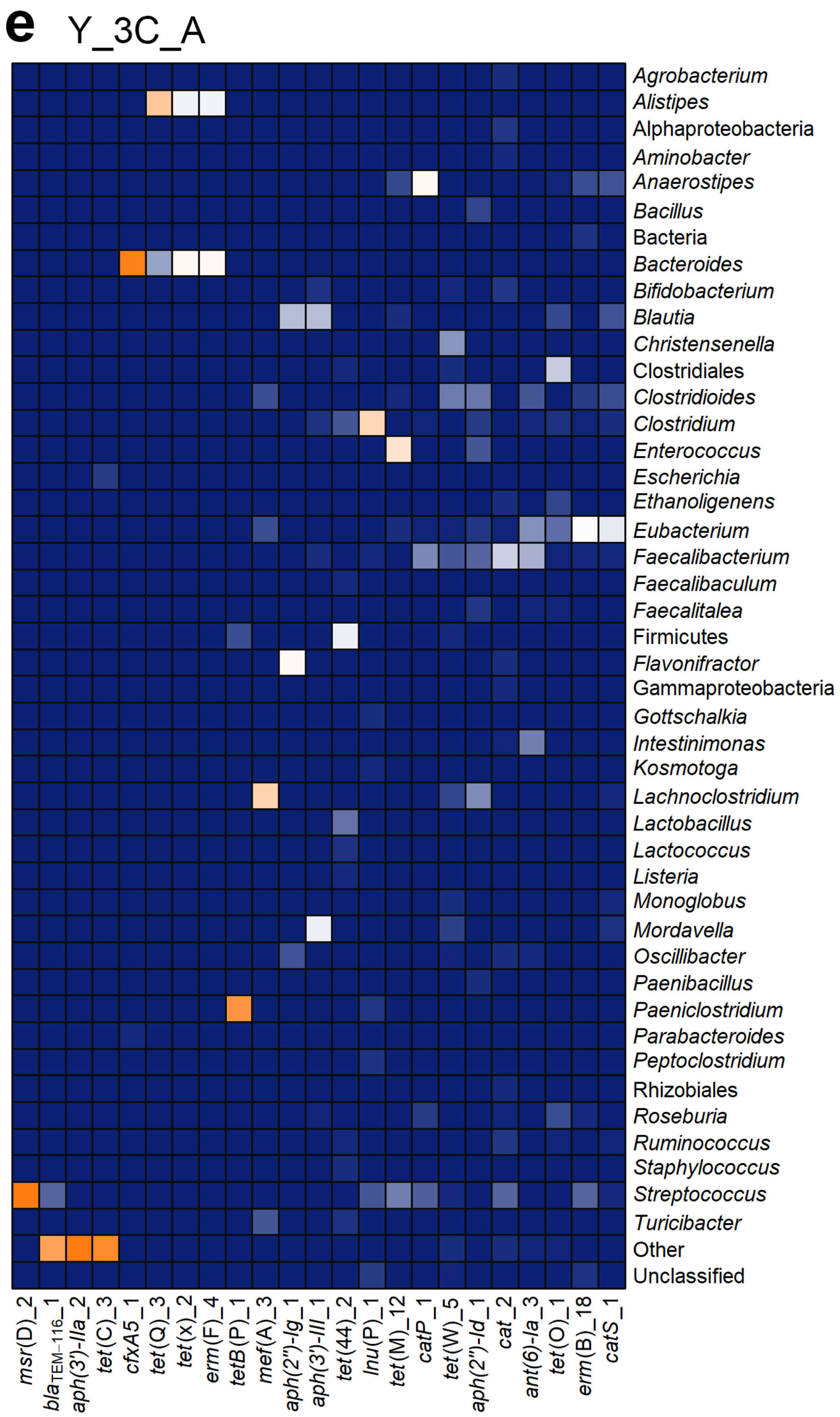


Figure 3.14. Heatmap showing antimicrobial resistance genes (ARGs) linked with their microbial hosts for downloaded 3C/Hi C datasets.
Contigs linked to ARG-containing contigs were taxonomically classified using Kraken2. The heatmaps show the proportion of contigs linked to each ARG that was classified as the taxon on the right. Where there were multiple taxa that made up a proportion of no more than 0.02 for any ARG in that dataset, they have been grouped into "Other".

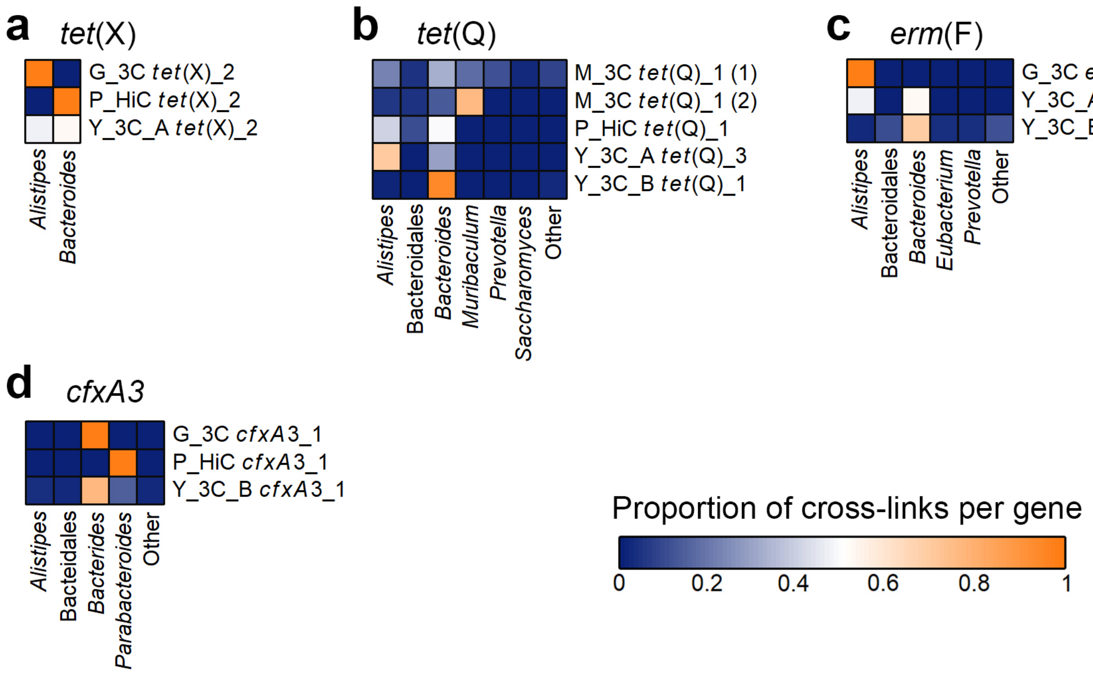


Figure 3.15. Individual ARG-host association heatmaps for any ARG that linked to similar hosts in at least 3 datasets.

Each heatmap represents an individual gene named above it. Each row represents a different dataset, and each column is a different taxon that was associated with the ARG. The gene name next to the dataset is the full name of the gene identified in that dataset according to the ResFinder database. Where there were multiple taxa that made up a proportion of no more than 0.02 linking to the ARG in any dataset, they have been grouped into "Other".

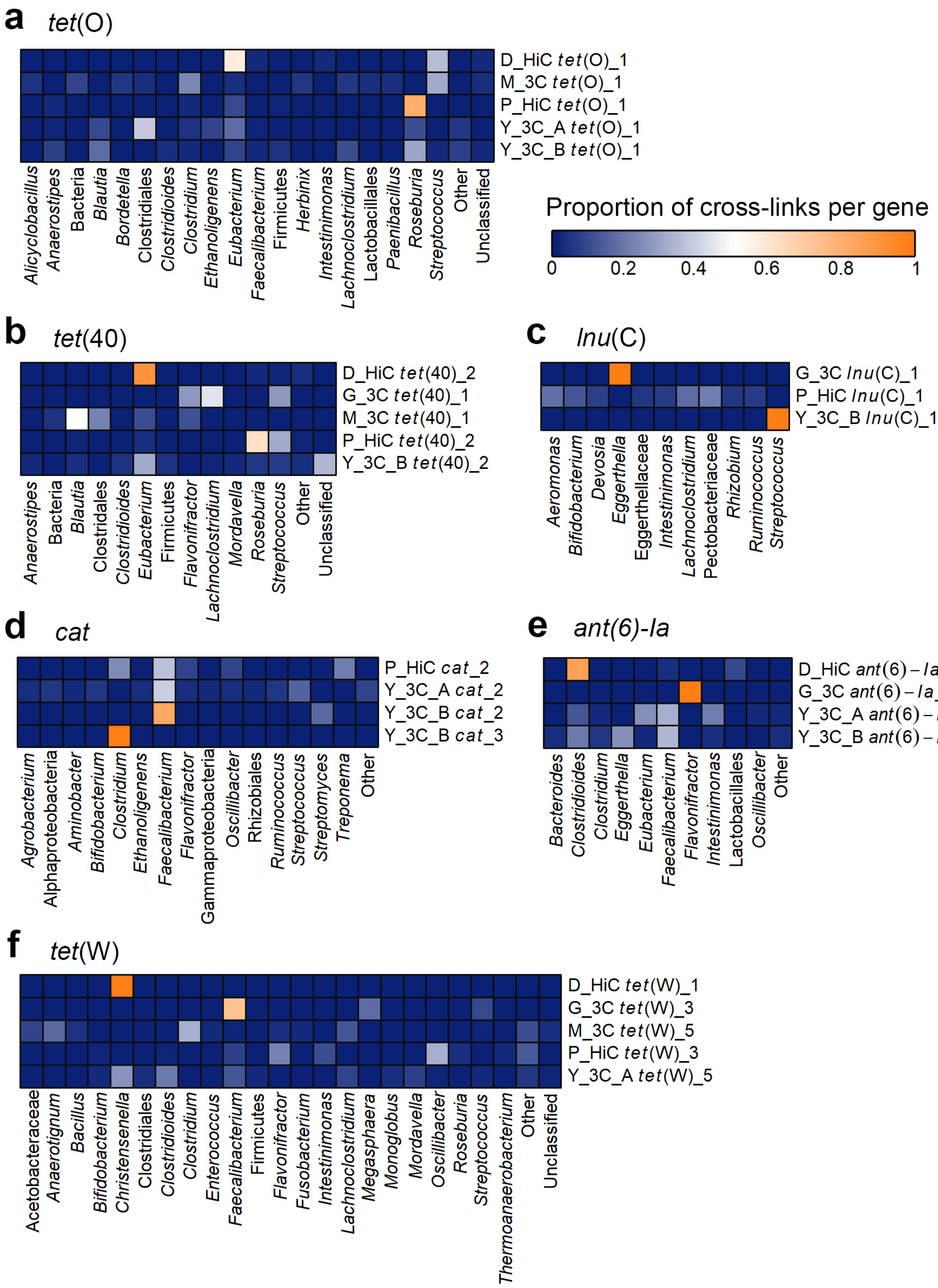


Figure 3.16. Individual ARG-host association heatmaps for any ARG that linked to different hosts in least 3 datasets.

Each heatmap represents an individual gene named above it. Each row represents a different dataset, and each column is a different taxon that was associated with the ARG. The gene name next to the dataset is the full name of the gene identified in that dataset according to the ResFinder database. Where there were multiple taxa that made up a proportion of no more than 0.02 linking to the ARG in any dataset, they have been grouped into “Other”.

3.3 Discussion

Previous studies have implemented 3C-based methods on the gut microbiota (Marbouty *et al.*, 2017; Press *et al.*, 2017; DeMaere *et al.*, 2020; Kent *et al.*, 2020; Yaffe and Relman, 2020). These studies have either not focused on linking ARGs to their microbial hosts (Press *et al.*, 2017; Yaffe and Relman, 2020), or only discussed ARG-host associations at the class-level (Kent *et al.*, 2020), and importantly, the issue of problematic noise arising from spurious intercontig reads has so far not been studied in detail. This study sought to implement meta3C on a human stool sample to link ARGs to their microbial hosts, as well as compare the 3C data generated here to previously published 3C/Hi-C datasets with the aim to optimise analysis methods for 3C/Hi-C data by reducing the impact of spurious intercontig reads.

The findings here reveal that published 3C/Hi-C-based gut microbiota studies show varying degrees of success of achieving high-quality 3C/Hi-C data from stool samples. The results demonstrate that all 3C/Hi-C datasets contain a background level of spurious intercontig reads, and present various steps that can be taken to eliminate many of these to reduce the amount of noise interfering with analysis of the 3C/Hi-C data. Filtering was followed by steps to link ARGs to their microbial hosts, to create a workflow that I have named H-LARGe (Host-Linkage to Antimicrobial Resistance Genes), outlined in Figure 3.17.

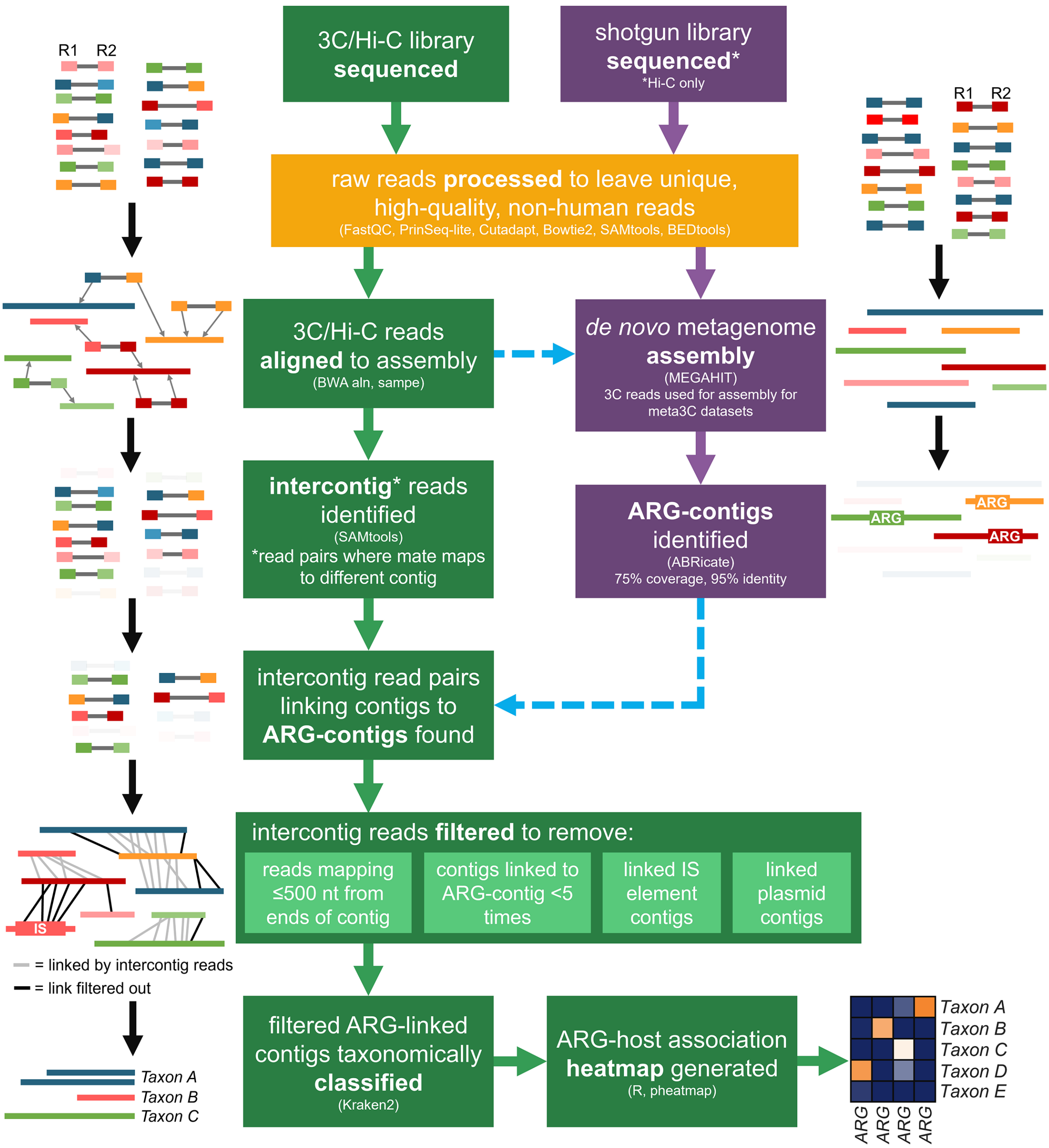

R1 R2
3C/Hi-C library sequenced
shotgun library sequenced*
*Hi-C only
R1 R2
raw reads processed to leave unique, high-quality, non-human reads
(FastQC, PrinSeq-lite, Cutadapt, Bowtie2, SAMtools, BEDtools)
3C/Hi-C reads aligned to assembly
(BWA aln, sampe)
de novo metagenome assembly
(MEGAHIT)
3C reads used for assembly for meta3C datasets
intercontig* reads identified
(SAMtools)
*read pairs where mate maps to different contig
ARG-contigs identified
(ABRicate)
75% coverage, 95% identity
ARG
ARG
ARG
intercontig read pairs linking contigs to ARG-contigs found
intercontig reads filtered to remove:
reads mapping ≤500 nt from ends of contig
contigs linked to ARG-contig <5 times
linked IS element contigs
linked plasmid contigs
IS
= linked by intercontig reads
= link filtered out
filtered ARG-linked contigs taxonomically classified
(Kraken2)
ARG-host association heatmap generated
(R, pheatmap)
Taxon A
Taxon B
Taxon C
Taxon A
Taxon B
Taxon C
Taxon D
Taxon E
ARG
ARG
ARG
ARG

Figure 3.17. Overview of the H-LARGe workflow.
The flowchart shows the major stages of the H-LARGe (Host-Linkage to Antibiotic Resistance Genes) workflow. Programs used for each stage are shown in brackets in each box. Stages using the 3C/Hi-C reads are in green boxes, and those using shotgun metagenomic reads are shown in purple (only applicable to Hi-C datasets). Blue dashed arrows indicate where stages of the workflow require outputs from another stage. A schematic representation of each stage is shown outside of the boxes (R1 and R2 = read 1 and read 2, respectively).

Using the H-LARGe workflow, 87 ARGs were linked to their microbial hosts across the six datasets, including 27 in the meta3C data first described in this thesis. These included 6 ARGs known to be plasmid-borne in two spike-in strains that were added to G_3C, showing that meta3C was able to link ARGs carried in plasmids to chromosomal DNA of their microbial hosts in a human stool sample.

3.3.1 Implementing meta3C on a human stool sample

It took several attempts to get the meta3C protocol to work. The main issue may have been using formaldehyde that was stabilised with methanol as changing this to methanol-free formaldehyde resulted in greater DNA yields. This could be due to the presence of methanol (added to stabilise the formaldehyde solution during storage) increasing the cell permeability (Hobro and Smith, 2017), causing over cross-linking of the DNA. Following this, all future 3C/Hi-C experiments should not use formaldehyde stabilised with methanol. Before sequencing, the meta3C library was quantified using gel electrophoresis as advised in the protocol (Foutel-Rodier *et al.*, 2018), and appeared to have been successful. However, it should be noted that the gel electrophoresis only indicates successful digestion and ligation and does not indicate that a meta3C library has been truly successfully generated as it gives no indication on whether the cross-linking step was effective. Therefore, although useful for showing that digestion and ligation was achieved, this method of quantification should be used with caution in future experiments, and additional checks of the quality of the meta3C library should be done before full-depth sequencing. For example, some low-depth sequencing of the sample could be performed, and the reads mapped back to a metagenomic assembly, preferably of the same sample that the meta3C procedure was carried out on, to estimate the proportion of intercontig reads present.

More importantly, a spike-in should be included. Here, a spike-in of two strains of *E. coli* and *E. faecium* were added to the stool sample prior to meta3C, which were useful for downstream analysis. The metagenomic profile of the reads from the meta3C library showed that these were added in equal amounts. This was the first study to add spike-ins during proximity ligation of a stool sample, although Marbouty *et al.* (2017) added meta3C reads post-sequencing from three bacterial species into the mouse faecal meta3C reads before downstream analysis, and a study implementing Hi-C on wastewater used an *E. coli* spike-in strain in one of the samples (Stalder *et al.*, 2019). Whilst the G_3C spike-ins were useful in analysis of the meta3C data generated here, the strains used may not have been optimal. Both spike-ins were species of bacteria that are commonly found in the human gut microbiota. Cross-sectional studies of healthy human adults have shown that *E. coli* is present in the microbiota of over 90% of individuals (Tenaillon *et al.*, 2010). Similarly, *E. faecium* is a common member of the human gut microbiota, with one study detecting it in 100% of human faecal samples tested (Layton *et al.*, 2010). This meant that any *E. coli* or *E. faecium* strains that were naturally present in the sample used would have been masked by the spike-in strains, complicating the detection of potential ARG-host links to these species. For future spike-ins, species that are less likely to be naturally present in the sample type that is being studies should be considered. Previous metagenomic studies have used spike-ins to aid estimation of species abundance in human faecal samples, and have used environmental bacteria that are not found in the human gut microbiota such as the extreme soil halophile *Salinibacter ruber* and other soil bacteria (Stämmler *et al.*, 2016). Spike-ins from these environmental bacteria would allow the chance for all naturally present hosts of ARGs in faecal samples to be linked in future 3C/Hi-C experiments. However, to be useful as a spike-in when linking ARGs to their hosts, the spike-in strains should have multiple plasmids containing ARGs, which may be harder to find in these environmental bacteria.

After generating a *de novo* metagenomic assembly, 37 ARGs were identified in the sample, with 17 ARGs originating from the spike-ins. The non-spike-in ARGs present included several erythromycin resistance genes (*erm*(B), *erm*(F), *erm*(Q)), aminoglycoside resistance genes (*aph(2'')-Ib*, *aac(6')-Im*, *ant(6)-Ia*), and tetracycline resistance genes (*tet*(32), *tet*(40), *tet*(M), *tet*(W), *tet*(X), *tetA*(P), *tetB*(P)). These families of resistance genes are prevalent in the gut microbiota, and have been commonly identified at high abundance in human faecal samples in several studies (Forslund *et al.*, 2013; Hu *et al.*, 2013).

The fractions of intercontig read pairs were 1.2% and 2.1% for the two meta3C libraries (prepared with restriction enzymes HpaII and MluCI, respectively) generated in this study. This is lower than expected from the protocol which suggested that 10-15% of the reads will be from cross-linked fragments (Foutel-Rodier *et al.*, 2018). However it is within the range of another study by the same authors using meta3C on human stool samples, which reported intercontig reads ranging from 1.92-14.58% (Marbouty *et al.*, 2021). Additionally, a study that tested the meta3C protocol on a synthetic community also reported that most of their experiments resulted in around 1% proximity ligation read rate (Darling and Liu, 2015).

3.3.2 Comparison of published gut microbiota 3C/Hi-C studies

Various studies have been published that use 3C-based techniques on human stool samples (Press *et al.*, 2017; DeMaere *et al.*, 2020; Kent *et al.*, 2020; Yaffe and Relman, 2020). The datasets generated from these, along with a meta3C study on mouse faeces (Marbouty *et al.*, 2017), were downloaded and analysed along with my own meta3C sequencing data.

The metagenomic profiles of all the datasets were quite comparable, with the identified taxa mostly comprised of bacteria belonging to the phyla Bacteroidetes and Firmicutes, which are thought to make up around 90% of all species in the gut microbiota of healthy humans (Eckburg *et al.*, 2005). There were some differences between the compositions of the shotgun metagenomic reads and 3C/Hi-C reads of the same samples (Figure 3.5). This same observation was made by Kent *et al.*, who suggested that this variation could be due to the levels of DNA-binding proteins available during the cross-linking step, genome-wide patterns and structures of the DNA, and the positions of restriction enzyme cut sites within the different chromosomes (Marbouty *et al.*, 2014; Crémazy *et al.*, 2018; Kent *et al.*, 2020). Differing methods for cell lysis during 3C/Hi-C and DNA extraction for shotgun metagenomic sequencing could also have affected the diversity of bacteria sequenced. The DNA extraction methods for shotgun metagenomic sequencing in most widely used protocols tend to use more vigorous mechanical disruption (Yang *et al.*, 2020) compared to many 3C/Hi-C protocols, which use more gentle mechanical disruption by vortexing (Press *et al.*, 2017; Kent *et al.*, 2020; Uritskiy *et al.*, 2021) or enzymatic cell lysis methods (Yaffe and Relman, 2020). Studies have shown that mechanical cell disruption by bead-beating results in improved DNA extraction from human faecal samples, and allows for higher bacterial diversity compared to enzymatic lysis methods, particularly for Gram-positive species (Salonen *et al.*, 2010). However, overly harsh mechanical lysis of the cells during 3C/Hi-C can lead to chromatin damage and significantly reduce the final yield of the 3C/Hi-C library (Belton *et al.*, 2012), so more gentle approaches of cell lysis appear to be a necessity.

For all the samples, including both shotgun and 3C/Hi-C reads, the majority of the taxonomic profile was unclassified by MetaPhlAn3. The 'unknown' estimation is a new feature introduced to version 3.0 of MetaPhlAn, and reports the portion of the taxonomic profile that do not correspond with taxa in any of the databases used, taking into account average genome length (Beghini *et al.*, 2021). Therefore, previous studies have not reported the unknown estimation of human faecal samples calculated using MetaPhlAn. However, it is unsurprising that around 60% of the taxonomic profiles were unknown in these samples, as it has been previously estimated that 40-50% of bacteria in the human gut microbiota lack a reference genome (Sunagawa *et al.*, 2013; Nayfach *et al.*, 2019). The M_3C dataset had the highest proportion of unknown taxa (94%), likely due to this dataset coming from a mouse faecal sample, which is less catalogued than the human gut microbiota (Xiao *et al.*, 2015).

The ARG profiles of the samples were largely comparable. Tetracycline resistance genes were the most abundant, being present in all samples and with 90/96 samples containing multiple *tet* genes. Macrolide resistance genes were also common, with 70/96 samples containing at least two. The majority of samples (60/96) contained the *ant(6)-Ia* gene, which was also identified in all 162 samples of another study (Hu *et al.*, 2013). There were 31 different *bla* genes present in 25/66 of the K_HiC samples that came from neutropenic patients who were prescribed various antibiotics including the broad-spectrum β-lactam antibiotic piperacillin (Kent *et al.*, 2020). Two *bla*$_{\mathrm{OXA}}$ genes were also present in P_HiC, however were not present in P_SG, the accompanying shotgun metagenomic data from the same sample. All of the samples saw variation in the amount of congruence between the ARG profiles of the 3C/Hi-C data compared to the accompanying shotgun data, and all showed some dissimilarity. This could be due to the same reasons as the differences in the taxonomic profiles and could indicate that the hosts of these genes generated less 3C/Hi-C reads. Additionally, both the taxonomic profiles and the ARGs present between the 3C/Hi-C and shotgun metagenomic datasets of the same sample may simply have been impacted by the composition of the section of stool sample used if the sample was not fully homogenised before being aliquoted. However, studies have shown that homogenisation has no significant effect on stool sample composition, and the cell lysis method plays a more important role (Krauth *et al.*, 2012; Santiago *et al.*, 2014).

In order to detect reads that originated from cross-linked fragments of DNA, intercontig read pairs, where each mate maps to a different contig in the assembly, are identified in the 3C/Hi-C data. To identify intercontig reads in each dataset, each read was mapped to the corresponding assembly. During this study, only the first 50 bp of each read were mapped to the assembly to minimise the chance of reads containing a cross-link being discarded. If, for example, a 150 bp read from the 3C/Hi-C dataset spanned a ligation site, it would map to multiple contigs and have a low mapping quality and thus be discarded. Therefore, only mapping the first 50 bp of all reads reduces the chance of this issue occurring. This method was also employed by Yaffe and Relman (2020), who removed the first 10 bp of each read before mapping the following 40 bp.

The proportion of intercontig reads calculated here during the H-LARGe workflow considerably varied between each dataset, ranging from 0.64% in the K_HiC dataset to 13.74% in P_HiC. These values also differed from the reported figures in the respective original studies. In the P_HiC study, the authors only found 2.2-3.6% of their Hi-C read pairs were intercontig (Press *et al.*, 2017). Similarly, DeMaere *et al.* (2020) predicted that their fraction of Hi-C pairs was only 0.36-0.67% using the program qc3C (DeMaere and Darling, 2021), whereas around 7.11% of the D_HiC reads were intercontig when analysed here. The two samples from Y_3C had 6.45% and 7.59% intercontig reads, similar to the 5.6% previously reported (Yaffe and Relman, 2020). The M_3C dataset was the only one found to have a lower proportion of intercontig reads here than what was predicted in the original study (9.73% here compared to ~25% reported by the authors (Marbouty *et al.*, 2017)). This was likely due to the assembly generated by Marbouty *et al.* (2017) being much more fragmented, as it was comprised of 553,310 contigs (513 Mb total) compared to only 116,057 contigs (481 Mb total) in this study. It was surprising that the 3C datasets seemed to have proportions of intercontig reads in similar ranges as the Hi-C datasets, despite Hi-C having an additional enrichment step that should, in theory, increase the number of reads originating from cross-linked fragments of DNA (Lieberman-Aiden *et al.*, 2009).

The most unexpected finding from the proportion of intercontig reads was that all samples of the K_HiC dataset contained very few intercontig reads. The authors did not mention the fraction of intercontig reads in their publication (Kent *et al.*, 2020). The particularly low-level of intercontig reads compared to the other datasets analysed here suggests that the Hi-C procedure may not have worked effectively in the K_HiC samples. This could explain some unusual results reported in their study, such as observing multiple occurrences of inter-phyla HGT, including from *Klebsiella* to *Enterococcus*, despite evidence of inter-phyla HGT in the human gut microbiome being very limited (Porse *et al.*, 2018). After analysing the shotgun metagenomic reads in the same way as the Hi-C reads by mapping them back to the assembly in each sample of K_HiC, the proportion of spurious intercontig reads was higher than the Hi-C intercontig reads, substantiating the idea that true cross-linking had not been achieved for this Hi-C dataset. This also demonstrated that there was a background noise of spurious intercontig reads that could interfere with the analysis of the intercontig reads in the 3C/Hi-C datasets, so the extent of this phenomenon was investigated further. These observations led to the development of a standardised method to minimise the number of spurious intercontig reads in meta3C/Hi-C datasets.

3.3.3 Filtering of spurious intercontig reads

The method for identifying intercontig reads in the H-LARGe workflow by mapping 3C/Hi-C reads back to their corresponding assemblies has been used in all metagenomic 3C/Hi-C studies (Marbouty *et al.*, 2017; Press *et al.*, 2017; Stalder *et al.*, 2019; DeMaere *et al.*, 2020; Kent *et al.*, 2020; Yaffe and Relman, 2020), albeit with some differences in methods of assembly and mapping. However, it is known that spurious intercontig reads create noise in 3C/Hi-C data (DeMaere and Darling, 2018; Yaffe and Relman, 2020).

Spurious intercontig reads can be the result of the formation of spurious ligation products between DNA that originated in different hosts during the experimental process of 3C/Hi-C (Nagano *et al.*, 2015). They can also occur from sequencing errors (DeMaere and Darling, 2018), and as the results in this study show, they are an inherent issue in short-read sequencing, being present in both the shotgun metagenomic sequence datasets and the WGS short-read data analysed here. The issue of spurious intercontig reads has been relatively underappreciated by the previous studies performing 3C-based techniques on the gut microbiota. On the basis of the analysis performed in this chapter, it is clear that they have potential to significantly disrupt the interpretation of data by misassigning hosts to functional genes being investigated. Indeed, when analysing Hi-C reads from wastewater samples, Stalder *et al.* (2019) found that several clusters of contigs characterised as Firmicutes, Alphaproteobacteria, and Betaproteobacteria were linked by Hi-C reads to the *E. coli* spike-in strain that had been added to the sample. This *E. coli* spike-in was also linked to several ARGs and plasmids that were not present in the spike-in strain, and the authors concluded that these Hi-C links were spurious and likely due to the high abundance of the spike-in strain (Stalder *et al.*, 2019). The authors suggested that these ARGs and plasmids were probably present in other strains of *E. coli* that were present in the sample (Stalder *et al.*, 2019), however this cannot be confirmed without culturing of the sample. Press *et al.* (2017) also observed results caused by spurious intercontig reads, including a *Eubacterium eligens* megaplasmid being linked by Hi-C reads to another large plasmid originating from a species in the Bacteroidetes phylum.

The majority of the original studies that generated the datasets analysed in this chapter did little to remove spurious intercontig reads during their analysis. Like the analysis pipeline used in this study, most studies removed reads aligning with a low MAPQ and reads mapping to multiple contigs (Press *et al.*, 2017; Kent *et al.*, 2020; Yaffe and Relman, 2020). Some also required the presence of restriction sites on the contigs being mapped to (Kent *et al.*, 2020; Yaffe and Relman, 2020), although as they were using 4-cutter restriction enzymes, these restriction sites could be quite common in the assembly. Yaffe and Relman (2020) did the most to reduce spurious intercontig reads from interfering with the data analysis by developing a pipeline that included probabilistic modelling of experimental noise to determine the likelihood of links made using the 3C data being real. This method allowed them to detect and remove thousands of spurious links (Yaffe and Relman, 2020). A recent study also developed a workflow to remove spurious links through normalisation of Hi-C data based on zero-inflated negative binomial regression frameworks (Du *et al.*, 2022), although this method has not been applied to 3C/Hi-C experiments on the human gut microbiota.

The results in this chapter show that spurious intercontig reads often account for ~2% of shotgun metagenomic reads, indicating that a considerable fraction of identified intercontig reads in meta3C/Hi-C datasets, even after removal of low-quality mapping, could be spurious reads not originating from cross-linked fragments. By comparing 3C data generated here to WGS reads of the spike-ins, it was observed that intercontig reads, both from 3C data and normal short-read data, had similar insert sizes when mapped to the spike-in genomes, suggesting that methods using insert sizes to differentiate real and spurious intercontig reads (such as the program qc3C (DeMaere and Darling, 2021)) may only work to remove a subset of spurious intercontig reads. Intercontig reads from both 3C data and WGS data were also more likely to map near to IS elements. This indicates that many spurious intercontig reads could be caused by repeats in the genome causing fragmentation of the assembly into smaller contigs. Repeat regions in the genome, like IS elements, that are longer than the read length cause breaks in the assembly as the assembly software will not be able to determine which sequences the repeat is between in the genome. This results in fragmented contigs and separate individual contigs for the repeats, or contigs that are flanked by the repeats (Acuña-Amador *et al.*, 2018). This is especially an issue for bacteria, as repeat regions are estimated to make up around 5-10% of the total genome (Shapiro and Von Sternberg, 2005). As IS elements can be fairly short (<2 kb) (Gonçalves *et al.*, 2020), it is possible that read pairs could be made up of reads that were sequenced either side of an IS element, which would result in them being identified as intercontig. Similarly, one read of the pair could map to an IS element contig, or even a contig flanked by repeats. This could cause not only spurious intercontig reads, but also false host-associations of contigs during analysis of 3C data, as the same IS elements can be present in different species (Siguier, Gourbeyre, and Chandler, 2014).

Furthermore, the results also showed that intercontig reads were significantly more likely to map within the first or last 500 nt of a contig compared to non-intercontig reads for both the 3C and WGS reads for the spike-ins. This supports the concept that many spurious intercontig reads are caused by fragmented assemblies, as it suggests that read pairs from regions of DNA that were fragmented into separate contigs during the assembly process were being identified as intercontig reads. Analysis of intercontig reads from the downloaded 3C/Hi-C datasets and the spurious intercontig reads in the accompanying shotgun datasets confirmed that spurious intercontig reads tended to map near the ends of contigs. Importantly, it showed that the 3C/Hi-C intercontig reads in datasets that had a high proportion of intercontig reads had lower levels of intercontig reads that mapped within 500 nt of the ends of a contig. This implied that, perhaps unsurprisingly, higher levels of true intercontig reads were better able to mask the noise from spurious intercontig reads. For the K_HiC dataset, the proportion of intercontig reads mapping near the start or end of a contig was the same for the Hi-C reads and the shotgun reads, further confirming that the Hi-C reads in K_HiC seemed to contain a particularly high number of spurious intercontig reads.

By filtering out reads that mapped to the first or last 500 nt of a contig, many of the spurious intercontig reads will be removed, as it will remove read pairs from regions of DNA that were fragmented in the assembly, as well as reads mapping to repeat regions that flank either end of a contig. Whilst this may remove some true intercontig reads that originated from cross-linked fragments of DNA, it is an important step to reduce the impact of spurious intercontig reads on host-ARG associations during further analysis. In addition, intercontig reads mapping to IS element contigs should be removed as these could also be spurious, and taxonomic classification of these contigs could be inaccurate. These methods are unlikely to remove all spurious intercontig reads but will at least reduce the level of noise interfering with analysis of the 3C/Hi-C reads.

3.3.4 Linking ARGs to their microbial hosts

By combining methods to reduce the number of spurious intercontig reads in each sample, a stringent approach was used whereby a contig had to be linked to an ARG-contig by at least five unique intercontig read pairs to be classed as truly linked (Figure 3.17). Similar methods have previously been used, but were less stringent, only requiring two (Kent *et al.*, 2020) or three (Stalder *et al.*, 2019) read pairs. A more rigorous approach was used here to maximise the reduction in interference of the noise from any remaining spurious intercontig reads. Using this approach, the majority of ARGs in all datasets were able to be linked to their microbial hosts.

Kraken2 was used to taxonomically classify the linked contigs to determine the hosts of the ARGs. This tool classifies sequences by finding the LCA (lowest common ancestor) of genomes containing an exact match to each *k*-mer in the sequence (Wood, Lu, and Langmead, 2019). The main limitation of using Kraken2 is that it heavily relies on correct annotations in the database being used, which is especially a problem when the query contigs differ greatly from sequences in the database (Von Meijenfeldt *et al.*, 2019). The hosts of some ARGs were very likely misclassified. For example, *tet*(Q) was linked to the fungi genus *Saccharomyces* in the M_3C dataset, and *lnu*(P) was linked to *Kosmotoga*, a thermophile found in hydrothermal systems in the ocean (DiPippo *et al.*, 2009), in the Y_3C_A dataset. It should be noted that these links represented less than 3% of the ARG-host cross-links for those genes. Another incorrectly linked ARG was *bla*$_{\text{TEM-116}}$ in the Y_3C_B dataset, which linked to a contig that Kraken2 identified as human mastadenovirus C. BLAST results for this contig reveal that it maps to synthetic constructs and expression vectors in the nt database.

Inconsistency in taxonomic resolution when using best-hit approaches can also be an issue, particularly for DNA that has been recently acquired via HGT that is classified in the database as a distantly related organism, compared to highly conserved DNA (Von Meijenfeldt *et al.*, 2019). In the G_3C dataset, several ARGs linked to multiple species of the same or similar genera, such as *erm*(Q) linking to *Clostridioides difficile*, *Clostridium perfringens*, and *Paeniclostridium sordellii*; and *tet*(X) linking to both *Alistipes finegoldii* and *Alistipes shahii*. Whilst it is possible that these genes had multiple hosts, they may have only been present in a single host closely related to these species. For the other datasets, the hosts classifications were only presented to genus-level.

Other 3C/Hi-C studies have used binning methods to improve the reliability of the gene-host link (Marbouty *et al.*, 2017; Press *et al.*, 2017; Kent *et al.*, 2020; Yaffe and Relman, 2020), as this will link genes to a group of contigs rather than just one, which could reduce the chance of misclassifying the host. Classifying these MAGs often uses phylogenetic trees of multiple marker genes, and whilst this is a well-established method, interpreting the resulting phylogeny and taxonomically classifying the MAGs still has the limitations of needing an accurate reference database (Von Meijenfeldt *et al.*, 2019).

The results of this study showed that ARGs were widespread amongst different microbial hosts, including in many known commensals in the gut microbiota. Genes that were present in multiple datasets were compared and many of them showed similar hosts across the datasets. The genes *tet*(Q), *tet*(X), and *erm*(F) were associated with the genera *Alistipes* and *Bacteroides* in nearly all datasets where those genes occurred. These three ARGs are known to be prevalent amongst *Bacteroides* species, and commonly occur together in the same strains, along with the presence of a conjugative transposon (Bartha *et al.*, 2011). These genes have also been observed in species of the *Alistipes* genus in a chicken gut microbiota (Duggett, 2016). The β-lactamase genes *cfxA3* and *cfxA5* were also always linked to contigs assigned to the *Bacteroides* genus, where these genes are known to be prevalent (Binta and Patel, 2016). Other genes were widespread and were linked to multiple hosts, including various tetracycline resistance genes. Again, this is in line with studies that show these genes are highly prevalent and widespread in the human gut microbiota (Scott *et al.*, 2000; Gueimonde, Salminen, and Isolauri, 2006).

Some novel observations were made too, such as the linkage the vancomycin resistance genes *vanHDX* to the genus *Eubacterium* in the Y_3C_B dataset. This gene has not been observed in *Eubacterium* in previous studies, although *vanD* has been found in *Ruminococcus*, a species from the same class as *Eubacterium*, in a human faecal sample (Domingo *et al.*, 2008), and a recent study found *vanD* genomic islands in five gut microbes from the order Eubacteriales (Hashimoto *et al.*, 2022). Furthermore, other vancomycin resistance genes have been found in gut commensals such as *Eggerthella lenta* and *Clostridium innocuum*, and it is thought that vancomycin resistant enterococci may arise from HGT involving these commensals (Stinear *et al.*, 2001).

3.3.5 Future work and conclusions

The results here indicated that the H-LARGe workflow used on 3C/Hi-C data could offer realistic insights into the hosts of the ARGs in the samples, despite the inherent limitations discussed above. It should be noted that these data only give an indication of the microbial hosts, and do not confirm the host of the ARG with absolute certainty. Similarly, the results do not offer much insight into the genomic context of the ARG. For that, the 3C/Hi-C data should be coupled with culturing where it can provide insights for targeted culturing and isolation of the hosts of ARGs in the sample to validate the 3C/Hi-C findings and allow the genomic context of the ARGs to be investigated. One could also explore the use of long-read sequencing to improve the metagenomic assembly to more reliably link resistance genes to larger contigs, for which taxonomic assignment will be more likely to be accurate.

From the results in this chapter, it is not obvious whether meta3C or Hi-C libraries are better for linking ARGs to their hosts in human stool samples. Two of the meta3C datasets (M_3C, Y_3C) had as high or greater proportions of intercontig reads as two of the Hi-C datasets analysed here (P_HiC, D_HiC). This is somewhat surprising, as meta3C libraries are thought to have less reads originating from cross-linked fragments of DNA due to the lack of an enrichment step. At the same time, the Hi-C data from Kent *et al.* (2020) seemed to have no reads originating from true cross-linked fragments of DNA, despite the enrichment step. The meta3C performed in this study (G_3C), which used a similar protocol as the Marbouty *et al.* (2017) study, also resulted in a rather low amount of intercontig reads. The same group recently described a derivative of meta3C called metaHiC. This method is reported to yield a greater number of cross-links than meta3C (up to 14.9% for metaHiC compared to up to 5.1% for meta3C), and is further improved when ligation is performed at room temperature, achieving an average proportion of 32% cross-linked reads (Marbouty *et al.*, 2021). However, the two Hi-C studies analysed in this present study that achieved a relatively high proportion of intercontig reads both used the ProxiMeta Hi-C protocol from PhaseGenomics (Press *et al.*, 2017; DeMaere *et al.*, 2020). Therefore, ProxiMeta Hi-C may be the best and most reliable option for future 3C-based experiments on the human gut microbiota.

Overall, the findings in this chapter demonstrate that 3C/Hi-C data contain a substantial background noise from spurious intercontig reads, that could confound host-ARG associations during analysis. Several steps can be taken to reduce the impact of these spurious intercontig reads, including discarding reads that map near to the ends of a contig, removing reads mapping to IS element contigs, and requiring at least five unique intercontig read pairs to link two contigs together. After developing and using a novel analysis workflow, H-LARGe, the hosts of many ARGs were identified, which showed that resistance genes are widespread among many species in the human gut microbiota. Future work will use Hi-C on human stool samples to link ARGs to their hosts using the H-LARGe workflow developed here, and culturing will then be performed to confirm the findings from the proximity ligation data.

CHAPTER 4

HI-C PROFILING OF RESERVOIRS OF ANTIBIOTIC RESISTANCE GENES IN THE HUMAN GUT MICROBIOTA

4.1 Introduction

In the previous chapter, I performed meta3C on a human faecal sample and reanalysed existing meta3C/Hi-C datasets. Analyses of these datasets allowed for the development of a bioinformatic workflow that filtered spurious intercontig reads to reduce the impact of noise in the data, and linked ARGs to their hosts. The findings also revealed that the meta3C protocol (Foutel-Rodier *et al.*, 2018) that I used did not lead to high-levels of cross-linking compared to datasets from previous studies, particularly those Hi-C studies that used the ProxiMeta kit from PhaseGenomics (Press *et al.*, 2017; DeMaere *et al.*, 2020).

One of the limitations of the H-LARGe workflow was the use of Kraken2 to taxonomically classify the ARG-linked contigs. Other studies have implemented binning methods on metagenomic assemblies using 3C/Hi-C data (Press *et al.*, 2017; Baudry *et al.*, 2019; Kent *et al.*, 2020; Yaffe and Relman, 2020), which could improve the accuracy of taxonomic classification as it allows clusters of contigs to be classified rather than a single contig, reducing the chance of misclassification. Furthermore, the addition of spike-in strains into the meta3C sample in Chapter 3 allowed validation of the bioinformatic workflow, as well as offering insights into the noise caused by spurious intercontig reads. However, the *E. coli* and *E. faecium* strains used as spike-ins in Chapter 3 meant that any strains of *E. coli* and *E. faecium* naturally present in the sample were likely to be masked by the spike-ins.

In this chapter, I aimed to use ProxiMeta Hi-C to link ARGs to their hosts in human faecal samples, using a strain of *Acinetobacter pittii* as a spike-in. This study also aimed to implement clustering of contigs into bins into the bioinformatic workflow from Chapter 3 to improve taxonomic classification of ARG-linked hosts.

4.2 Results

4.2.1 Preparation of Hi-C libraries and shotgun libraries

Hi-C was performed on four human stool samples, including the sample used for meta3C in Chapter 3. As a control for the experimental stage and bioinformatic analysis of the Hi-C sequencing data, an ARG-plasmid carrying strain of *Acinetobacter pittii* (OB7) (GenBank accession GCA_002999215.3) was added as a spike-in to the stool samples at approximately 0.5%. *A. pittii* is a rare coloniser of the human gut (Dijkshoorn *et al.*, 2005). Due to the spike-ins in the G_3C meta3C library prepared in Chapter 3 being present at a higher level than expected, the approximate number of cells/g of stool estimated by Sender, Fuchs and Milo (2016) (1×10^{11} cells/g) was used to calculate the number of cells for the spike-in control. This resulted in approximately 5×10^{7} cells being added to 100 mg of stool before preparation of the Hi-C libraries. During the cell lysis stage of the ProxiMeta Hi-C protocol, fatty content in samples H1 and H3 caused the Recovery Beads to be weakly adhered to the magnet, resulting in some loss of the pellet. Subsequently, these samples had to undergo additional PCR cycles (16 instead of 12) during the library amplification step. Prepared Hi-C libraries were quantified on an Agilent TapeStation (Figure 4.1a). This confirmed the DNA fragment sizes were in the desired range of 300-1,000 bp, with the majority of DNA being between 290-650 bp for all libraries.

DNA extractions were also performed on the same stool samples for shotgun sequencing, with 1.25×10^{8} cells of the OB7 spike-in being added to 250 mg of stool prior to DNA extraction. The extracted DNA fragment sizes were analysed on the Agilent TapeStation to check for degradation. The results showed that the majority of DNA for each sample was above 7 kbp, however two samples did show some degradation, with H1 having a strong smear down to 250 bp, and H4 showing a smear between 7,000-1,200 bp (Figure 4.1b). The extracted DNA was then prepared for shotgun sequencing and both the shotgun libraries and Hi-C libraries were sequenced on an Illumina NovaSeq 6000 2x150 bp paired-end platform.

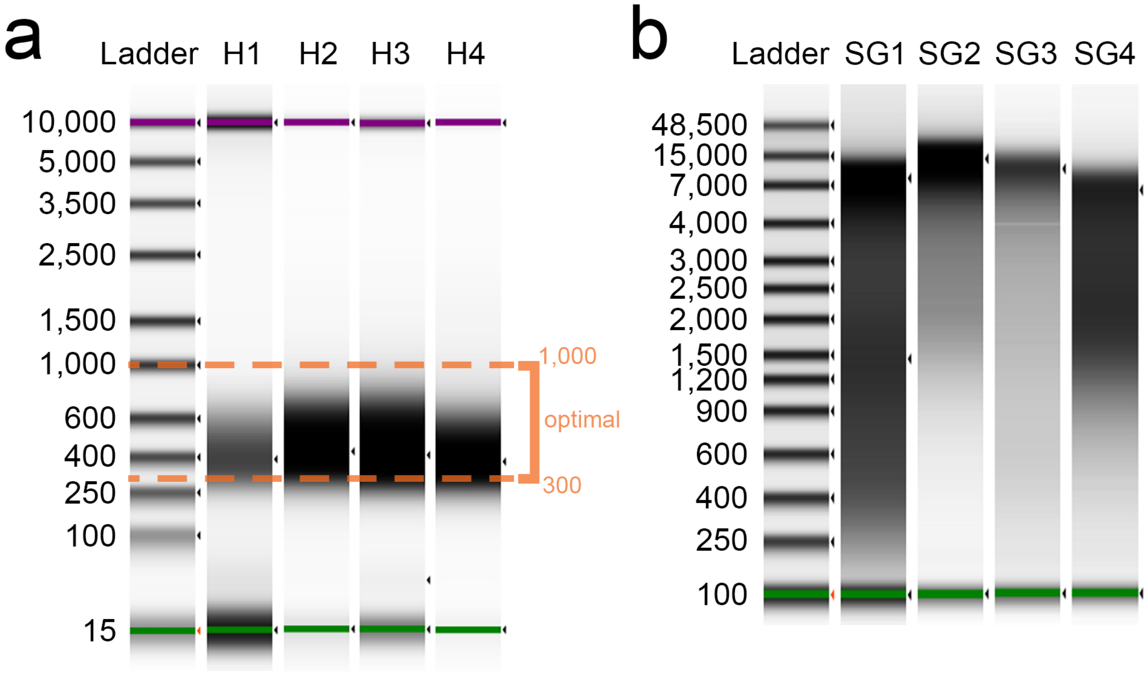


Figure 4.1. Quantification of Hi-C libraries and extracted DNA.
DNA was quantified on an Agilent TapeStation after **a)** ProxiMeta Hi-C library preparation and **b)** DNA extraction from the faecal samples using the FastDNA Spin Kit for Soil. The number to the left of the lanes represent the band size in base pairs (bp). The lane names indicate the sample number (1-4) and type of sample loaded (H/SG), with H indicating the Hi-C library, and SG indicating the DNA extracted for shotgun metagenomic sequencing. Orange dashed lines indicate the optimal fragment size range for the Hi-C libraries (between 300-1,000 bp).

4.2.2 Shallow sequencing to check the Hi-C libraries

Shallow sequencing of the Hi-C libraries was performed to assess the quality before carrying out full sequencing. Over 10 million raw reads were generated for each Hi-C library. After processing the reads, 67.2%, 86.5%, 82.1%, and 87.0% of reads remained for H1, H2, H3, H4, respectively (Table A4.1). The greatest loss of reads was during the deduplication step, where 32.2% of H1 and 17.1% of H3 reads were lost, compared to 12.5% and 12.3% for H2 and H4, respectively. This difference is likely due to the extra PCR cycles that were used for H1 and H3 during the library amplification stage of the Hi-C protocol.

Taxonomic profiling of the processed reads confirmed the OB7 spike-in was present in all libraries. The taxonomic profiles were generally in line with gut microbiota, although 96.7% of H2 reads could not be classified (Figure A4.1.2). The reads were then mapped to the G_3C assembly from Chapter 3. To estimate the number of intercontig reads (read pairs where each read maps to a different contig), the proportion of reads mapping with a MAPQ >20 that were also intercontig was calculated (number of intercontig reads / total number of reads mapping with MAPQ >20). This method was first tested on the P_HiC and D_HiC datasets from Chapter 3, which showed it could be used to accurately predict intercontig reads (Table A4.2). The resulting estimations showed promising results for the Hi-C libraries, with intercontig reads estimated to be between 10.5% and 51.6% across the four samples (Table A4.3). The libraries were subsequently sent for deep sequencing.

4.2.3 Processing of deeply sequenced Hi-C and shotgun data

Between 411-476 million raw paired-end reads were generated from the shotgun libraries (Table 4.1). After processing of the raw reads, 87.5-88.2% of the shotgun reads remained, which were subsequently assembled into contigs. The number of contigs varied between 88,597-117,632 across the four samples, and the difference in the number of contigs was reflected by the total length, with H1 having 88,587 contigs and a total length of 583,763,938 bp, compared to H3 which had 177,632 contigs and a total length of 677,892,565 bp. The assembly quality was good for all samples, with similar N50s ranging between 19,125-24,951 bp (Table 4.1).

Table 4.1. Read counts and assembly statistics for the shotgun reads

Sample	H1	H2	H3	H4
Raw shotgun reads	411,149,324	411,391,722	476,680,418	428,723,690
Processed reads	361,399,114	362,992,032	417,192,014	377,396,260
% remaining reads	87.9%	88.2%	87.5%	88.0%
Total length of assembly (bp)	583,763,938	663,836,881	677,892,565	578,372,808
Number of contigs	88,587	112,108	117,632	90,812
N50	24,951	19,125	20,856	21,440
ARGs	28	34	36	21

bp = base pairs; ARGs = antimicrobial resistance genes

The taxonomic profiles of all processed reads from both the shotgun and Hi-C libraries were assessed using MetaPhlAn3 (Figure 4.2). For the shotgun reads, 49.1±3.8% (average ± standard deviation) were unclassified. The proportion of unclassified reads was greater for the Hi-C reads (77.1±12.4%), and particularly high for the Hi-C reads for sample H2 (96.2% unclassified) (Figure 4.2). Of the classified reads, the phylum Firmicutes was predominant in all samples (56.5±20.7% of Hi-C reads, 58.3±9.4% of shotgun reads on average) except for the Hi-C reads of H3, where Verrucomicrobia was the most common phylum, making up 35.2% of the reads, despite only making up 5.9% of the classified reads of the corresponding shotgun reads. Other common phyla present in the reads were Actinobacteria, Bacteroidetes, and Proteobacteria, making up 16.9±11.3%, 18.5±11.2%, and 1.9±1.2% of all classified reads, respectively. The phylum Bacteroidetes was generally more common in the classified shotgun reads (26.3±9.7%) compared to the Hi-C reads (10.8±5.9%) (P = 0.008, paired t test), and this was reflected at class-level too, where Bacteroidia had a higher relative abundance in the shotgun reads than Hi-C reads (P = 0.008, paired t test) (Figure 4.2). Similarly, the phylum Actinobacteria was more abundant in Hi-C reads (22.1±12.1%) than the shotgun reads (11.6±7.3%) (not significant, paired t test), which was also reflected at class-level (Figure 4.2). The *A. pittii* OB7 spike-in was present in all sets of reads. In the shotgun reads, OB7 was present as 0.8-1.3%, slightly higher than the estimated 0.5% spike-in. Likewise, in the Hi-C reads for sample H1, 1.1% of the reads were identified as *A. pittii*. However, this value was lower than 0.5% in the other sets of Hi-C reads, ranging from 0.02-0.2% (Figure 4.2).

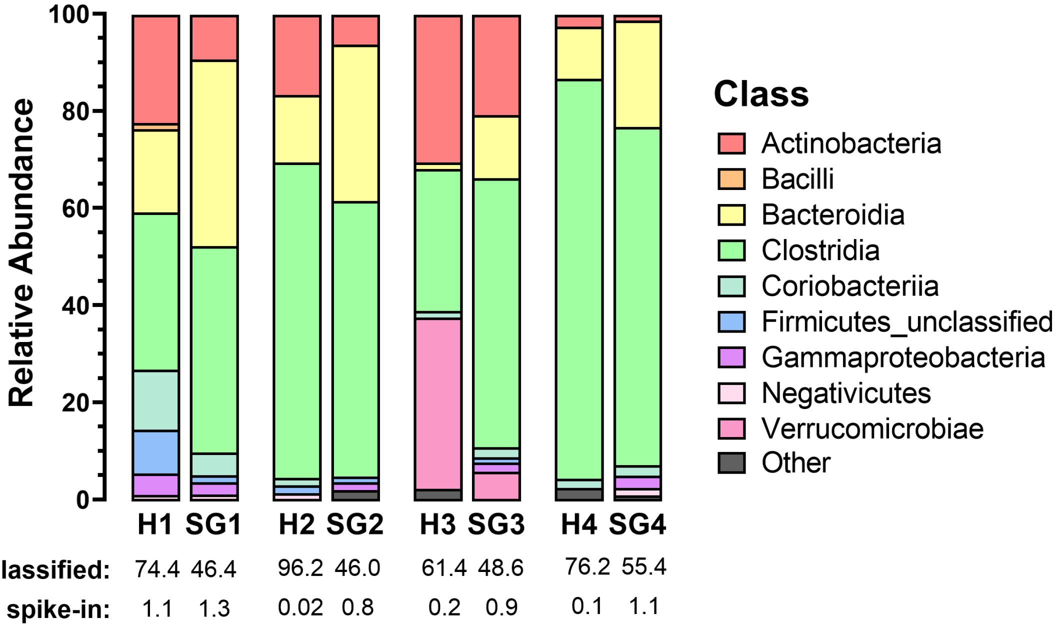


Figure 4.2. Class-level composition of Hi-C and shotgun reads.
The reads from each dataset (Hi-C (H) and corresponding shotgun reads (SG)) were taxonomically profiled using MetaPhlAn3. The stacked bars show the relative abundance (%) of each class for the classified reads. Reads that could not be classified by MetaPhlAn3 (proportion unclassified showed below bars) are excluded here. The relative abundance of the *Acinetobacter pittii* strain OB7 spike-in are shown below the bars.

ABRicate identified a total of 119 contigs containing a total of 62 unique ARGs in the shotgun metagenomic assemblies. The number of ARG-contigs in each assembly varied, with 28, 34, 36, and 21 ARGs in the shotgun metagenomic assemblies for samples H1, H2, H3, and H4, respectively (Table 4.1). Genes conferring resistance to tetracycline antibiotics were highly prevalent in the samples, with 22 different tetracycline resistance genes being found in 40 contigs in the four assemblies. Macrolide (23 contigs containing 12 different genes) and aminoglycoside (19 contigs containing 11 different genes) resistance genes were also common. The abundance (defined as RPKM) of the ARGs were calculated for both the shotgun and Hi-C reads for each dataset (Figure 4.3). The RPKM values were relatively similar between the shotgun and Hi-C reads, and all ARGs in the shotgun metagenomic assemblies had Hi-C reads mapping to them with the exception of *tetB*(P) in sample H2, which had no Hi-C reads mapping to it despite being present in the shotgun reads at 0.15 RPKM (Figure 4.3). ARGs from the OB7 spike-in were present in all sets of reads, with the chromosomal ARG *bla*$_{\text{OXA-500}}$ being present at an average of 3.22±0.31 RPKM in the shotgun reads and 2.81±3.59 RPKM in the Hi-C reads. The OB7 ARGs carried in a plasmid had a 10-fold higher relative abundance (37.28±5.52 and 26.48±32.50 RPKM for *mph*(E) and 36.89±5.34 and 47.11±59.78 RPKM for *msr*(E) for the shotgun and Hi-C reads, respectively), likely due to the plasmid these genes are present in having a higher copy number than the chromosome.

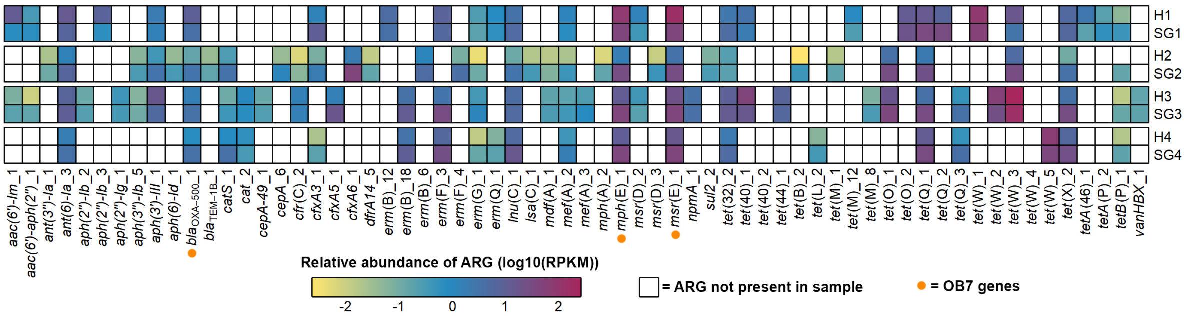


Figure 4.3. Relative abundance of antimicrobial resistance genes (ARGs) in Hi-C datasets.
The ARG sequences from the shotgun metagenomic assemblies of each dataset were isolated, and both the Hi-C (H[*]) and shotgun (SG[*]) reads from that dataset (rows) were mapped to the ARGs (columns). The relative abundance was calculated as reads per kilobase per million mapped reads (RPKM). White cells mean the ARG was not present, and coloured cells show that the ARG was present, with the colour relating to the relative abundance of the ARG within that set of reads (log(10) transformed RPKM values). ARGs highlighted with an orange dot are from the *Acinetobacter pittii* OB7 spike-in added to the samples.

Notably, *bla*$_{\text{OXA-58}}$ was not present in the metagenomic assemblies despite being carried on the same plasmid in OB7 as *mph*(E) and *msr*(E). However, mapping reads to the *bla*$_{\text{OXA-58}}$ gene revealed that both the Hi-C and shotgun data from all samples had reads mapping to the gene, with average RPKM values of 31.8±38.2 for the Hi-C reads, and 40.0±5.8 for the shotgun dataset. These values are comparable to the RPKM values for *mph*(E) and *msr*(E) on the same plasmid, therefore suggesting that *bla*$_{\text{OXA-58}}$ was still present in the OB7 plasmid during spike-in. The lack of presence of the gene in the assembly is, thus, likely due to an assembly error, where the assembler merged the full and partial copies of IS*Aba3* either side of *bla*$_{\text{OXA-58}}$ (Figure 4.4). A separate contig containing *bla*$_{\text{OXA-58}}$ (893 bp) with a partial IS*Aba3* sequence would likely have been less than 1 kbp in length and therefore would have been filtered out during the assembly process.

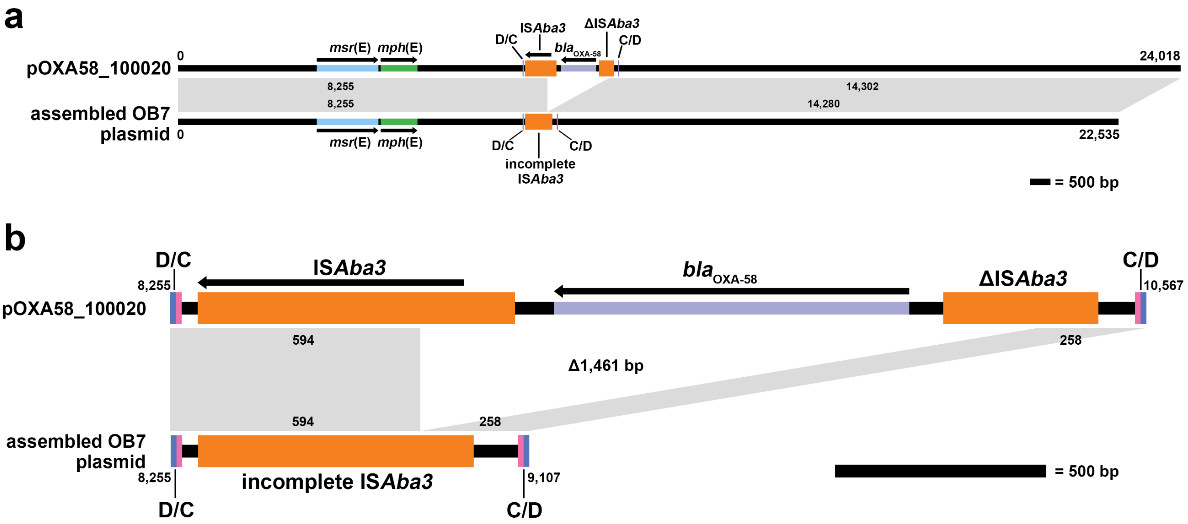


Figure 4.4. Comparison of OB7 spike-in plasmid and assembled plasmid in Hi-C datasets.

After shotgun metagenomic assembly, the assembled contig for pOXA58_100020 from the *Acinetobacter pittii* OB7 spike-in was different to the reference strain. **a)** shows the alignment of the plasmid sequences with a scale bar underneath. Numbers above and below the sequences show the length in base pairs. ARGs are highlighted, with arrows indicating the direction of the open reading frames. Grey boxes represent identical regions, with the numbers inside representing the length of the aligned region. **b)** shows bases 8,255 to 10,567 of pOXA58_100020 between the XerD/XerC (D/C) and XerC/XerD (C/D) sites, and the alignment to the assembled plasmid-contig missing bases 8,848 to 10,309.

4.2.4 Isolating intercontig reads in the Hi-C data

Between 388-478 million raw paired-end reads were generated for the Hi-C libraries (Table 4.2). During processing of the raw Hi-C reads, a significant number of reads for H1 and H3 were filtered out, leaving only 28.6% and 37.3% of the reads, respectively. The majority of reads were removed because they were duplicates, presumably caused by the additional amplification steps performed on these samples during the Hi-C procedure. Even so, as the sequencing depth was high, there was a sufficient number of processed reads left to analyse (97 million and 140 million for H1 and H3, respectively). The majority of H2 and H4 reads remained after processing, leaving over 300 million Hi-C reads each (Table 4.2).

Table 4.2. Read counts during Hi-C analysis

Sample	H1	H2	H3	H4
Raw Hi-C reads	388,159,156	405,692,574	397,350,814	478,932,328
Processed reads	111,035,162	345,912,546	150,145,886	338,806,344
% remaining	28.6%	85.3%	37.8%	70.7%
Reads mapped (MAPQ>20)	93,015,737	302,082,760	135,712,651	287,893,610
Percentage mapped	83.8%	87.3%	90.4%	85.0%
Intercontig reads	25,186,211	209,255,528	32,279,723	122,473,008
Percentage intercontig	22.7%	60.5%	21.5%	36.1%
Filtered intercontig reads	23,094,881	196,736,517	30,041,556	112,093,920
% intercontig after filtering	20.8%	56.9%	20.0%	33.1%

MAPQ = mapping quality

The first 50 bp of the Hi-C were then mapped to the corresponding metagenomic assemblies generated from the shotgun data, resulting in 83.8%, 87.3%, 90.4%, and 85.0% of reads mapping with a mapping quality score >20 for H1, H2, H3, and H4, respectively (Table 4.2). Intercontig read pairs, where both reads of the pair mapped to different contigs, were then identified. The proportions of intercontig reads from total processed reads were high (20.0-56.9%), in line with the estimations from the shallow sequencing data. Despite sample H3 having the lowest proportion of intercontig reads from these four datasets (21.5%), it had nearly double the proportion observed from the Hi-C dataset with the highest intercontig read proportion analysed in Chapter 3 (P_HiC, 13.74%). The highest proportion of intercontig reads was sample H2, with 60.5% of processed reads being intercontig (Table 4.2). In comparison, when mapping the first 50 bp of the shotgun reads to the metagenomic assemblies, only 1.5-2.0% of the reads were identified as intercontig across the four datasets.

To filter out spurious intercontig reads that mapped near the ends of contigs, the positions in the metagenomic assembly that the reads mapped to were checked for both the shotgun an Hi-C reads (Figure 4.5). The majority (65.3±4.3% average) of intercontig reads from the shotgun data (which, by definition, are spurious as they cannot be originating from cross-linked fragments of DNA) mapped within the first or last 500 nt of a contig. In comparison, the proportion of non-intercontig shotgun reads mapping near the end of contigs (5.6±0.7%) was significantly lower (P=0.0004). For the Hi-C reads, there was no significant difference between proportions of non-intercontig (5.5±1.3% average) and intercontig (7.4±1.0% average) reads which mapped near the ends of contigs (P = 0.0733) (Figure 4.5). Nonetheless, intercontig reads mapping within 500 nt of the ends of a contig were removed from further analysis, leaving 23.1 million, 196.7 million, 30.0 million, and 112.1 million intercontig reads for H1, H2, H3, and H4, respectively (Table 4.2).

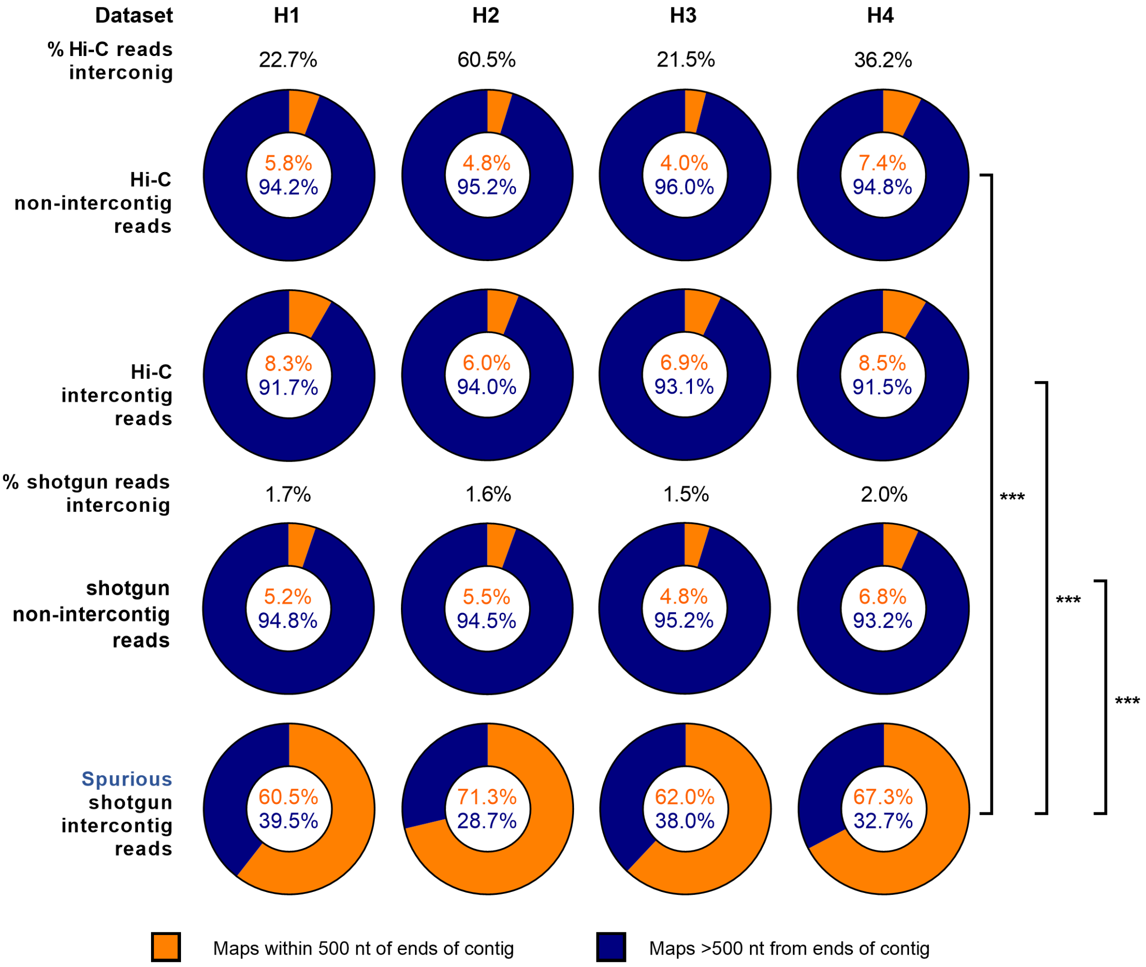


Figure 4.5. Proportions of intercontig reads mapping within the first or last 500 nucleotides (nt) of a contig in their respective assemblies.
The position of the alignment to contigs was checked for all Hi-C and shotgun reads. Orange shows the proportion of reads mapping within 500 nt of the ends of a contig. Blue shows reads mapping greater than 500 nt away from the ends of a contig. The proportions for both the intercontig reads and non-intercontig reads for the Hi-C (first two rows) and shotgun (last two rows) datasets. For each row, average proportions of reads mapping within 500 nt of the ends of a contig were compared to every other row using a one-way ANOVA with Tukey's multiple comparisons test. Rows that were significantly different are indicated on the right-hand side (*** = P < 0.001).

4.2.5 Contigs were successfully clustered into bins using the Hi-C reads

The processed Hi-C reads and metagenomic assemblies were uploaded to the ProxiMeta platform for binning and other cloud-based ProxiMeta analyses. This resulted in contigs clustering into 122-199 bins across the samples (Figure 4.6). Fewer bins were generated for samples H1 and H3 (Figure 4.6a). Likewise, a lower percentage of contigs were clustered for H1 (6.05%) and H3 (8.54%) than H2 (14.25%) and H4 (26.10%) (Figure 4.6b). The same was true for the proportion of assembly length clustered into bins, where the majority of the assembly length was clustered for H2 and H4, compared to around a third for H1 and H3 (Figure 4.6b). Bins were checked for completeness (percentage of marker genes) and contamination (marker gene overrepresentation), and were quality scored (completeness - 5 × contamination) (Figure 4.6c). Bins that had <10% contamination, >50% completeness, and >50 quality score were considered reasonable quality bins and were kept. Bins below this threshold were discarded. This left 58, 88, 71, and 97 bins for H1, H2, H3, and H4, respectively (Figure 4.6a).

The remaining bins were classified using GTDB-Tk. All samples had a bin that GTDB-Tk classified as *Acinetobacter pittii*, the spike-in that had been added prior to Hi-C, all with 98.9% completeness and 0.27% contamination. As the GTDB, the reference database for GTDB-Tk, contains MAGs and sequenced isolates that have no validly or effectively published name, some of the classifications by GTDB-Tk contain placeholder names. Therefore, all bins classified with placeholder names were relabelled as the lowest ranking validly or effectively published name given by GTDB-Tk. For example, “Lachnospiraceae CAG-95 sp000438155” was relabelled as “Lachnospiraceae”, and “*Prevotella* sp003447235” was relabelled as “*Prevotella*”. After relabelling, 240 out of 314 bins were classified to species- or genus-level across the four samples (Figure 4.6d).

4.2.5.1 Plasmid annotation

ProxiMeta also clustered contigs into plasmid bins (Figure 4.7). Samples H1 and H2 had 5 plasmid bins, and samples H3 and H4 had 7 plasmid bins each. The plasmid bins were filtered using the same completeness (percentage of reference covered by plasmid bin), contamination (percentage of plasmid bin that has mutually-conflicting alignments to reference), and quality score (completeness – 5 × contamination) thresholds as the rest of the bins, leaving between 1-2 plasmid bins for each sample (Figure 4.7). During ProxiMeta analysis, contigs in the metagenomic assemblies were also annotated as plasmids if they mapped to plasmid sequences in the NCBI RefSeq database with >80% coverage and >90% identity. In total, 200, 684, 267, and 192 contigs were annotated as plasmids for H1, H2, H3, and H4, respectively. These contigs were then filtered to remove any contigs that covered less than 80% of the reference sequence, leaving 11, 12, 11, and 7 plasmid-contigs for H1, H2, H3, H4, respectively. Contigs were also predicted to be from plasmids using ABRicate and the PlasmidFinder database, which identified plasmid replicons in 1, 11,12, and 11 contigs for H1, H2, H3 and H4, respectively. After removing any overlap from the different plasmid annotation methods, 14, 22, 29, and 12 plasmid-contigs remained for H1, H2, H3, and H4, respectively (Figure 4.8).

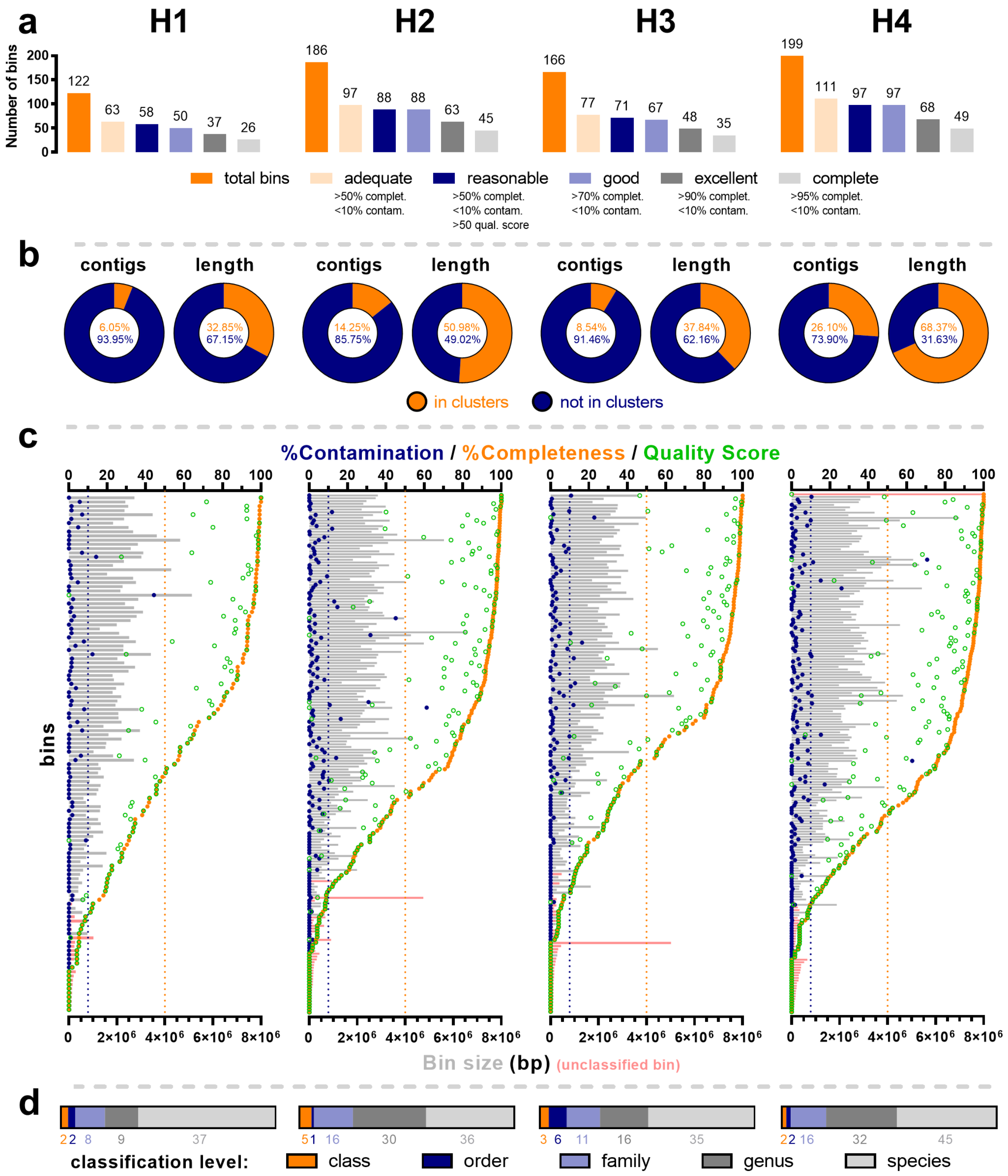


Figure 4.6. Results of ProxiMeta binning for the Hi-C datasets.
Contigs from the shotgun metagenomic assemblies were binned using the Hi-C reads. Each column shows the results for the dataset shown at the top. Figure shows **a)** the number of bins for each dataset, displaying the total number and the numbers in each category of assessed quality (qual.) from the completeness (complet.) and contamination (contam.) scores. **b)** the proportion of contigs and assembly length clustered into bins. **c)** each row represents an individual bin's contamination (% of marker gene overrepresentation, blue dot), completeness (% of marker genes, orange dot), and quality score (completeness - 5 × contamination, green circle). Bars show the bin length in base pairs (bp), with pink bars showing bins that could not be classified by GTDB-Tk. **c)** proportion of achieved GTDB-Tk classification levels for filtered bins.

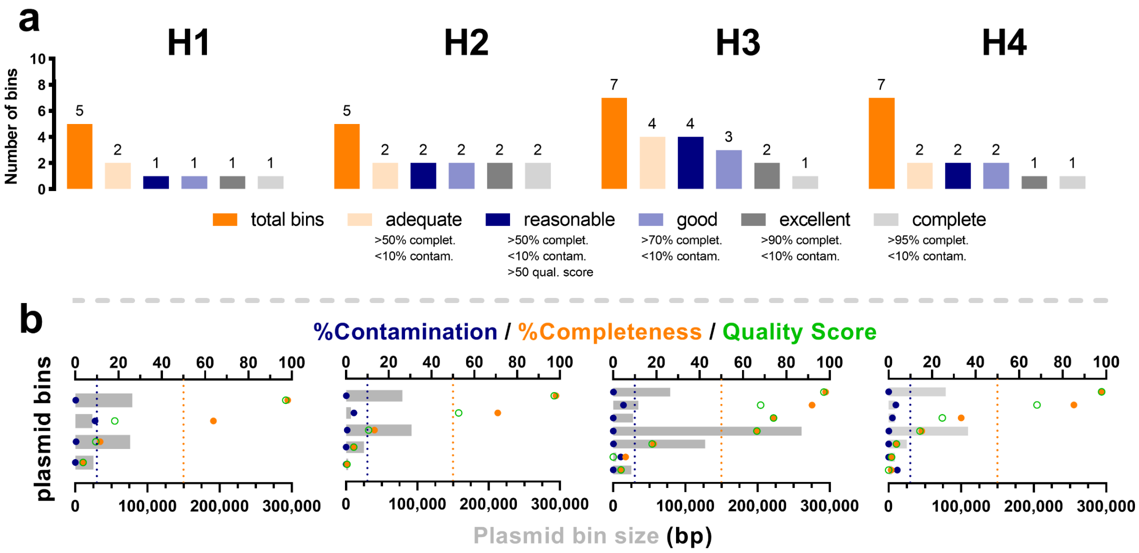


Figure 4.7. Results of ProxiMeta plasmid binning for the Hi-C datasets. Contigs from the shotgun metagenomic assemblies were binned into plasmid bins using the Hi-C reads. Each column shows the results for the dataset shown at the top. **a)** shows the number of plasmid bins for each dataset, displaying the total number and the numbers in each category of assessed quality (qual.) from the completeness (complet.) and contamination (contam.) scores. **b)** each row represents an individual plasmid bin's contamination (percentage of plasmid bin that has mutually-conflicting alignments to reference, blue dot), completeness (percentage of reference covered by plasmid bin, orange dot), and quality score (completeness - 5 × contamination, green circle). Bars show the bin length in base pairs (bp).

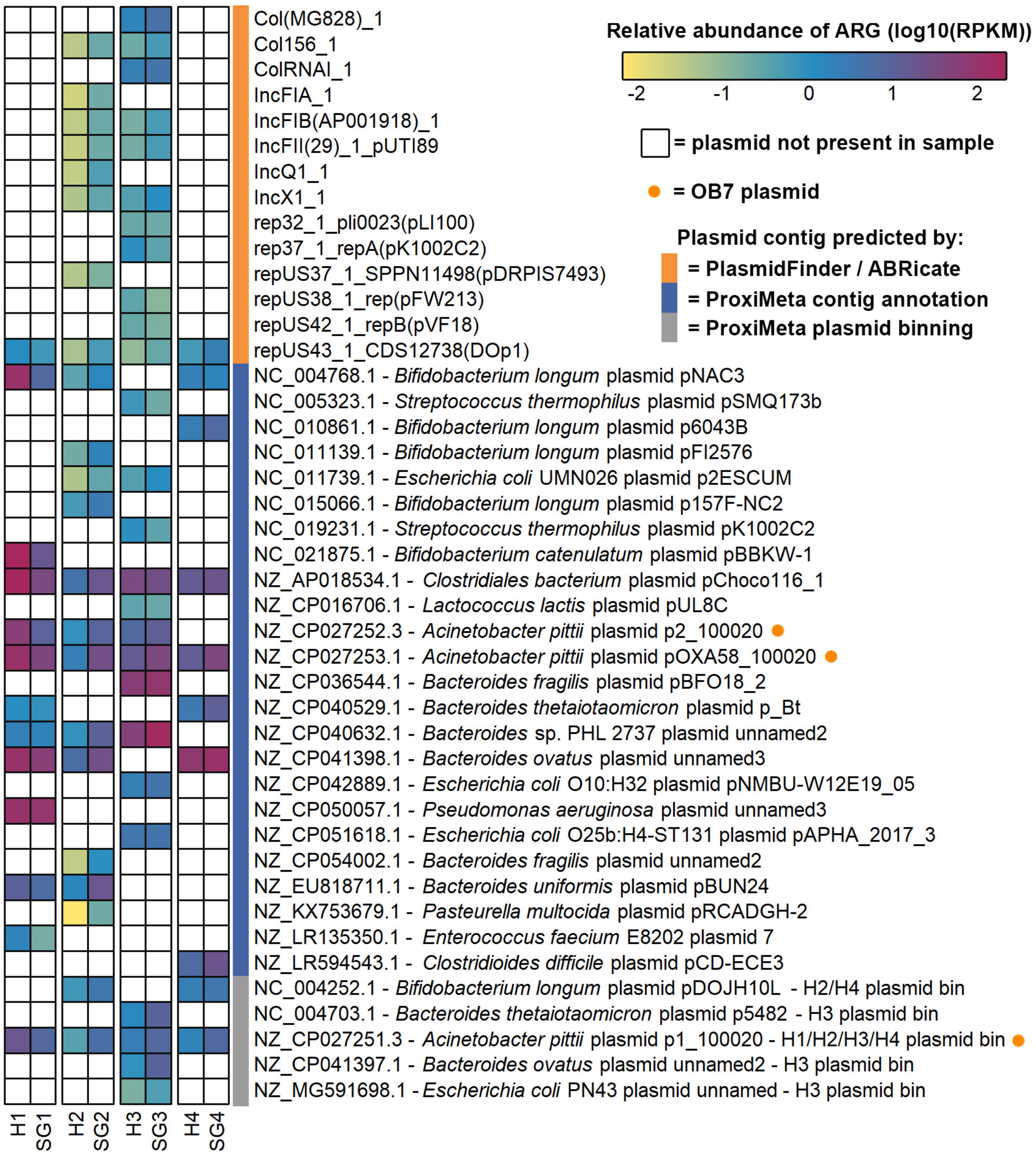


Figure 4.8. Relative abundance of plasmid-contigs in Hi-C datasets. Plasmid-contigs from the shotgun metagenomic assemblies of each dataset were predicted using the PlasmidFinder database (orange bar), ProxiMeta contig annotation (blue bar), and ProxiMeta plasmid binning (grey bar). Both the Hi-C (H[*]) and shotgun (SG[*]) reads from each dataset (columns) were mapped to the respective plasmid-contigs (rows). The relative abundance was calculated as reads per kilobase per million mapped reads (RPKM). White cells mean the plasmid-contig was not present, and coloured cells show that the plasmid-contig was present, with the colour relating to the relative abundance of the plasmid-contig within that set of reads (log(10) transformed RPKM values). For the plasmid-contigs predicted by ProxiMeta, the row label shows the NCBI accession number and title of the reference sequence that the contig or plasmid bin mapped to. Plasmids highlighted with an orange dot are from the *Acinetobacter pittii* OB7 spike-in added to the samples.

4.2.6 Linking ARGs to hosts

Before linking ARGs using the workflow developed in Chapter 3, filtered bins were screened for ARGs using ABRicate. Across the samples, 17 ARGs were found in 21 different bins (Figure 4.9).

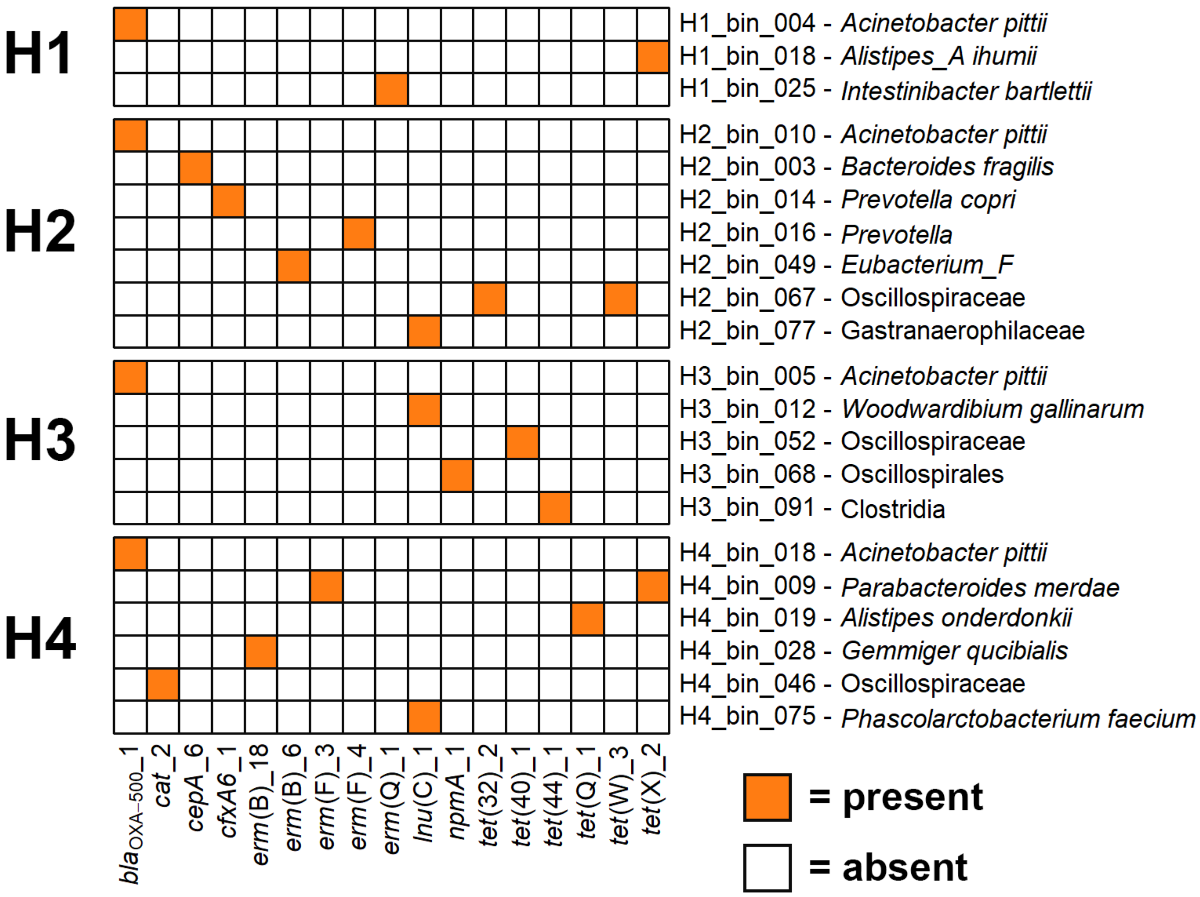


Figure 4.9. Antimicrobial resistance genes (ARGs) present in bins.
Heatmap for each dataset (labelled on left) shows the presence (orange) or absence (white) of the ARG (column) for each bin with that contained one or more ARG (rows) according to ABRicate using the ResFinder database. The row label shows the bin name and GTDB-Tk classification.

4.2.6.1 Linking ARGs to hosts using the H-LARGe workflow

To further identify host-ARG links, the Hi-C data were used to link ARG-contigs to classified bins. Filtered intercontig read pairs where one read mapped to an ARG-contig were identified using the H-LARGe workflow developed in Chapter 3. Using the same stringent approach to minimise noise from any remaining spurious intercontig reads as in Chapter 3, ARG-contigs were considered linked to another contig in the metagenomic assembly only if they were linked by at least five unique Hi-C intercontig read pairs. In addition, links to IS element contigs were also removed (Table 4.3). For the H-LARGe v2 workflow used in this chapter, links to other ARG-contigs were also removed to reduce any noise caused by contigs of similar ARGs being linked to each other.

Table 4.3. Number of contigs linking to ARG-contigs in Hi-C datasets

Sample	H1	H2	H3	H4
Filtered intercontig reads	23,094,881	196,736,517	30,041,556	112,093,920
Percentage intercontig after filtering	20.8%	56.9%	20.0%	33.1%
Reads linking contigs to an ARG-contig	174,840	46,328	134,287	255,867
Unique contigs linked to ARG-contig	2,611	5,997	4,265	18,228
Linked ≥5 times & no IS element/ARG-contig	2,128	557	1,807	2,923
Number of ARGs linked to host(s) (/ number of ARGs in sample)	27 / 28	15 / 34	28 / 36	17 / 21

ARG = antimicrobial resistance gene; IS element = insertion sequence element

Despite having the highest number of intercontig reads, sample H2 had the lowest number of reads linking contigs to an ARG-contig, and consequently only 15 out of 34 ARGs were linked to contigs from their hosts for H2. For H1, H3, and H4, the majority of ARGs were able to be linked to contigs from their hosts (Table 4.3).

Across all samples, 87 ARGs were linked to contigs from 216 different bins. The hosts of the ARGs were then classified by checking the taxonomic classification of the bin that the linked contig had been clustered into. After classification, the 87 ARGs linked to 107 different hosts (Figure 4.10). For H1 and H3, most of the linked contigs were not clustered into bins by ProxiMeta, so were not taxonomically classified. This limitation did not impact the analyses of samples H2 and H4 to the same extent, as the majority of the metagenomic assemblies had been successfully clustered into bins, although 6 out of 17 ARGs for H4 were mostly linked to bins which had been discarded due to their quality.

ARGs that were present in a bin (Figure 4.9) often linked to other contigs from the same bin, or to contigs from a bin classified as a similar host such as *tet*(X) in H1 linking both to contigs in its own *Alistipes_A ihumii* bin and contigs in an *Alistipes putredinis* bin (Figure 4.10a). As well as this, the H-LARGe workflow was, importantly, able to link unbinned ARG-contigs to contigs that were clustered into classified bins.

Excluding unbinned contigs, most ARGs were linked to a single host. Some genes, however, were linked to several hosts. For example, *erm*(B) in sample H4 was linked to *Faecousia*, *Gemmiger qucibialis*, *Negativibacillus*, and a Lachnospiraceae bin.

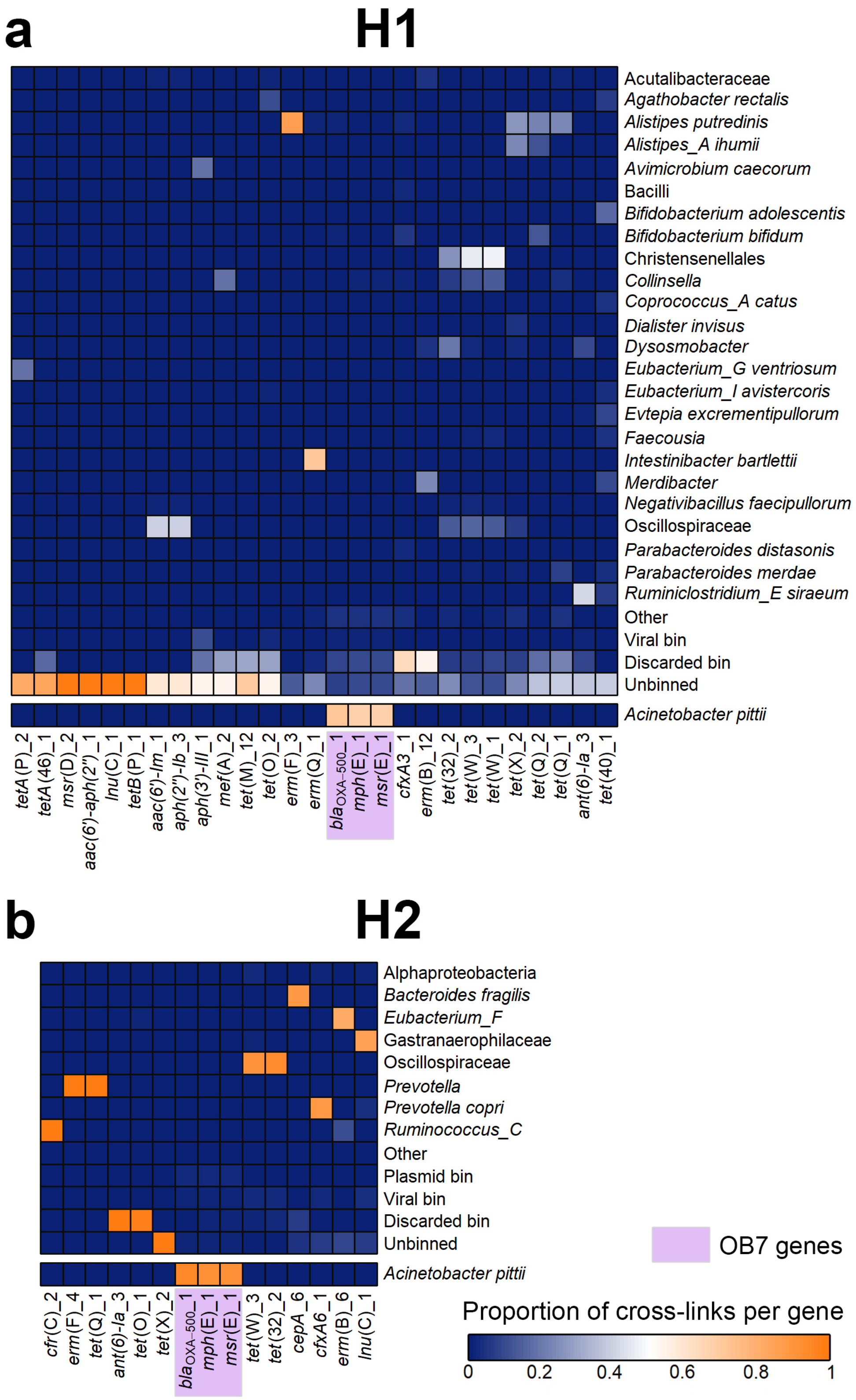


Figure 4.10 continued on next page.

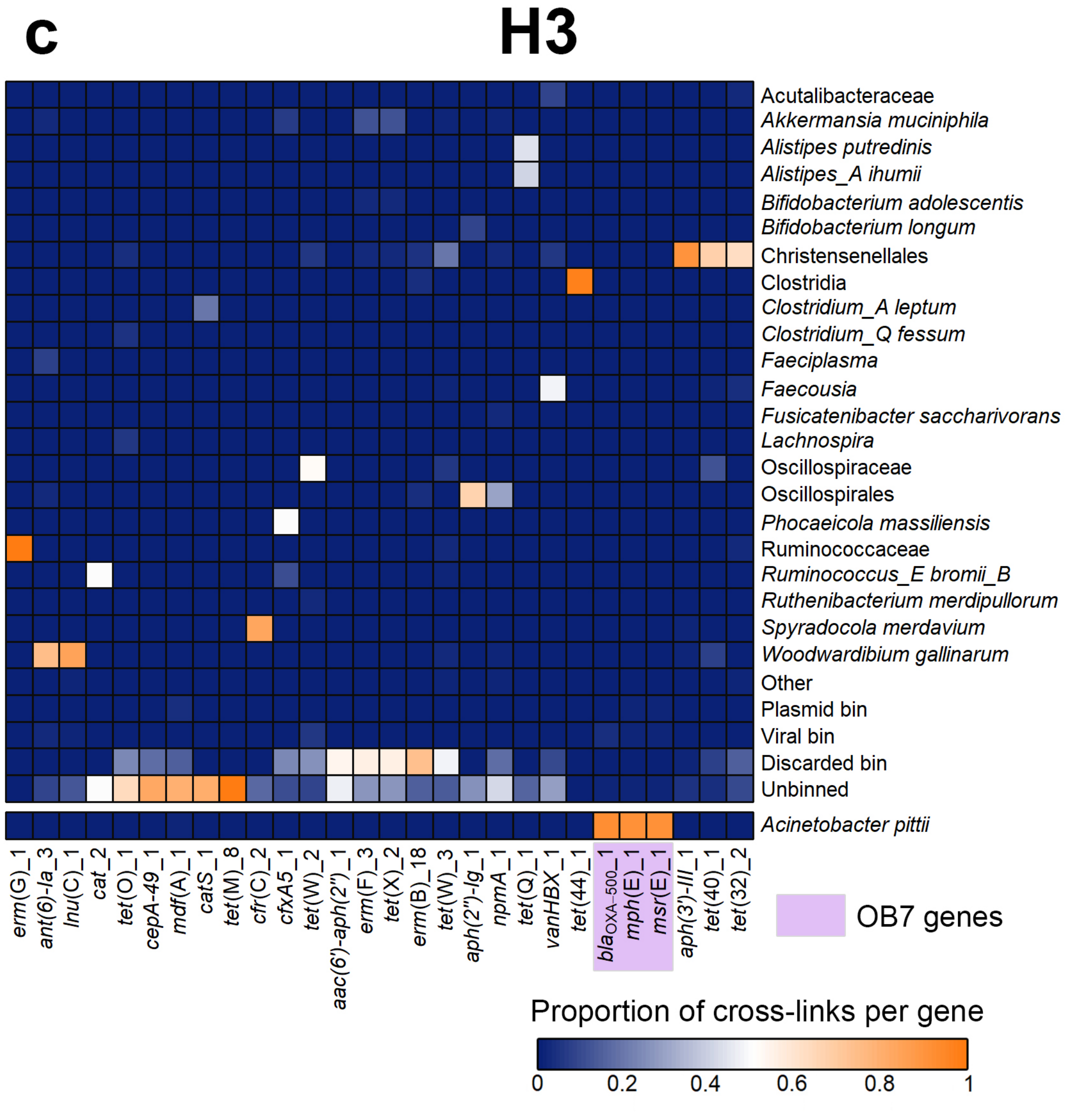


Figure 4.10 continued on next page.

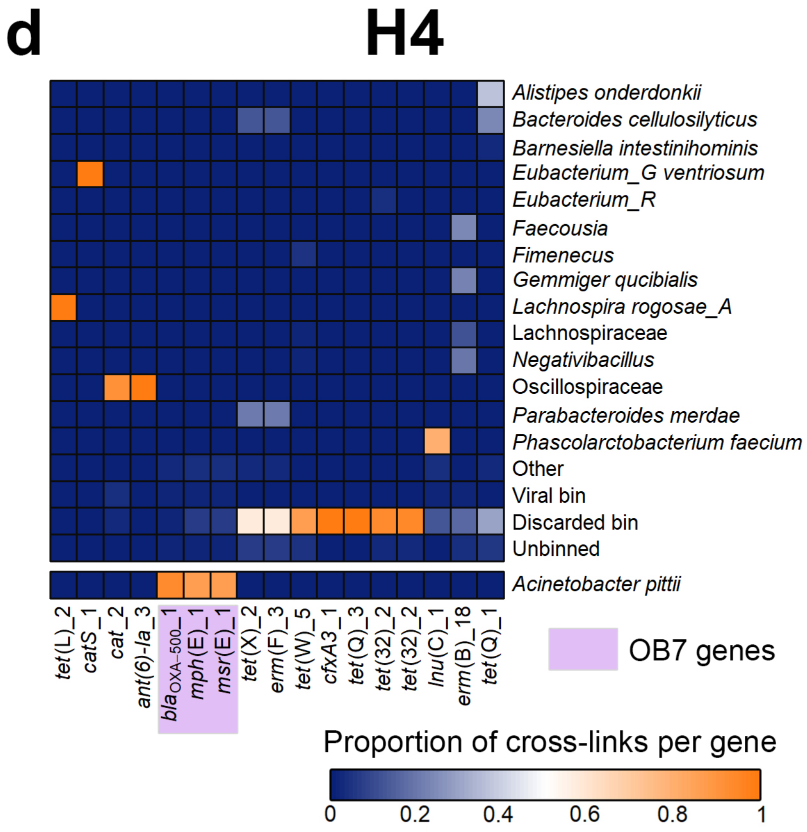


Figure 4.10. Heatmaps showing antimicrobial resistance genes (ARGs) linked with their microbial hosts.

Contigs linked to ARG-containing contigs were taxonomically classified according to the bin they were present in, which were classified by GTDB-Tk. The heatmaps show the proportion of contigs linked to each ARG that were classified as the taxon on the right for each dataset (**a)** H1, **b)** H2, **c)** H3, **d)** H4). Where there were multiple taxa that made up a proportion of no more than 0.02 for any ARG in that dataset, they have been grouped into "Other". ARG-linked contigs that were not clustered into a bin are labelled as "Unbinned". ARG-linked contigs that were clustered into a bin that was subsequently discarded due to being low-quality are labelled as "Discarded bin". ARGs highlighted in purple originate from the spike-in *Acinetobacter pittii* OB7 strain that was added to the samples prior to performing Hi-C.

4.2.6.2 Linking plasmids to hosts using the H-LARGe workflow

Plasmids were linked to their hosts using the same method (Table 4.4). In total, 64/77 plasmid-contigs were linked to a host (Figure 4.11).

Table 4.4. Number of contigs linking to plasmid-contigs in Hi-C datasets

Sample	H1	H2	H3	H4
Filtered intercontig reads	23,094,881	196,736,517	30,041,556	112,093,920
Percentage intercontig after filtering	20.8%	56.9%	20.0%	33.1%
Reads linking contigs to a plasmid-contig	271,376	32,004	42,397	126,593
Unique contigs linked to plasmid-contig	4,319	5,139	1,588	10,272
Linked ≥5 times & no IS element/plasmid-contig	3,589	361	670	982
Number of plasmid-contigs linked to host(s) (/ number of plasmid-contigs in sample)	14 / 14	12 / 22	26 / 29	12 / 12

IS element = insertion sequence element

All plasmid-contigs from the OB7 spike-in correctly linked to *A. pittii* in all samples (Figure 4.11). Similar to the ARG-host links, the majority of plasmid-links in samples H1 and H3 were to unbinned contigs or discarded bins (Figure 4.11a,c). Where plasmids were linked to contigs present in bins, the classification of the bin aligned with the species or genus that the plasmid was from according to its NCBI reference sequence. For example, *Bifidobacterium longum* plasmids in samples H1, H2, and H4 all linked to *Bifidobacterium longum* bins (Figure 4.11a,b,d). Likewise, two *Streptococcus thermophilus* plasmids in sample H3 both linked to a *Streptococcus thermophilus* bin (Figure 4.11c).

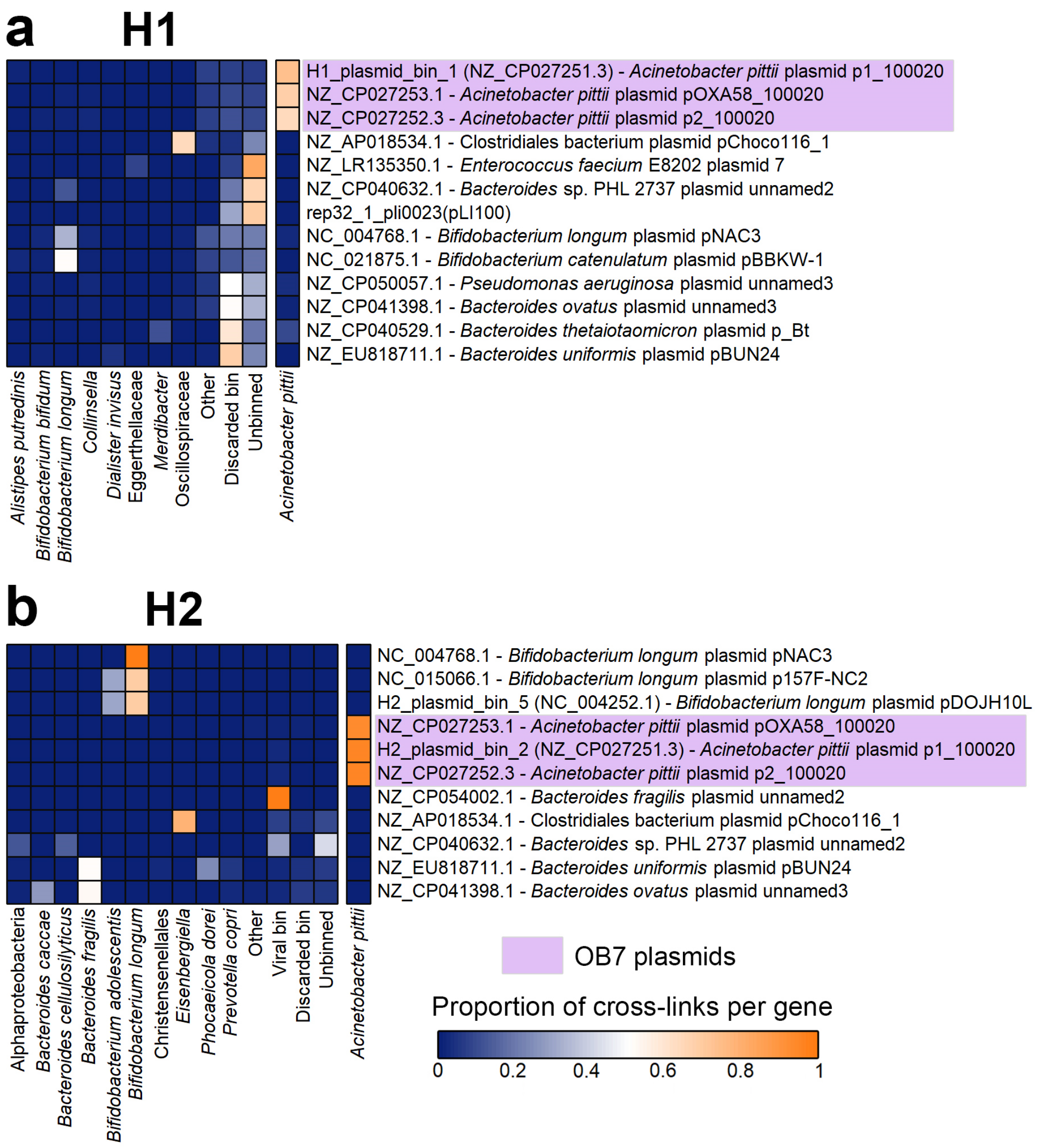


Figure 4.11 continued on next page.

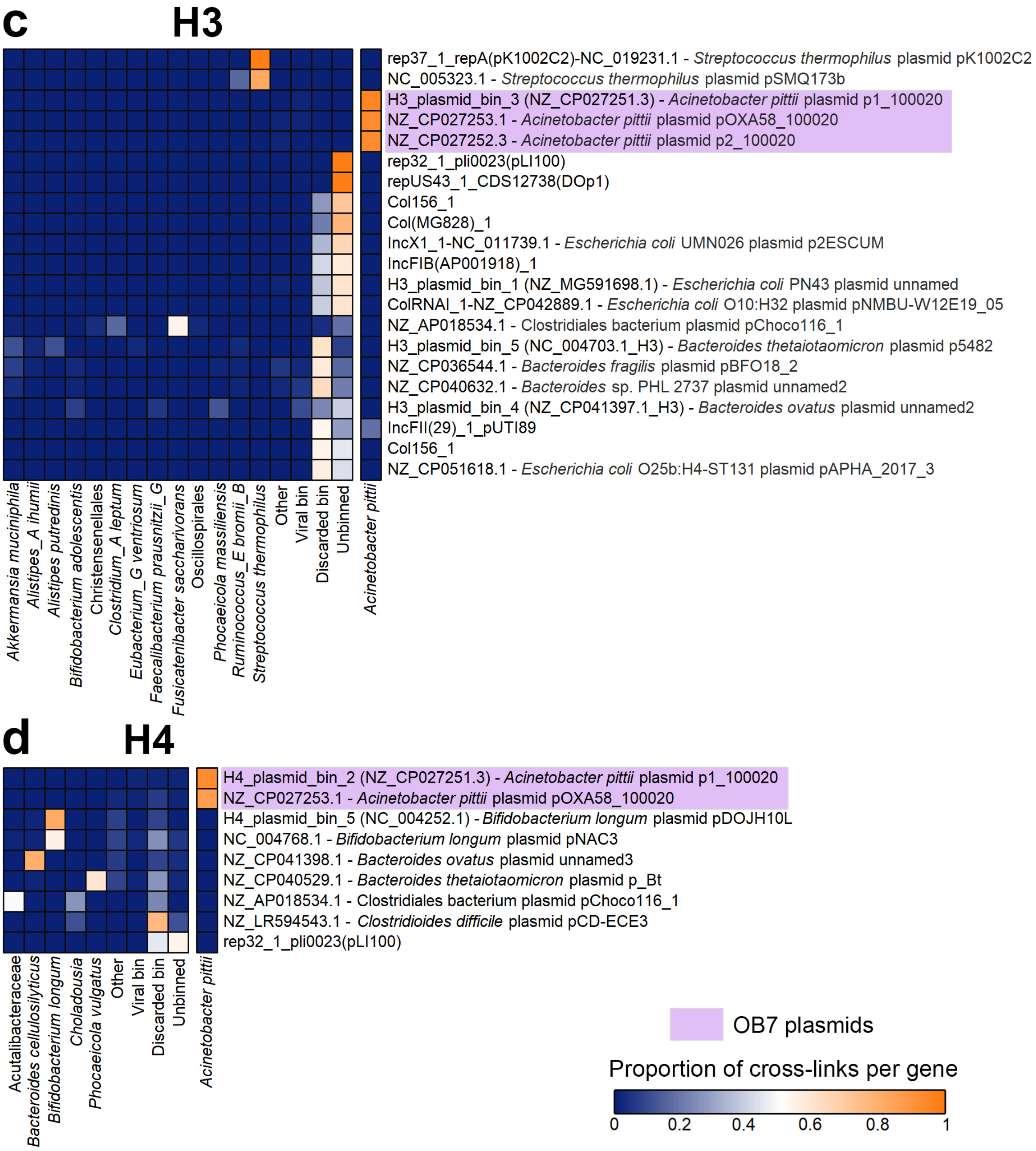


Figure 4.11. Heatmaps showing plasmids linked with their microbial hosts. Contigs linked to plasmid-contigs from each dataset (**a)** H1, **b)** H2, **c)** H3, **d)** H4), were taxonomically classified according to the bin they were present in, which were classified by GTDB-Tk. The labels to the right of the heatmaps show the NCBI accession number and title of the reference sequence that the contig or plasmid bin mapped to. The heatmaps show the proportion of contigs linked to each plasmid that were classified as the taxon labelled under the heatmap. Where there were multiple taxa that made up a proportion of no more than 0.02 for any plasmid in that dataset, they have been grouped into “Other”. Plasmid-linked contigs that were not clustered into a bin are labelled as “Unbinned”. Plasmid-linked contigs that were clustered into a bin that was subsequently discarded due to being low-quality are labelled as “Discarded bin”. Plasmids highlighted in purple originate from the spike-in *Acinetobacter pittii* OB7 strain that was added to the samples prior to performing Hi-C.

4.3 Discussion

Previous 3C-based studies on the human gut microbiome have not focused on linking ARGs to their microbial hosts (Press *et al.*, 2017; DeMaere *et al.*, 2020; Yaffe and Relman, 2020; Marbouty *et al.*, 2021). The work in this chapter aimed to implement Hi-C on four faecal samples and optimise the H-LARGe workflow to incorporate metagenomic binning to improve the accuracy of taxonomic classification.

The H-LARGe v2 workflow added additional filtering steps and integrated processing of metagenomic bins into the original H-LARGe workflow developed in Chapter 3 (highlighted in yellow, Figure 4.12). The results demonstrate that Hi-C paired with the optimised H-LARGe v2 workflow can link ARGs to their hosts, and indicate that commensals in the gut are a reservoir of ARGs. The Hi-C datasets contained more reads originating from cross-linked fragments of DNA compared to the meta3C protocol performed in Chapter 3, reducing the impact of any noise from spurious intercontig reads. Across the four samples, the H-LARGe v2 workflow was able to link 87 ARG-contigs and 22 plasmids to contigs from their microbial host, 45.7% (3,545/7,456 linked contigs) of which were in classified bins. Whilst this workflow could link ARGs and plasmids to contigs from their host, the ability to identify the host is limited by the ability to cluster the majority of contigs from a metagenomic assembly into high-quality, complete bins with a reliable, taxonomically valid, classification.

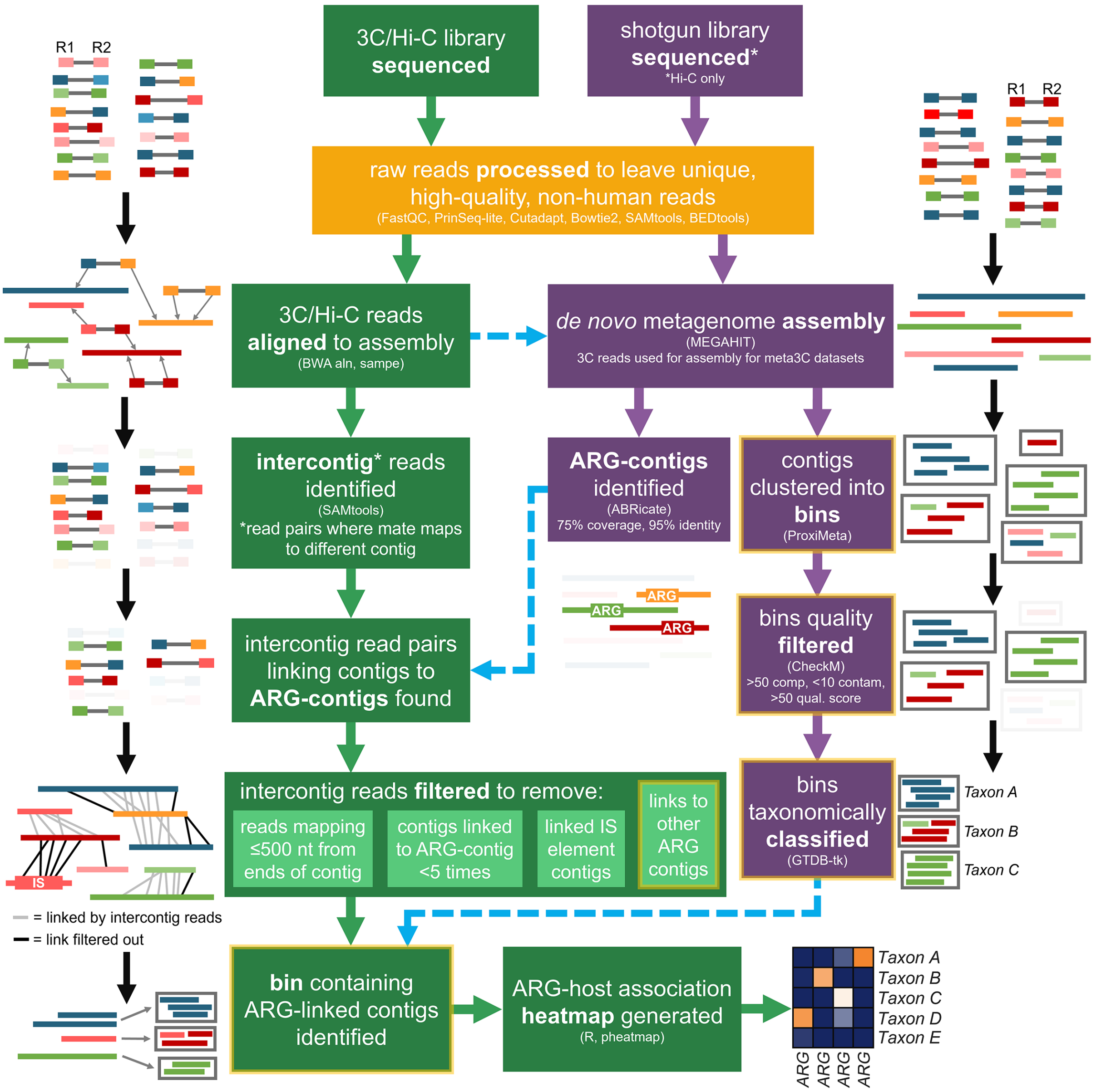

R1 R2
3C/Hi-C library sequenced
shotgun library sequenced*
*Hi-C only
R1 R2
raw reads processed to leave unique, high-quality, non-human reads
(FastQC, PrinSeq-lite, Cutadapt, Bowtie2, SAMtools, BEDtools)
3C/Hi-C reads aligned to assembly
(BWA aln, sampe)
de novo metagenome assembly
(MEGAHIT)
3C reads used for assembly for meta3C datasets
intercontig* reads identified
(SAMtools)
*read pairs where mate maps to different contig
ARG-contigs identified
(ABRicate)
75% coverage, 95% identity
contigs clustered into bins
(ProxiMeta)
ARG
ARG
ARG
bins quality filtered
(CheckM)
>50 comp, <10 contam, >50 qual. score
intercontig read pairs linking contigs to ARG-contigs found
intercontig reads filtered to remove:
reads mapping ≤500 nt from ends of contig
contigs linked to ARG-contig <5 times
linked IS element contigs
links to other ARG contigs
bins taxonomically classified
(GTDB-tk)
Taxon A
Taxon B
Taxon C
IS
= linked by intercontig reads
= link filtered out
bin containing ARG-linked contigs identified
ARG-host association heatmap generated
(R, pheatmap)
Taxon A
Taxon B
Taxon C
Taxon D
Taxon E
ARG
ARG
ARG
ARG

Figure 4.12. Overview of the H-LARGe version 2 workflow.
The flowchart shows the major stages of version 2 of the H-LARGe (Host-Linkage to Antibiotic Resistance Genes) workflow. Programs used for each stage are shown in brackets in each box. Stages using the 3C/Hi-C reads are in green boxes, and those using shotgun metagenomic reads are shown in purple. Blue dashed arrows indicate where stages of the workflow require parts of another stage. A schematic representation of each stage is shown outside of the boxes (R1 and R2 = read 1 and read 2, respectively). Boxes highlighted in yellow are stages that are new in version 2 of the workflow.

4.3.1 Implementation of ProxiMeta Hi-C on human faecal samples

The ProxiMeta Hi-C protocol was more streamlined than the meta3C protocol performed in Chapter 3 (Foutel-Rodier *et al.*, 2018). The only issue encountered that affected final library yields was during the cell lysis stage of the protocol, which involves vortexing the sample with 0.1 mm glass beads followed by the addition of magnetic DNA recovery beads to the lysate. Particularly for samples H1 and H3, the DNA-bound recovery beads failed to strongly adhere to the magnet during the recovery steps due to a viscous slime-like substance surrounding the beads. Subsequent vortexing, diluting, and washing of the beads allowed some beads to be recovered. However, the final yields of these Hi-C libraries were low, and additional amplification cycles had to be performed before sequencing to account for this, which consequently led to most reads being removed during the deduplication step when processing the Hi-C reads. The viscous substance surrounding the beads may have been due to extracellular glycoproteins produced by bacteria present in these samples. Extracellular glycoproteins can be problematic during DNA extraction of stool samples as they can inhibit DNA modifying enzymes, and extraction methods with additional glycan degradation steps have been tried to minimise this issue (Angelakis *et al.*, 2016). The glycan degradation steps used in these extraction methods involves boiling the lysate (Angelakis *et al.*, 2016). Whilst this is effective for glycan degradation and may have helped here, boiling the lysate at this step could denature DNA-bound proteins and thus remove cross-links, and therefore is not feasible during the Hi-C protocol. In this case, although not ideal, vortexing and pipetting of the magnetic beads was enough to shear some of the viscous glycoprotein to recover many of the beads.

Shallow sequencing of the Hi-C libraries allowed accurate prediction of the proportion of intercontig reads in each sample before samples were deeply sequenced, and this should be continued in future Hi-C work to minimise the significant expenditure of low-quality Hi-C libraries.

A strain of *Acinetobacter pittii* (GenBank accession GCA_002999215.3) was selected as a spike-in due to the presence of ARGs on both its chromosome and plasmids, and the low prevalence of *Acinetobacter* spp. in the gut of healthy individuals. During a study sampling faecal samples from 100 random healthy individuals from the UK, only 3% contained any species of *Acinetobacter*, none of which were *A. pittii* (Dijkshoorn *et al.*, 2005). The same study found that 30/126 healthy faecal samples from the Netherlands contained *Acinetobacter* spp., although, again, none of these samples contained *A. pittii* (Dijkshoorn *et al.*, 2005). Due to this low prevalence, *A. pittii* was a better option than the *E. coli* and *E. faecium* strains used as a spike-in in Chapter 3, which are both highly prevalent in the gut of healthy adults (Layton *et al.*, 2010; Tenaillon *et al.*, 2010).

4.3.2 Analysis of sequencing data

As seen in the published 3C/Hi-C datasets analysed in Chapter 3, most (77.1±12.4% average ± standard deviation) of the Hi-C reads were unclassified by MetaPhlAn3. Fewer reads were unclassified for the shotgun data (49.1±3.8% unclassified), and this value was more in line with the average mappability of human stool metagenomic reads when aligned to 80,990 microbial reference genomes (around 60% mapped) (Pasolli *et al.*, 2019). The classified reads had taxonomic profiles in line with gut microbiota. The most common phyla were Firmicutes, Actinobacteria, Bacteroidetes, Proteobacteria, and Verrucomicrobia, which are the phyla known to dominate the gut microbiota (Arumugam *et al.*, 2011; Rinninella *et al.*, 2019).

The shotgun metagenomic reads from the four faecal samples were assembled into high-quality assemblies. The N50s for these samples (21.5±2.1 kbp) were greater than the shotgun reads for nearly all datasets analysed in Chapter 3 (16.0±5.7 kbp). This can be attributed to the high sequencing depth of the shotgun metagenomic sequencing, resulting in an average of 379.7±22.5 million 150 bp reads remaining after pre-processing for each sample. The only assembly in Chapter 3 with a higher N50 was Y_SG_A (22.5 kbp), which was also the only sample to have higher sequencing depth (410.3 million 160 bp reads) than the four samples in this chapter.

The ARG-profiles of the four samples studied here were also in line with the human gut resistome. All four samples contained multiple tetracycline, aminoglycoside, and macrolide resistance genes, which are highly abundant in healthy faecal samples (Feng *et al.*, 2018; Ho *et al.*, 2020). The samples also all contained *cfxA* β-lactam resistance genes, which are prevalent in *Bacteroides* spp. (Veloo *et al.*, 2019), a common commensal species in the human gut microbiota, and known to have high abundance in the healthy human gut microbiota (Feng *et al.*, 2018).

4.3.3 Optimisation of H-LARGe workflow

Using the H-LARGe workflow developed in Chapter 3, Hi-C reads originating from cross-linked fragments of DNA were identified by isolating intercontig read pairs after mapping the Hi-C reads to their corresponding shotgun metagenomic assembly. The Hi-C samples all had high proportions of intercontig reads, with the lowest being 20.0% in sample H3, nearly double the highest proportion from the published datasets analysed in Chapter 3 (13.7% in P_HiC). In Chapter 3, meta3C was performed on sample H1 (named dataset G_3C in Chapter 3) and achieved 1.7% intercontig reads. Here, Hi-C was performed on the same sample (H1), which resulted in a more than 10-fold increase in proportion of intercontig reads (20.8%). This was unsurprising, as the Hi-C protocol has an additional step compared to meta3C to enrich for ligated fragments of DNA prior to sequencing (Lieberman-Aiden *et al.*, 2009; Marbouty *et al.*, 2014). The proportion of intercontig reads in samples H1-H4 were considerably higher than the published datasets studied in Chapter 3 that also used the ProxiMeta kit. This is likely due to further optimisation of the proprietary ProxiMeta kit since the P_HiC (Press *et al.*, 2017) and D_HiC (DeMaere *et al.*, 2020) datasets were published.

A major part of the H-LARGe workflow developed in Chapter 3 was the identification and filtering of spurious intercontig reads to reduce the impact of problematic noise in the Hi-C data. Compared to datasets analysed in Chapter 3, the Hi-C data from the samples studied in this chapter had a smaller proportion of spurious intercontig reads mapping within 500 nt of the ends of a contig. This is likely a consequence of having a much greater proportion of intercontig reads making up the Hi-C reads, meaning that the noise from the spurious intercontig reads was masked by reads that truly originated from cross-linked fragments of DNA.

4.3.3.1 Implementation of binning into the H-LARGe workflow

One of the main limitations identified in Chapter 3 of using 3C/Hi-C to link ARGs to their hosts was the taxonomic classification of contigs linked to ARGs. To try and address this limitation, the H-LARGe workflow was adapted in this chapter to implement binning of the contigs. During binning, contigs from the same or closely related organisms are clustered into groups referred to as bins (Sangwan, Xia, and

Gilbert, 2016). By clustering the assembled contigs into bins, the taxonomic classifications are based off groups of contigs rather than single contigs (Sangwan, Xia, and Gilbert, 2016), which, in theory, should increase the reliability of the classifications. This is particularly true when attempting to classify single contigs made up of sequences from highly conserved regions or regions recently acquired via HGT (Von Meijenfeldt *et al.*, 2019), compared to classifying a bin those contigs have been clustered into. To cluster the contigs into bins, the ProxiMeta platform was used. ProxiMeta maps the Hi-C reads to the shotgun metagenomic assembly to find links between contigs, which, along with proprietary post-processing steps, are used to cluster the contigs into bins (Press *et al.*, 2017). Whilst the clustering algorithm is proprietary, it is based on the MetaPhase technique developed by Burton *et al.* (2014). MetaPhase works by mapping the Hi-C reads to the assembly, and building a graph representing the number of Hi-C read pairs linking each contig together (Burton *et al.*, 2014). From this graph, the Jarvis-Patrick nearest-neighbour clustering algorithm (Jarvis and Patrick, 1973) is applied, before merging the nodes of the graph (representing individual contigs) using hierarchical agglomerative clustering (Burton *et al.*, 2014). This nearest-neighbour clustering approach was able to accurately deconvolve two synthetic metagenomic samples (Burton *et al.*, 2014), and ProxiMeta has been applied to several metagenomic samples with promising results, yielding high-quality bins in human faecal samples (Press *et al.*, 2017), cow rumen samples (Stewart *et al.*, 2018), wastewater samples (Stalder *et al.*, 2019), and canine faeces (Cuscó *et al.*, 2022). When applied to the samples in this chapter, ProxiMeta successfully clustered contigs into 122-199 bins for each of the four metagenomic assemblies. More bins were generated for H2 and H4 compared to H1 and H3, likely due to samples H1 and H3 having fewer processed Hi-C reads than H2 and H4, resulting in fewer intercontig reads and therefore fewer edges connecting each node when building the graph for clustering.

After binning, the quality of the bins was assessed using CheckM (Parks *et al.*, 2015). CheckM estimates the completeness and contamination of the bins based on the presence of single-copy marker genes. Completeness is evaluated by the proportion of correct marker genes presents, and contamination is estimated by the overrepresentation of these marker genes (Parks *et al.*, 2015). Using these values, a quality score was also calculated (completeness – 5 × contamination), as used in the genome quality control stages during construction of the GTDB (Parks *et al.*, 2018). Bins were then filtered using these scores to remove any bins that were less than 50% complete, contained greater than 10% contamination, and/or had a quality score less than 50. These criteria were chosen as a balance between ensuring the quality of the bins is substantial for accurate classification, whilst not discarding the vast majority of the bins. The MIMAG (Minimum Information about a Metagenome-Assembled Genome) criteria from the Genomic Standards Consortium state that >50% completeness and <10% contamination can be considered medium-quality (Bowers *et al.*, 2017), and other Hi-C studies have used these criteria to filter out low-quality bins (Yaffe and Relman, 2020; Cuscó *et al.*, 2022). The addition of the quality score here ensured that the contamination threshold was proportional to the completeness of the bin, so that bins with, for example, only 50% completeness could not have close to 10% contamination and still be considered medium-quality. These quality criteria are also employed during curation of the GTDB (Parks *et al.*, 2018).

The remaining bins were classified using the GTDB-Tk classify workflow, a popular toolkit for assigning classifications to MAGs/bins using the GTDB (Yang *et al.*, 2021). After classification, the bins were named based on the lowest ranking validly or effectively published name given by GTDB-Tk as GTDB-Tk assigns placeholder names based on NCBI accession numbers to genomes in the GTDB that do not cluster into groups with taxonomically valid named representatives (Parks *et al.*, 2018). Some genus and species names in the GTDB are sometimes labelled with an alphabetic suffix (e.g. *Ruminococcus_E bromii_B*), indicating genera or species that are polyphyletic and therefore subdivided into these alphabetic groups in the GTDB (Parks *et al.*, 2018, 2020). These suffixes were kept on the bin classification names to indicate that the bin may not be from that specific taxon, but from a closely related genus or species.

All bins were able to be classified to at least class-level, with 76.4% of the bins classified to at least genus-level, and 48.7% of the bins successfully fully classified to species-level. In a previous study, when GTDB-Tk was performed on the Unified Human Gastrointestinal Genome (UHGG) collection, comprised of 204,938 genomes from human gut bacteria, only around 30% of the gut bacteria MAGs were assigned to species-level, indicating the lack of representation of the majority of species in the human gut in the reference database (Almeida *et al.*, 2020). However, since that UHGG study, the number of reference genomes in the GTDB has almost doubled from 145,904 (Parks *et al.*, 2020) to 258,406 genomes (Parks *et al.*, 2022).

4.3.3.2 Linking ARGs to their hosts using H-LARGe v2

In total, 87 ARGs were linked to contigs from their bacterial hosts across the four samples. The taxonomic classification of the bacterial hosts was determined by the GTDB-Tk classification of the bins that the linked contigs were clustered into, resulting in 107 different classified hosts linked to the 87 ARGs.

H-LARGe v2 successfully linked the ARGs from the spike-in to the correct host, with the majority of the contigs linked to the spike-in ARGs being present in an *A. pittii* bin for each sample. However, there were spurious links to these ARGs, particularly in sample H1 where up to 32.7% of the linked contigs were not present in the *A. pittii* bin. For the other samples, the proportion of incorrectly linked contigs to the spike-in ARGs were lower, around 8.4±2.5%. In all four samples, there were no specific bins that were particularly contributing to the incorrect linking of the spike-in ARGs. In fact, the incorrect links to the spike-in ARGs seemed to be randomly distributed across many diverse bins, with the average proportion of Hi-C links to each non-*A. pittii* bin being only 0.14% of the total links to those genes (0.25% average when including links to the "Unbinned" and "Discarded bin" groups). Due to the spuriousness of these incorrect links, it is difficult to determine why they occurred. It is possible that during the Hi-C experimental protocol, fragments of free DNA cross-linked together to cause this low-level noise. Prior to preparation of the Hi-C libraries, the four stool samples had been frozen for several years without cryopreservation, which could lead to bacterial lysis and DNA degradation (Wu *et al.*, 2019; Bilinski *et al.*, 2022). Ideally, Hi-C should be performed on fresh stool samples to avoid this. Other, non-spike-in, ARGs in the samples did not have this widespread noise of low-level spurious links to many bins, so the spurious links may have occurred due to the high abundance of the spike-ins.

Indeed, sample H1 had both the highest abundance of the *A. pittii* OB7 spike-in, as well as the highest level of spurious incorrect links to the spike-in ARGs. However, due to the low-level of the noise, and the fact that no classified bin represented greater than 2% of the links to the spike-in ARGs (cut-off for appearing on the heatmaps), the spurious links were not concerning as they did not affect the overall results indicating that the host of these genes was *A. pittii*.

The non-spike-in ARGs linked to a diverse range of gut microbiota. Genes that were linked to bins were often linked to hosts that were classified to genus- or species-level. Across the samples, many ARG-host links were consistent with previous studies of antibiotic resistant isolates from the human gut microbiome. For example, *cepA* and *cfxA6* in sample H2 were linked to *Bacteroides fragilis* and *Prevotella copri*, respectively, both of which are known to frequently carry these genes (Veloo *et al.*, 2019). Previous studies have found that, unlike *cfxA* genes, the *cepA* gene is not widespread within *Bacteroides* and *Prevotella* and is only found in *Bacteroides fragilis* strains (García *et al.*, 2008; Tran, Tanaka, and Watanabe, 2013; Veloo *et al.*, 2019). Two other genes widespread within these species, *tet*(Q) and *erm*(F), were linked to *Alistipes putredinis* in sample H1, *Prevotella* in sample H2, and *Bacteroides cellulosilyticus* in sample H4. These genes are known to be present in the majority of *Bacteroides* strains due to the widespread conjugative transposon CTnDOT, which harbours both *tet*(Q) and *erm*(F) (Waters and Salyers, 2013). Whilst CTnDOT has mostly been studied in *Bacteroides* strains, it has been detected in *Prevotella* strains (Arzese, Tomasetig, and Botta, 2000; Sherrard *et al.*, 2014), and Veloo *et al.* (2019) hypothesised that the prevalence of this CTn was increasing among *Prevotella* strains after detecting *tet*(Q) in over 30% of their *Prevotella* clinical isolates. Strains of

Alistipes spp. have not been observed to harbour CTnDOT, although both *tet*(Q) and *erm*(F) have previously been detected in different species of *Alistipes* present in canine saliva (Tóth *et al.*, 2022). *Alistipes*, like *Prevotella* and *Bacteroides*, is in the order Bacteroidales, so it is likely that horizontal transfer of CTnDOT is possible between these genera.

The H-LARGe v2 workflow also revealed some novel ARG-host associations. In sample H3, *cfr*(C) was linked to *Spyradocola merdavium*, an uncultured gut bacterium in the class Clostridia. The *cfr*(C) gene has been shown to cause resistance to linezolid and phenicol antibiotics in the opportunistic gut pathogen *Clostridioides difficile* (Marín *et al.*, 2015; Stojković *et al.*, 2019), and is thought to be present in around 7% of *C. difficile* strains, often in a transposon (Candela *et al.*, 2017). In sample H2, *cfr*(C) was linked to the gut commensal *Ruminococcus_C* (a genus closely related to *Ruminococcus*), which is also in the class Clostridia. Whilst *cfr*(C) has mainly been studied in clinical isolates of opportunistic pathogens originating from the gut, the findings from this chapter indicate that *cfr*(C) could be widespread in commensal bacteria of the Clostridia class in the human gut microbiota. A recent study identified a new plasmid harbouring *cfr*(C) in clinical isolates of *C. difficile* (Chatedaki *et al.*, 2019), so it is possible that plasmid-mediated transfer of this ARG could happen in the gut, including between commensals and opportunistic pathogens. Although linezolid is not currently used for CDI treatment (van Prehn *et al.*, 2021), it has been shown to have a protective role in preventing *C. difficile*-associated diarrhoea in patients with ventilator-associated pneumonia (Valerio *et al.*, 2012). Furthermore, linezolid is used for intra-abdominal infections, which are often caused by opportunistic pathogens originating from the gut, such as species of *Clostridioides* (Candela *et al.*, 2017) and

Enterococcus (You *et al.*, 2022). Therefore, the widespread resistance to linezolid in gut commensals from the Clostridia class found in this chapter may be clinically concerning.

In all four samples, the macrolide resistance gene *erm*(B) linked to multiple hosts which were all from the Firmicutes phylum, and nearly all from the class Clostridia. In sample H1, *erm*(B) linked to bins classified as *Dysosmobacter*, a recently described gut commensal genus in the Ruminococcaceae family (Clostridia class) (Le Roy *et al.*, 2020), and *Merdibacter*, another recently described genus in the Erysipelotrichaceae family (Erysipelotrichia class) isolated from the gut (Ricaboni, Mailhe, Cadoret, *et al.*, 2017). In H2, 81.3% of the links to *erm*(B) were to *Eubacterium_F*, and 10.3% were to *Ruminococcus_C*, both from the class Clostridia. In sample H3, where *erm*(B) had linked to classified bins, the linked hosts were all in the class Clostridia, and the same was true for H4, where links to *erm*(B) were evenly distributed between four hosts. In H4, the *erm*(B) gene linked to a host from the family Lachnospiraceae, as well as hosts from the recently named genera *Faecousia* (Gilroy *et al.*, 2021) and *Negativibacillus* (Ricaboni, Mailhe, Vitton, *et al.*, 2017), and was linked to the gut commensal *Gemmiger qucibialis*. The results here show that *erm*(B) is widespread within anaerobic gut commensals from the class Clostridia. The *erm*(B) gene is prevalent in opportunistic pathogens in the gut from the Clostridia class such as *C. difficile* (Tang-Feldman *et al.*, 2005), as well as other Firmicutes opportunistic pathogens such as *Enterococcus* (Portillo *et al.*, 2000). In addition, wild-type plasmids harbouring *erm*(B) are able to transfer from *Lactobacillus plantarum* to *Enterococcus faecalis* in the gut of rats (Jacobsen *et al.*, 2007), showing the potential for transfer in the human gut from commensals to opportunistic pathogens.

Sample H1 in this chapter is the same sample that meta3C was performed on in Chapter 3. In comparison, 24 non-spike-in ARGs were linked to their host using Hi-C, nearly double the number linked by meta3C which only led to 13 ARG-host associations of non-spike-in ARGs. For the ARGs that both Hi-C and meta3C linked to their host, the host associations were similar, such as *erm*(F) and *tet*(X) linking to *Alistipes* in both methods. However, although Hi-C linked many more ARGs to contigs from their hosts, most contigs in this sample were unbinned and therefore the taxonomic classification of the linked host could not be made.

4.3.3.3 Plasmid-host associations

The majority of plasmids (64/77) were linked to contigs from their host using the H-LARGe v2 workflow. Many of the plasmids were linked to hosts with similar classifications as the species they were originally sequenced from. For example, multiple plasmids named as *Bifidobacterium* plasmids in the NCBI nt database were linked to *Bifidobacterium longum* bins in sample H1, H2, and H4. The same was true for *Bacteroides* plasmids in all four samples, which linked to bins classified as various *Bacteroides* species. A small (6 kbp) plasmid named “Clostridiales bacterium plasmid pChoco116_1” was present in all four samples and was interestingly linked to multiple hosts in several of the samples. It was mainly linked to an Oscillospiraceae bin and *Eisenbergiella*, a genus in the Clostridiales class (Amir *et al.*, 2014), in samples H1 and H2, respectively. In sample H3, pChoco116_1 was linked to *Fusicatenibacter saccharivorans*, a gut commensal in the Clostridiales class, as well as *Clostrdium_A leptum*. It also linked to multiple hosts in H4, where it was linked to bins classified as *Choladousia* and Acutalibacteraceae, both in the Clostridia class. These findings indicate that this plasmid is prevalent in the human gut microbiome and is widespread among gut bacteria in the Clostridiales class. The results of the plasmid-host linkage also show that Hi-C and the H-LARGe v2 workflow is able to successfully link plasmids to their hosts. However, as was the case for ARG-host links, links to unbinned contigs meant that the taxonomic classification of the linked host could not always be made.

4.3.3.4 Limitations

The results here show that the H-LARGe v2 workflow can successfully link the majority of ARGs to their hosts in human faecal samples. However, ARG-host association failed for 28/119 ARGs, predominantly from sample H2 were 19 ARGs were not linked to any contigs. ARGs that could not be linked to their host included the clinically relevant ARGs *bla*$_{\text{TEM-1B}}$ in H2 and *vanHBX* in H3.

In addition, the results were limited by the binning process, specifically the ability to cluster the majority of contigs into high-quality bins. Across the four samples, 54.3% of linked contigs were either unbinned or in a discarded bin. Where a plasmid or ARG was linked to unbinned contigs, or contigs present in discarded bins, the taxonomic classification of the linked host could not be made, despite the gene being linked to its host contigs. Unbinned contigs were mainly a problem in samples H1 and H3 where only 6-8.5% of contigs in the assemblies (33-38% of the total assembly length) had been clustered into bins. Contigs in discarded bins were also an issue in samples H1 and H3, but even more so in sample H4 where the majority of links in 6/17 ARGs were to contigs in discarded bins.

To improve these limitations, alternative binning methods should be tested in future studies using the H-LARGe workflow. There are several recently developed opensource binning programs for Hi-C data using various clustering algorithms. One option is MetaTOR, developed to be used on both meta3C and Hi-C datasets, which splits contigs into 1,000 bp chunks before clustering them into bins using the 3C/Hi-C reads (Baudry *et al.*, 2019). There are also recently developed workflows that, like ProxiMeta, attempt to cluster assembled contigs into bins, including bin3C (Demaere and Darling, 2019). When compared to ProxiMeta, bin3C had a 70% improvement in the number of high-quality bins obtained from a human faecal sample (Demaere and Darling, 2019). Another recently developed program, HiCBin, employs some bin3C code but uses a different clustering algorithm (Du and Sun, 2022). The HiCBin authors compared their binning results from a human faecal Hi-C dataset to both bin3C and ProxiMeta, showing an improvement on both for number of bins and number of high-quality bins obtained (Du and Sun, 2022). The recently published HAM-ART bioinformatic pipeline also developed an optimised binning algorithm (Kalmar *et al.*, 2022), which could also be considered in future studies, although this pipeline has so far not been benchmarked against the other alternatives. A package for the assembly program SPAdes called hicSPAdes is also being developed and includes a binning module for extracting MAGs from metagenomic assemblies using Hi-C reads. Although this package has not yet been fully released, Ivanova *et al.* (2022) recently demonstrated the results of a pre-release version using Hi-C data from human faecal samples, and showed that hicSPAdes was able to generate more high-quality, complete MAGs than bin3C and standard (non-Hi-C) metagenomic binning tools. The use of bin3C on the Hi-C data from this chapter was briefly explored, however the program failed to run. Nevertheless, future Hi-C studies should consider using these programs to improve the binning stage of the H-LARGe workflow.

Another method for improving the binning process is using long-read sequencing to improve the length of contigs in the assembly. The presence of regions of repeated DNA sequences is a known challenge for genome assembly using only short-reads (Koessler *et al.*, 2010). Repeats are even more problematic during metagenome assembly because as well as intragenomic repeats caused by repeats in IS elements and other mobile elements, there are also intergenomic repeats from multiple species of bacteria sharing highly conserved genomic regions (Lapidus and Korobeynikov, 2021). These repeats lead to fragmentation of the assembly, and thus cause shorter contigs. Long-read sequencing allows sequencing of reads that are able to span repetitive regions of DNA (Amarasinghe *et al.*, 2020), allowing these regions to be resolved during assembly and reducing the amount of fragmentation. By using long-read sequencing to achieve considerably longer contigs, the binning process could be improved as there would be less short contigs consisting of repetitive DNA, or contigs consisting of highly conserved regions, which are difficult or not possible to cluster into a single bin. Indeed, several recent studies have used hybrid assembly of both short- and long-reads to achieve high-quality binning in both canine faeces (Cuscó *et al.*, 2021), and human faecal samples, where 475 high-quality MAGs were assembled across 12 faecal samples, including 44 complete, circularised MAGs (Jin *et al.*, 2022). Hybrid assembly is desirable as long-read sequencing currently has a significantly higher error rate compared to short-read Illumina sequencing (M. Jain *et al.*, 2018). However, as the technology improves for long-read sequencing, hybrid assembly may no longer be needed in the near future. Long-read sequencing coupled with Hi-C data has been used to improve binning for cow rumen (Stewart *et al.*, 2018), canine faeces (Cuscó *et al.*, 2022), and sheep faeces (Bickhart *et al.*, 2022). However, this sequencing combination has so far not been used for human faecal samples. Future studies using the H-LARGe workflow should consider using long-read metagenomic sequencing to improve the binning process and allow further improvement of the host classification process.

4.3.4 Conclusions

Overall, the results in this chapter showed that ARGs are widespread in commensal bacteria, including genes that are prevalent in clinical isolates of opportunistic pathogens. The H-LARGe v2 workflow was further developed to improve host classification, however, was limited by the binning process, and future studies should work to further optimise and refine this stage of the workflow to increase the amount of ARG-host associations. Additionally, whilst the majority of ARGs were linked to their hosts, there were still 28 ARGs that were not able to be linked to their host. To fully characterise the human gut resistome, Hi-C data should be coupled with culturing to identify the hosts of the unlinked ARGs and provide insights into the genomic context of the ARGs linked to their hosts by Hi-C.

CHAPTER 5

CULTURING THE HOSTS OF ANTIBIOTIC RESISTANCE GENES

5.1 Introduction

In the previous results chapters, I implemented meta3C/Hi-C on human faecal samples and developed the H-LARGe bioinformatic workflow to link ARGs to their microbial hosts. The findings of Chapter 4 revealed that ARGs are widespread in commensal bacteria. However, not all ARGs in the metagenomic sample were able to be linked to a host, showing that Hi-C alone is unable to fully characterise the resistome in a sample.

Hi-C is often praised for being a culture-free approach (Press *et al.*, 2017; Demaere and Darling, 2019; Yaffe and Relman, 2020; Kalmar *et al.*, 2022). However, in this chapter, I aimed to explore whether culture-based approaches could be coupled with Hi-C to expand the results and further characterise the resistome. Isolating and sequencing the hosts of ARGs in the same faecal samples that Hi-C was performed on would also allow for better understanding of the genomic context of the ARGs, as well as allowing validation of the Hi-C ARG-host associations.

No previous Hi-C studies have coupled the technique with culturing. Therefore, the work in this chapter aims to be the first to validate Hi-C results using culturing. Several studies have aimed to culture and isolate many organisms from stool samples (Browne *et al.*, 2016; Lagier *et al.*, 2016), and have successfully cultured a diverse range of bacteria that were, until recently, thought to be "unculturable". As the gut microbiota is dominated by obligate anaerobes, I focussed on anaerobic culturing.

In this chapter, I aimed to culture the hosts of ARGs from the same faecal samples used for Hi-C in Chapter 4. By implementing WGS on cultured isolates, I aimed to validate the Hi-C results, and understand the genomic context of the ARGs.

5.2 Results

5.2.1 Enrichment of the hosts of ARGs from human faecal samples

The four stool samples (H1, H2, H3, and H4) were added to mGAM (modified Gifu Anaerobic Medium) broth supplemented with and without antibiotic to enrich for the hosts of ARGs. After overnight growth at 37°C in anaerobic conditions, DNA was extracted from the cultures, and the remainder was transferred to fresh broth supplemented with and without antibiotic to repeat the progress. Serial dilutions were also spread onto mGAM agar in order to grow single colonies. To check for ARG enrichment, qPCR was performed on the DNA extracts for each ARG present.

First, samples were enriched for tetracycline resistance genes as these were the most prevalent ARGs, with 40 *tet*-type genes across the samples. Genes were selected for colony PCR screening if they reached a relative quantity of 0.1 copies/copy of 16S rRNA gene after enrichment. Using this criteria, 13 tetracycline resistance genes were considered enriched across the four samples (4, 2, 4, and 3 in H1, H2, H3, and H4, respectively) (Figure 5.1). Many of these enriched *tet* genes were also present at high relative quantities in the DNA extracts from the stool cultures that contained no antibiotics, indicating the high abundance of tetracycline resistance in the samples. The *tet*(Q) gene was enriched in all samples, showing the high abundance and prevalence of this gene in human faecal samples. Some *tet* genes such as *tetA*(P) and *tetB*(P) in sample H1 decreased in relative quantity compared to the no antibiotic controls (Figure 5.1). Strains carrying these *tet*(P)-type genes often show a minimum inhibitory concentration of around 8 µg/mL tetracycline (Vidor *et al.*, 2019), which is the concentration used to enrich for *tet* genes here, so it is likely that the hosts of these genes could not grow in the antibiotic-supplemented broth.

Erythromycin was used to enrich for macrolide resistance genes. Fewer macrolide resistance genes were enriched for (6/27), compared to the tetracycline resistance genes, across the samples (Figure 5.1). The *erm*(F) gene was enriched in all samples except for H3 where the relative quantity decreased compared to the controls. However, due to the high abundance of this gene in the H3 controls (>0.1 copies/copy of 16S rRNA gene), H3 colonies were still screened for *erm*(F) in the next stage. Some genes such as *erm*(B) in sample H1 showed enrichment compared to the control (over 10-fold increase in relative quantity), however were not selected for colony PCR screening due to the low relative quantity after both 24 and 48 hours (<10^{-4} copies/copy of 16S rRNA gene) (Figure 5.1). Overall, the macrolide resistance genes showed lower relative abundance compared to the tetracycline resistance genes.

For enrichment of aminoglycoside resistance genes, both streptomycin and kanamycin were used, however no ARGs were enriched except for *aph(6)-Id* in H2 (Figure 5.1).

Ampicillin was used to enrich for β-lactamases. Most targeted (5/8) β-lactamase genes were enriched (Figure 5.1). Although there were 3 ARGs that were not enriched above 0.1 copies/copy of 16S rRNA gene (*cfxA6* in H2, *cepA-49* in H3, and *cfxA3* in H4), all β-lactamase genes were selected for colony PCR screening due to the low total number of them and their high relative abundance.

For samples H2, H3, and H4, chloramphenicol was used to enrich for chloramphenicol resistance. No chloramphenicol resistance genes were enriched for, and the relative quantities of all genes decreased after 48 hours (Figure 5.1), indicating that the host of these genes may not grow in mGAM or that the concentration of chloramphenicol used was too high to select for relevant resistance genes.

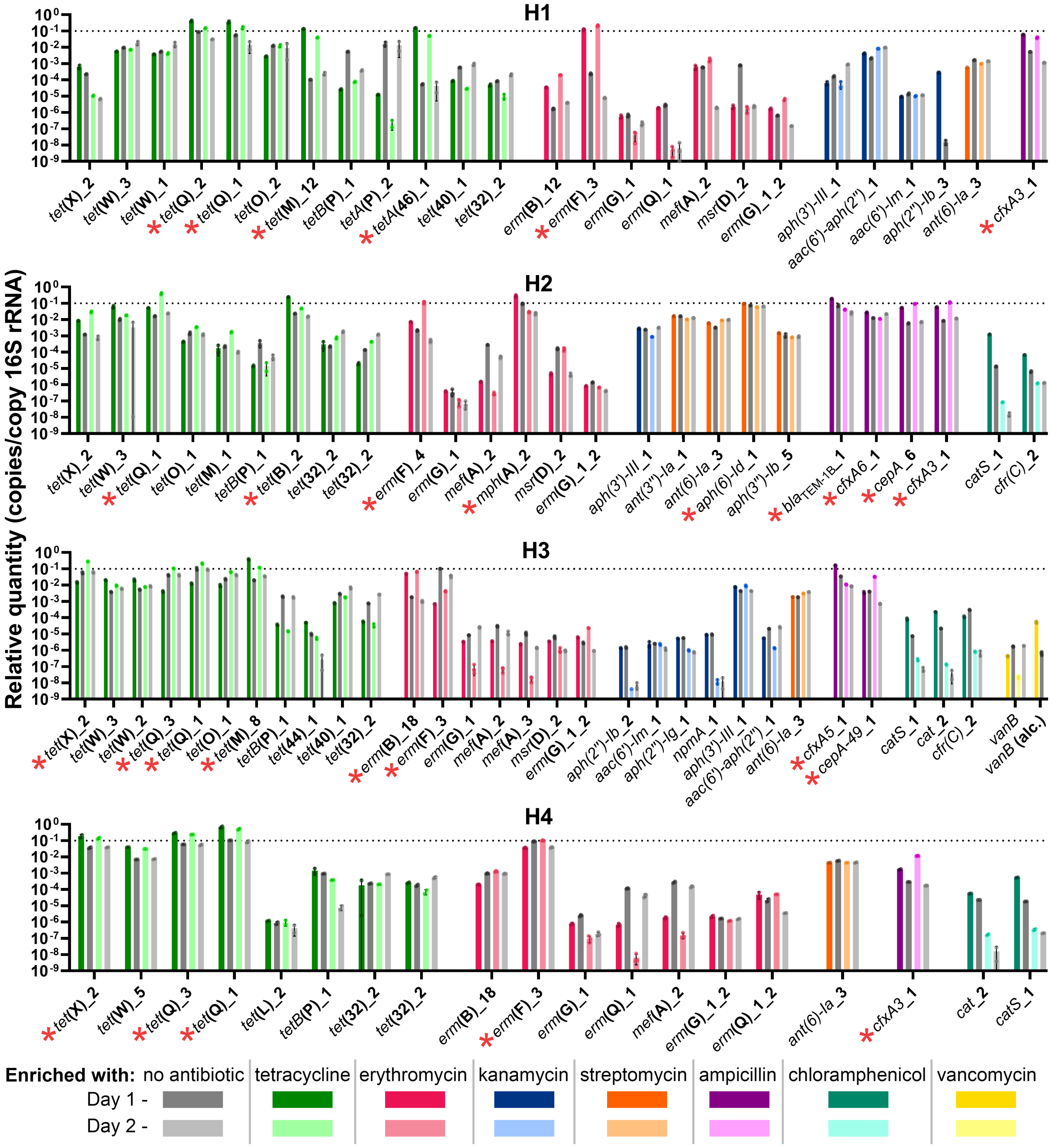


Figure 5.1. Enrichment of antimicrobial resistance genes.

Four stool samples (H1, H2, H3, H4) were added to modified Gifu Anaerobic Medium (mGAM) broth supplemented with and without antibiotic and grown at 37°C in anaerobic conditions. After 24 hours (Day 1), the enrichment process was repeated by adding the culture to fresh broth for a further 24 hours (Day 2). Quantitative polymerase chain reaction (qPCR) was performed on DNA extracted from the cultures to quantify the abundance of antimicrobial resistance genes (x-axis). The 16S rRNA gene was used as a reference gene to quantify the relative abundance against. Relative quantities of the ARGs were calculated using the formula $2^{-\Delta Ct}$, where $\Delta Ct = Ct_{ARG} - Ct_{16S\ rRNA}$. Bars show the average of duplicate or triplicate qPCR results (dots at top of bars show individual values). Error bars show the standard deviation. Enriched genes highlighted with a red asterisk (*) were selected for colony PCR on single colonies after spreading the overnight cultures on mGAM agar.

Vancomycin (in combination with colistin to kill Gram-negative bacteria in the sample) was used to try and enrich for *vanB* in sample H3, however this gene was not enriched after 24 or 48 hours (Figure 5.1). The controls showed that *vanB* was present at very low relative abundance in the sample. In case the host of *vanB* was a spore-forming strain, the stool sample was treated with an equal volume of 70% ethanol as described by Browne *et al.* (2016) to kill vegetative cells, before being added to the vancomycin-supplemented mGAM broth. Although there seemed to be some enrichment after 24 hours (Figure 5.1), the *vanB* gene was still at low relative quantity (<10^{-3} copies/copy of 16S rRNA gene), and no colonies grew on mGAM agar, so this gene was not selected for colony PCR screening.

5.2.2 Colony PCR to check for presence of enriched ARGs

After growing single colonies of the ARG-enriched faecal cultures on mGAM agar, between 16-50 colonies were randomly picked and patch-plated for each enrichment antibiotic in each sample. Colony PCR was then performed on each colony to screen for all respective selected enriched ARGs based on the qPCR results.

Colony PCR screening of the ARGs revealed that 219/353 colonies (74/98 H1, 74/104 H2, 41/79 H3, and 30/72 H4 colonies) were positive for at least one ARG. Some of these PCR positive colonies (19/219 across all four samples) only had a faint positive band following gel electrophoreses of the PCR product, compared to the strong PCR positive bands of the other colonies (Figure 5.2).

For the tetracycline-enriched cultures, 47/50, 28/32, 26/31, and 18/32 colonies were PCR positive for at least one screened *tet* gene in samples H1, H2, H3, and H4, respectively (Figure 5.2). For Sample H1, the majority (42/50) of the screened colonies were positive for *tet*(Q), which aligned with the qPCR results showing that there was a high relative quantity of *tet*(Q) in this sample after enrichment (0.38±0.05 copies/copy of 16S rRNA gene, Figure 5.1). The remaining colonies in sample H1 were nearly all PCR positive for both *tet*(M) and *tetA*(46) (Figure 5.2).

For colonies growing in erythromycin-supplemented mGAM, 12/24, 17/24, 11/24, and 9/24 screened colonies were positive for macrolide resistance genes for H1, H2, H3, and H4, respectively (Figure 5.2). Two out of 24 colonies were positive for *erm*(F) in sample H3. This was unsurprising as this gene had the lowest relative quantity in sample H3 compared to the rest of the samples (Figure 5.1).

Only sample H2 had an enriched aminoglycoside resistance gene, and of the colonies screened, 7/24 were PCR positive for *aph(6)-Id*. All samples were screened for enriched β-lactamase genes. Samples H1 and H2 had many positive colonies (16/24 and 22/24, respectively). Colonies from sample H2 were positive for *cfxA*, *bla*$_{\text{TEM-1B}}$, both *cfxA* and *cepA*, or all three ARGs (Figure 5.2). Fewer colonies were positive for β-lactamase genes in samples H3 and H4 (4/24 and 3/16, respectively). No colonies were positive for *cepA-49* in H3, and the three *cfxA*-positive colonies in sample H4 only had a faint positive band following gel electrophoresis (Figure 5.2). The low number of positive colonies for these genes was unsurprising, as both had low relative quantity after enrichment (Figure 5.1).

Colonies were then selected for classification by 16S rRNA gene sequencing based on their PCR positive ARGs and colony morphology. Overall, 54 colonies were selected for 16S rRNA gene sequencing (14 H1, 20 H2, 13 H3, and 7 H4 colonies, Figure 5.2).

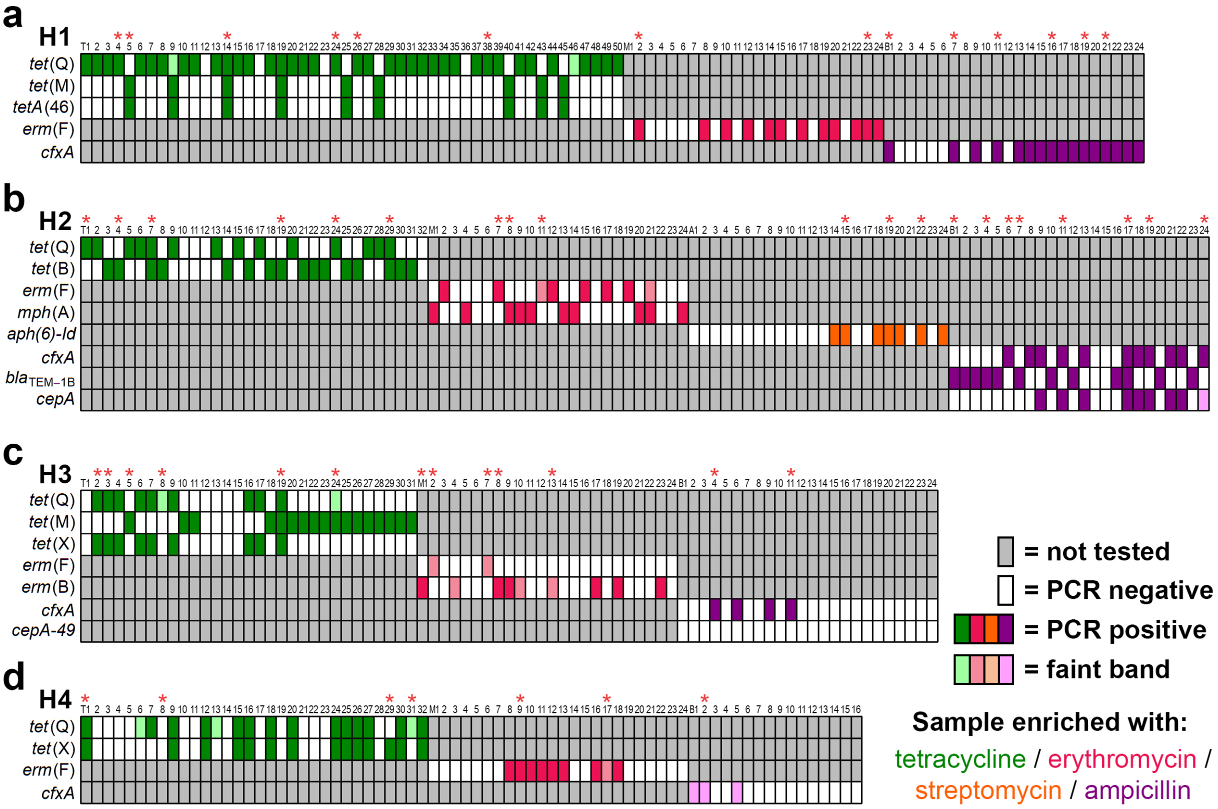


Figure 5.2. Colony PCR to check for presence of enriched antimicrobial resistance genes.

After enrichment for antimicrobial resistance genes (ARGs) in modified Gifu Anaerobic Medium (mGAM) broth, serial dilutions of the cultures were spread on mGAM agar to grow single colonies. Single colonies were screened for the presence of an ARG using colony polymerase chain reaction (PCR). Each heatmap represents a different stool sample: **a)** H1; **b)** H2; **c)** H3; **d)** H4. Each numbered column represents a different colony. Letters before the first colony of each enriched group represent the class of antibiotics enriched for (T = tetracycline; M = macrolide; A = aminoglycoside; B = beta-lactams). Each row represents a different ARG (shown on the left). Grey cells indicate the colony was not screened for the ARG of that row. White cells show colonies that were PCR negative for the ARG of that row. Coloured cells indicate colonies that were PCR positive for the ARG of that row. Colours of the cells correspond to the antibiotic they were enriched with (green = tetracycline; red = erythromycin; orange = streptomycin; purple = ampicillin). Lighter coloured cells indicate that the colony had a faint positive band following gel electrophoreses of the PCR product, whereas the darker colours indicate a strong positive band. Red asterisks (*) above colonies show colonies that were selected for 16S rRNA gene sequencing.

5.2.3 16S rRNA gene sequencing of ARG-positive colonies

Colony PCR to amplify the 16S rRNA gene (position 27 to 1,492) was performed on selected colonies. The amplicons were then sent for Sanger sequencing, and the resulting DNA sequences were aligned to the NCBI 16S rRNA gene database using BLAST.

For sample H1, the 16S rRNA gene BLAST results revealed that 13/14 of the colonies were species of *Bacteroides*. All *cfxA*-positive colonies were *Bacteroides uniformis*, and all *erm*(F)-positive colonies were *Bacteroides clarus*. A colony positive for *tet*(M) and *tetA*(46) was identified as *Streptococcus sinensis*. The colonies positive for *tet*(Q) were identified as *Bacteroides vulgatus*, including colony H1_T_14 which was also PCR-positive for *tet*(M) and *tetA*(46) (Table 5.1). All *tet*(Q)-positive colonies in sample H2 were species of *Bacteroides*. The two colonies positive for only *tet*(B) were identified as *Pseudescherichia vulneris*. Colonies positive for *mph*(A), *aph6-Id*, or *bla*$_{\text{TEM-1B}}$ were identified as *Escherichia* species. Like sample H1, all *cfxA*-positive colonies in H2 were species of *Bacteroides*. The same was true for *cfxA*-positive colonies in samples H3 and H4 (Table 5.1). The colonies in sample H3 positive for both *tet*(Q) and *tet*(X) were both *Bacteroides uniformis*. The colony with a faint-positive band for *tet*(Q) was *Dorea longicatena*, and two *tet*(M)-positive colonies were identified as *Enterococcus gallinarum*. There was also a *Bacteroides uniformis* colony containing all three *tet* genes (Table 5.1). Two *Bacteroides dorei* and *Lachnoclostridium pacaense* colonies contained *erm*(B). Finally, sample H4 had a *tet*(X) and *tet*(Q)-positive *Bacteroides cellulosilyticus* colony, as well as several other *Bacteroides* colonies containing *tet*, *erm*(F), and *cfxA* genes (Table 5.1). Following these results, 25 colonies were chosen for WGS based on the ARG PCR results and 16S rRNA gene results.

Table 5.1. 16S rRNA gene sequencing results

Colony	PCR-positive ARGs	16S rRNA gene BLAST result	COV (%)	ID (%)	WGS
H1_T_04	*tet*(Q)	*Bacteroides vulgatus*	98	99.3	
H1_T_05	*tet*(M), *tetA*(46)	*Streptococcus sinensis*	99	98.6	*
H1_T_14	*tet*(Q), *tet*(M), *tetA*(46)	*Bacteroides vulgatus*	97	100.0	*
H1_T_24	*tet*(Q)	*Bacteroides vulgatus*	98	99.3	*
H1_T_26	*tet*(Q)	*Bacteroides eggerthii*	98	99.8	
H1_T_38	*tet*(Q)	*Bacteroides vulgatus*	98	99.5	
H1_M_02	*erm*(F)	*Bacteroides clarus*	99	99.8	*
H1_M_23	*erm*(F)	*Bacteroides clarus*	99	99.4	
H1_B_01	*cfxA*	*Bacteroides uniformis*	99	99.2	
H1_B_07	*cfxA*	*Bacteroides uniformis*	100	98.8	
H1_B_11	*cfxA*	*Bacteroides uniformis*	99	99.2	*
H1_B_16	*cfxA*	*Bacteroides uniformis*	99	99.5	
H1_B_19	*cfxA*	*Bacteroides uniformis*	99	99.2	
H1_B_21	*cfxA*	*Bacteroides uniformis*	99	99.6	
H2_T_01	*tet*(Q)	*Bacteroides uniformis*	99	98.8	*
H2_T_04	*tet*(B)	*Pseudescherichia vulneris*	99	98.8	*
H2_T_07	*tet*(Q), *tet*(B)	*Bacteroides xylanisolvens*	98	99.8	*
H2_T_19	*tet*(B)	*Pseudescherichia vulneris*	98	99.3	
H2_T_24	*tet*(Q)	*Bacteroides dorei*	98	99.8	
H2_T_29	*tet*(Q), *tet*(B)	*Bacteroides uniformis*	99	99.8	
H2_M_07	*erm*(F)	*Bacteroides salyersiae*	99	90.1	*
H2_M_08	*mph*(A)	*Escherichia fergusonii*	99	99.0	*
H2_M_11	*erm*(F)	*Ruminococcus faecis*	99	99.6	
H2_A_15	*aph6-Id*	*Escherichia fergusonii*	99	98.4	
H2_A_19	*aph6-Id*	*Escherichia marmotae*	99	97.8	
H2_A_22	*aph6-Id*	*Escherichia fergusonii*	100	99.0	
H2_B_01	*bla*$_{\text{TEM-1B}}$	*Escherichia fergusonii*	99	98.0	
H2_B_04	*bla*$_{\text{TEM-1B}}$	*Escherichia marmotae*	99	98.0	
H2_B_06	*cfxA*	*Bacteroides thetaiotaomicron*	99	99.6	*
H2_B_07	*bla*$_{\text{TEM-1B}}$	*Escherichia marmotae*	99	99.0	
H2_B_11	*cfxA*, *cepA*	*Bacteroides fragilis*	99	99.4	*
H2_B_17	*cfxA*, *bla*$_{\text{TEM-1B}}$, *cepA*	*Bacteroides fragilis*	99	99.5	
H2_B_19	*cfxA*, *cepA*	*Bacteroides fragilis*	99	99.7	
H2_B_24	*cfxA*, *cepA*	*Parabacteroides distasonis*	99	99.0	*
H3_T_02	*tet*(Q), *tet*(X)	*Bacteroides uniformis*	98	99.3	*
H3_T_03	*tet*(Q), *tet*(X)	*Bacteroides uniformis*	98	100.0	
H3_T_05	*tet*(M)	*Enterococcus gallinarum*	98	99.3	*
H3_T_08	*tet*(Q)	*Dorea longicatena*	99	98.8	*
H3_T_19	*tet*(Q), *tet*(M), *tet*(X)	*Bacteroides uniformis*	98	100.0	*
H3_T_24	*tet*(Q), *tet*(M)	*Enterococcus gallinarum*	98	98.8	
H3_M_01	*erm*(B)	*Bacteroides dorei*	99	99.4	*
H3_M_02	*erm*(F)	*Dorea longicatena*	93	98.7	
H3_M_07	*erm*(F)	*Lachnoclostridium pacaense*	99	91.9	
H3_M_08	*erm*(B)	*[Clostridium] innocuum*	100	98.4	*
H3_M_13	*erm*(B)	*Lachnoclostridium pacaense*	99	92.0	
H3_B_04	*cfxA*	*Bacteroides uniformis*	99	99.8	*
H3_B_11	*cfxA*	*Bacteroides uniformis*	99	99.7	
H4_T_01	*tet*(Q), *tet*(X)	*Bacteroides cellulosilyticus*	99	98.8	*
H4_T_08		*[Eubacterium] rectale*	97	99.5	*
H4_T_29	*tet*(X)	*Eggerthella lenta*	99	98.8	*
H4_T_31	*tet*(Q)	*Bacteroides uniformis*	98	99.0	
H4_M_09	*erm*(F)	*Bacteroides uniformis*	99	99.3	*
H4_M_17	*erm*(F)	*Phascolarctobacterium faecium*	100	98.0	
H4_B_02	*cfxA*	*Bacteroides cellulosilyticus*	99	99.6	*

rRNA = ribosomal RNA; PCR = polymerase chain reaction; ARGs = antimicrobial resistance genes; COV = coverage; ID = identity; WGS = whole genome sequence; ARGs coloured blue indicate faint positive band after PCR; colour of colony name indicates antibiotic used for enrichment: **tetracycline** / **erythromycin** / **streptomycin** / **ampicillin**

5.2.4 Whole-genome sequencing of ARG-host isolates

Colonies with a unique taxonomic classification and ARG-profile were re-streaked and single colonies were picked for culture before DNA extraction and whole-genome sequencing. In total, 25 isolates (H1 = 5, H2 = 8, H3 = 7, H4 = 5) were short-read sequenced. From preliminary assembly using only the short-reads, four assemblies had unexpected classifications compared to the 16S rRNA gene classifications (Table 5.2). In sample H1, colony H1_T_14 was identified as *Bacteroides vulgatus* containing *tet*(Q), *tet*(M), and *tetA*(46). However, the sequenced isolate from this colony (H1-X) was *Streptococcus parasanguinis* containing *tet*(M) and *tetA*(46). This isolate was identical to isolate H1-01. This was likely due to the H1_T_14 colony being a mix of both *B. vulgatus* containing *tet*(Q), and *S. parasanguinis* containing *tet*(M) and *tetA*(46) during the colony PCRs. Similarly, isolate H1-02 was classified as *Collinsella*, despite the colony being identified as *B. vulgatus*. This issue also existed for sample H3, where two isolates expected to be *Bacteroides uniformis* were duplicates of isolate H3-01 *Enterococcus gallinarum* and H3-02 *Dorea longicatena* (Table 5.2).

As well as these unexpected duplicates arising from mixed colonies during colony PCR, some isolates were identical as they were selected from two different antibiotic class-enrichments. For example, *Escherichia coli* in sample H2 was sequenced twice as it was isolated during tetracycline enrichment (colony H2_T_04 (preliminarily identified as *Pseudescherichia vulneris* on the basis of 16S rRNA gene sequencing) containing *tet*(B)), and macrolide enrichment (colony H2_M_08 (*Escherichia fergusonii* on the basis of 16S rRNA gene sequencing) containing *mph*(A)). Likewise, *Bacteroides cellulosilyticus* was sequenced twice in sample H4 as it was isolated during both tetracycline resistance and β-lactamase enrichment (Table 5.2).

Table 5.2. Whole-genome sequencing of selected isolates

Colony name	16S rRNA gene BLAST result	# short-reads (2x150 bp)	# long-reads	Long-read N50 (bp)	# contigs	Assembly length (bp)	Whole-genome classification	Isolate name
H1_T_05	*Streptococcus sinensis*	572,054	16,377	17,587	2	2,035,100	*Streptococcus parasanguinis*	H1-01
H1_T_14	*Bacteroides vulgatus*	3,753,286	-	-	54	2,017,393	*Streptococcus parasanguinis*	H1-X (H1-01)
H1_T_24	*Bacteroides vulgatus*	2,109,438	17,575	25,003	2	2,304,762	*Collinsella*	H1-02
H1_M_02	*Bacteroides clarus*	3,073,776	96,153	11,843	3	4,171,424	*Bacteroides clarus*	H1-03
H1_B_11	*Bacteroides uniformis*	974,136	3,657	14,634	4	4,576,701	*Bacteroides uniformis*	H1-04
H2_T_01	*Bacteroides uniformis*	2,320,274	10,677	14,681	7	5,155,539	*Bacteroides uniformis*	H2-01
H2_T_04	*Pseudescherichia vulneris*	1,656,740	-	-	222	5,083,205	*Escherichia coli*	H2-X (H1-04)
H2_T_07	*Bacteroides xylanisolvens*	1,537,134	104,985	13,000	3	6,592,940	*Bacteroides xylanisolvens*	H2-02
H2_M_07	*Bacteroides salyersiae*	3,296,494	102,161	8,778	3	4,737,465	*Bacteroides sp014385165*	H2-03
H2_M_08	*Escherichia fergusonii*	3,091,714	9,345	19,211	24	5,082,493	*Escherichia coli*	H2-04
H2_B_06	*Bacteroides thetaiotaomicron*	2,171,802	8,651	12,781	53	6,784,289	*Bacteroides thetaiotaomicron*	H2-05
H2_B_11	*Bacteroides fragilis*	2,376,806	7,779	15,665	3	5,393,106	*Bacteroides fragilis*	H2-06
H2_B_24	*Parabacteroides distasonis*	2,053,814	3,099	20,033	2	5,286,814	*Parabacteroides distasonis*	H2-07
H3_T_02	*Bacteroides uniformis*	3,473,618	-	-	75	3,294,612	*Dorea longicatena*	H3-X1 (H3-02)
H3_T_05	*Enterococcus gallinarum*	2,279,508	5,562	21,394	1	3,391,120	*Enterococcus gallinarum*	H3-01
H3_T_08	*Dorea longicatena*	4,533,796	104,122	19,621	1	3,347,707	*Dorea longicatena*	H3-02
H3_T_19	*Bacteroides uniformis*	590,902	-	-	55	3,361,631	*Enterococcus gallinarum*	H3-X2 (H3-01)
H3_M_01	*Bacteroides dorei*	1,965,310	24,454	10,051	5	5,485,324	*Phocaeicola dorei*	H3-03
H3_M_08	*[Clostridium] innocuum*	1,535,918	5,140	23,593	7	5,961,434	*Enterocloster aldenensis*	H3-04
H3_B_04	*Bacteroides uniformis*	1,228,248	2,071	24,384	12	5,237,150	*Bacteroides uniformis*	H3-05
H4_T_01	*Bacteroides cellulosilyticus*	3,707,316	146,156	15,685	2	7,070,460	*Bacteroides cellulosilyticus*	H4-01
H4_T_08	*[Eubacterium] rectale*	373,948	183,317	5,813	1	3,302,144	*Agathobacter rectalis*	H4-02
H4_T_29	*Eggerthella lenta*	1,127,926	5,650	18,842	4	3,689,147	*Eggerthella lenta*	H4-03
H4_M_09	*Bacteroides uniformis*	3,824,358	-	-	79	4,997,438	*Bacteroides uniformis*	H4-04
H4_B_02	*Bacteroides cellulosilyticus*	4,259,424	-	-	57	7,013,020	*Bacteroides cellulosilyticus*	H4-X (H4-01)

rRNA = ribosomal RNA; bp = base pair; # = number of; colour of colony name indicates antibiotic used for enrichment: **tetracycline** / **erythromycin** / **streptomycin** / **ampicillin**; highlight colour of the 16S and whole-genome classification indicate discrepancies: expected classification and antimicrobial resistance genes (ARGs), conflict in ARGs present in whole-genome sequence vs colony polymerase chain reaction, conflict in both classification and ARGs present; assemblies statistics in blue indicate isolates assembled using short-reads only, the rest were hybrid assemblies of both short- and long-reads; isolate names in red are duplicates, with their identical genome shown (in brackets)

After discarding duplicate isolates, remaining isolates (n = 4, 7, 5, 4 for H1, H2, H3, and H4, respectively) were long-read sequenced and the genomes were hybrid assembled. All assemblies were >99% complete and <1% contaminated according to CheckM scores based on marker gene presence. Half of the assemblies (10/20) were fully complete assemblies with circular chromosome and plasmid contigs (Table 5.2, Figure 5.3). Remaining isolates were nearly fully complete, with the assembly graphs connecting contigs but unable to fully resolve a complete, circular chromosome (Figure 5.3). The most fragmented assembly was isolate H4-04 due to long-read sequencing failing for this isolate (Figure 5.3, Table 5.2).

The isolates contained a total of 54 ARGs. The ARGs present aligned with the colony PCR results in all but three isolates (Table 5.2). Isolate H2-02 (*Bacteroides xylanisolvens*) contained two copies of *tet*(Q) (Figure 5.3), however the colony PCRs were positive for both *tet*(Q) and *tet*(B) (Figure 5.2). This could have been due to the colony being mixed with a different strain containing *tet*(B), for example the *E. coli* isolate in the same sample. Isolate H3-02 (*Dorea longicatena*) contained *tet*(O) instead of *tet*(Q) (Figure 5.3), however it was only positive with a faint band for *tet*(Q) during colony PCR (Figure 5.2), so it was likely nonspecific amplification or due to a mixed colony. Similarly, H4-03 (*Eggerthella lenta*) was PCR-positive for *tet*(X), but the assembly only contained *tet*(W).

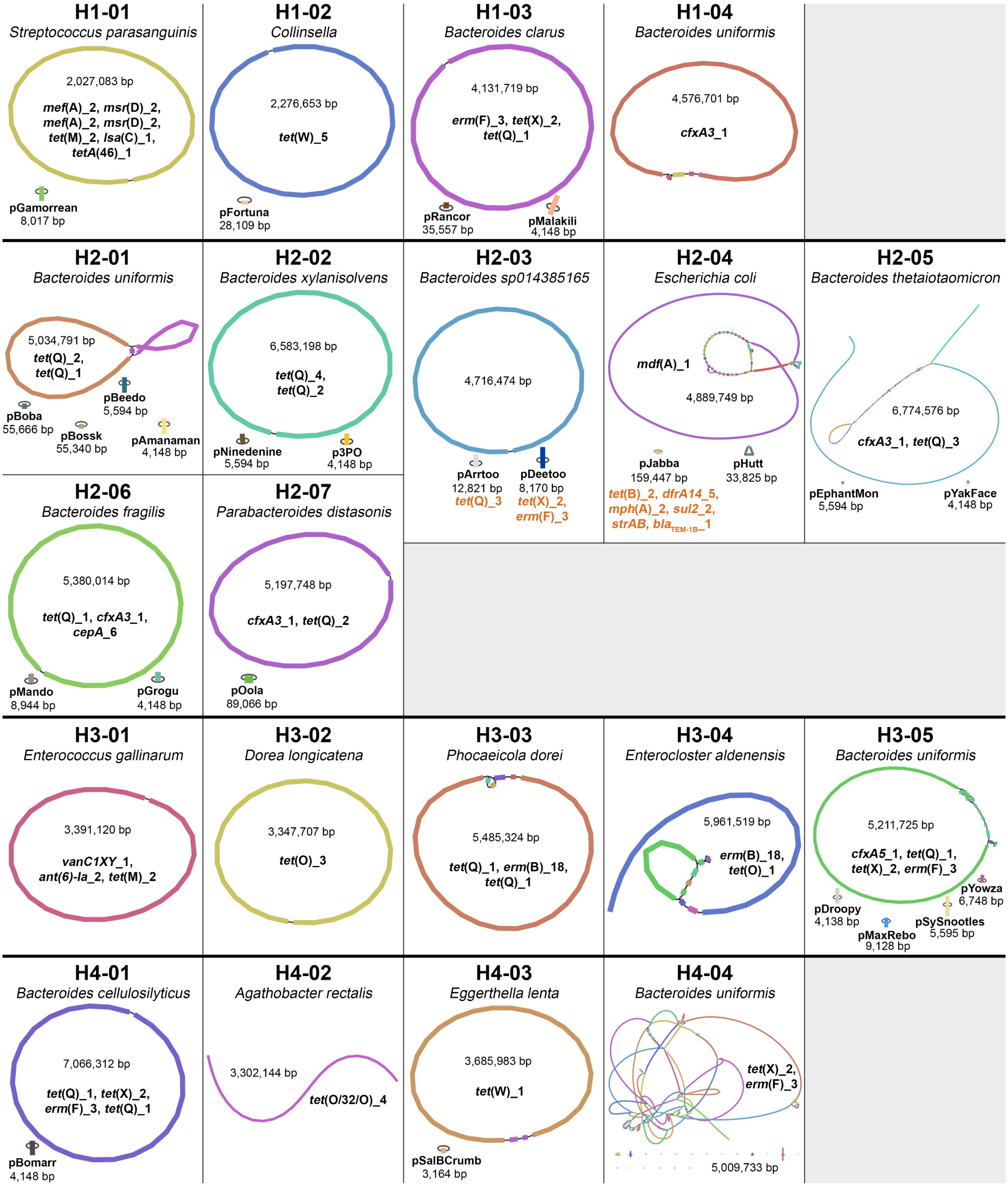


Figure 5.3. Assembly graphs of isolate genomes visualised using Bandage. Boxes shows the assembly graphs of each isolate (isolate name and taxonomic classification shown above). Each (randomly) coloured node represents an individual contig, with black lines representing connections between contigs. Contigs connecting to themselves indicate circular contigs. Lengths of connected contigs making up the chromosome are shown (bp = base pairs). Smaller circular contigs represent complete, circular plasmids (name and size shown below), with height of the node representing the sequencing depth relative to the chromosome. Antimicrobial resistance genes present in the genomes are shown (orange = present in plasmid).

As well as the expected ARGs from the colony PCR results, many of the isolates contained other ARGs that had not been screened for, with all isolates containing between 1-7 ARGs. The H1-01 *Streptococcus parasanguinis* isolate contained 7 ARGs (including two copies of both *mef*(A) and *msr*(D)) conferring resistance to macrolide, lincosamide, and tetracycline antibiotics. The *E. coli* isolate H2-04 contained 7 different ARGs (Figure 5.3) conferring resistance to tetracycline, diaminopyrimidine, macrolide, sulphonamide, aminoglycoside, penicillin, and first-generation cephalosporin antibiotics.

5.2.4.1 Classification of sequenced isolates

Using GTDB-Tk, 18/20 of the isolates were classified to species-level (Table 5.2, Figure 5.3). Using MLST, the MDR *E. coli* isolate, H2-04, was further classified as *E. coli* sequence type (ST) 69, a lineage of *E. coli* causing an increasing number of extraintestinal infections in humans (Riley, 2014; Shawa *et al.*, 2022).

The two isolates which were not classified to species-level, H1-02 and H2-03, were classified by GTDB-Tk as *Collinsella* and *Bacteroides sp014385165*, respectively (Table 5.2, Figure 5.3). The classification of H1-02 as *Collinsella* means that GTDB-Tk failed to place the genome beyond genus-level during the classification workflow. To further classify this to species-level, a core-genome alignment of H1-02 to reference genomes of all species in the *Collinsella* genus was performed, and this alignment was used to generate a phylogenetic tree (Figure 5.4a). H1-02 shared a clade with *Collinsella aerofaciens* on the tree, however only shared an ANI of 93.54% with the reference genome for *C. aerofaciens*. As the cut-off ANI to be considered the same species is >95% (Richter and Rosselló-Móra, 2009), H1-02 was compared with both core-genome alignment and ANI to all *C. aerofaciens* genomes uploaded to NCBI

GenBank. This revealed that H1-02 could be placed within a clade with other *C. aerofaciens* genomes and had an ANI above 95% with one other genome (95.04% with *C. aerofaciens* P10wA7) (Figure 5.4b). Therefore, H1-02 may be a strain of *C. aerofaciens* or a closely related novel species, but further taxonomic studies are required to substantiate this.

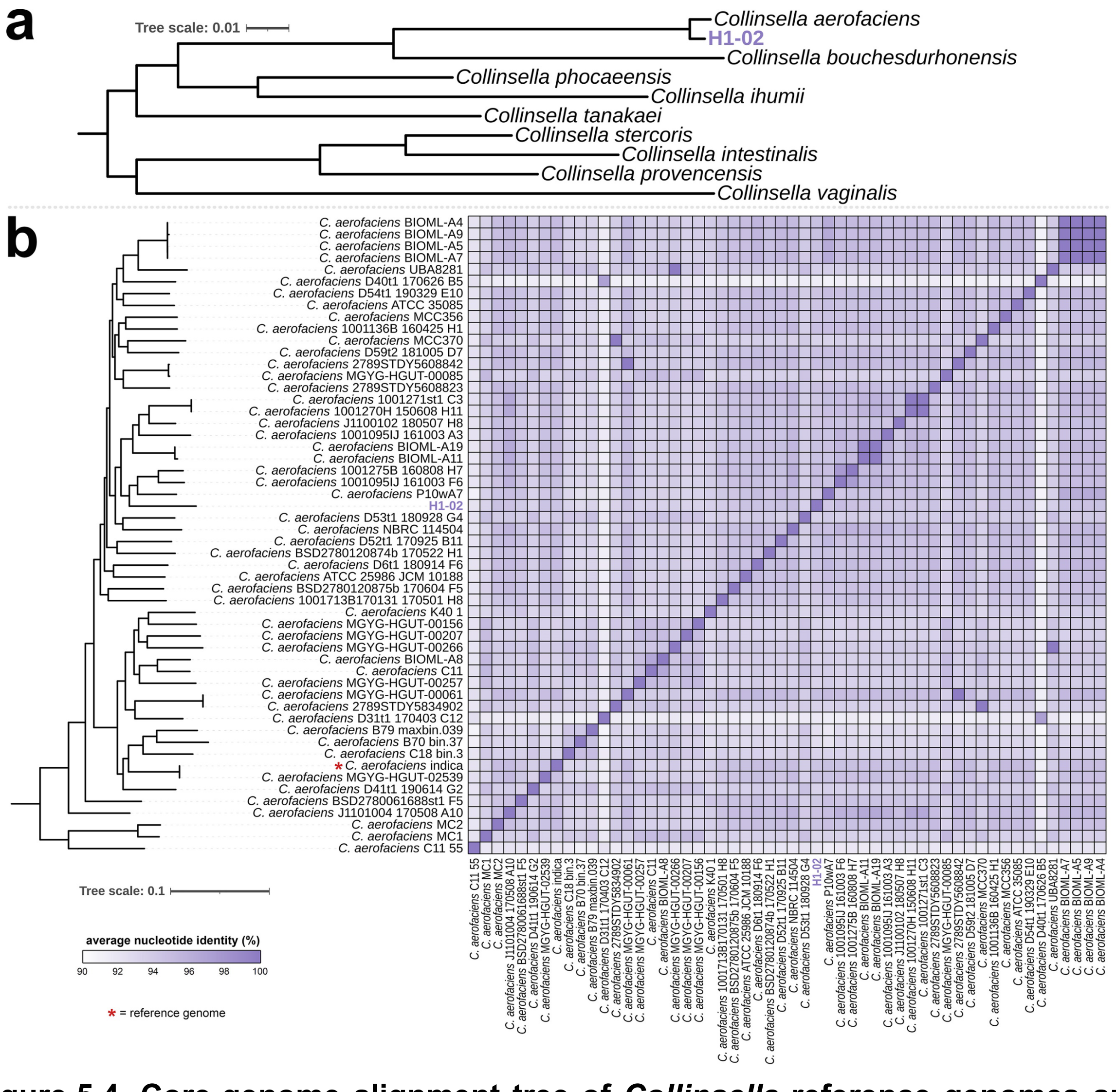


Figure 5.4. Core-genome alignment tree of *Collinsella* reference genomes and comparison to H1-02.

a) A core-genome alignment was generated between all *Collinsella* species reference genomes (downloaded from NCBI) and H1-02 (highlighted in purple). H2-03 is highlighted with a red asterisk as it does not share a clade with a reference genome. **b)** A core-genome alignment tree of isolate H1-02 and all *C. aerofaciens* genomes (downloaded from NCBI) is shown on the left. The heatmap shows the average nucleotide identity (%) of all genomes vs all genomes.

Isolate H2-03 was classified by GTDB-Tk to the species-level, but using a placeholder species name (*Bacteroides sp014385165*), which was generated for the GTDB based on NCBI accession numbers of the closest related genome that does not have a validly published name (Parks *et al.*, 2018). To identify its closest related *Bacteroides* species, a core-genome alignment tree of H2-03 and the reference genomes of all species in the *Bacteroides* genus was generated (Figure 5.5). The other sequenced *Bacteroides* isolates from this chapter were also included, and these all clustered with the reference genome of the species that GTDB-Tk had classified them as (Figure 5.5). However, H2-03 did not share a clade with a reference genome. The closest *Bacteroides* species to H2-03 in the tree were *Bacteroides nordii* and *Bacteroides salyersiae* (Figure 5.5), therefore a comparison to all genomes from those species was also performed (Figure 5.6). This showed that H2-03 still had a distinct clade from these *Bacteroides* species, and the highest ANI to H2-03 was 80.12% with *B. salyersiae* BFG-256 (Figure 5.6).

To further examine H2-03, every published *Bacteroides* species (sp.) genome was downloaded from NCBI (n = 3905), and the ANI with H2-03 was calculated for all vs H2-03. This revealed that H2-03 had an ANI of 99.90% and 99.01% with two MAGs uploaded to GenBank (accession numbers GCA_905199595.1 and GCA_934718525.1, respectively), both named “uncultured *Bacteroides* sp.”. There were no other *Bacteroides* genomes with an ANI to H2-03 greater than 80.12%. Although it shares a high ANI with two published genomes, both of these are MAGs of an uncultured and unclassified *Bacteroides* species. Isolate H2-03 can thus be considered to represent a novel species of *Bacteroides*.

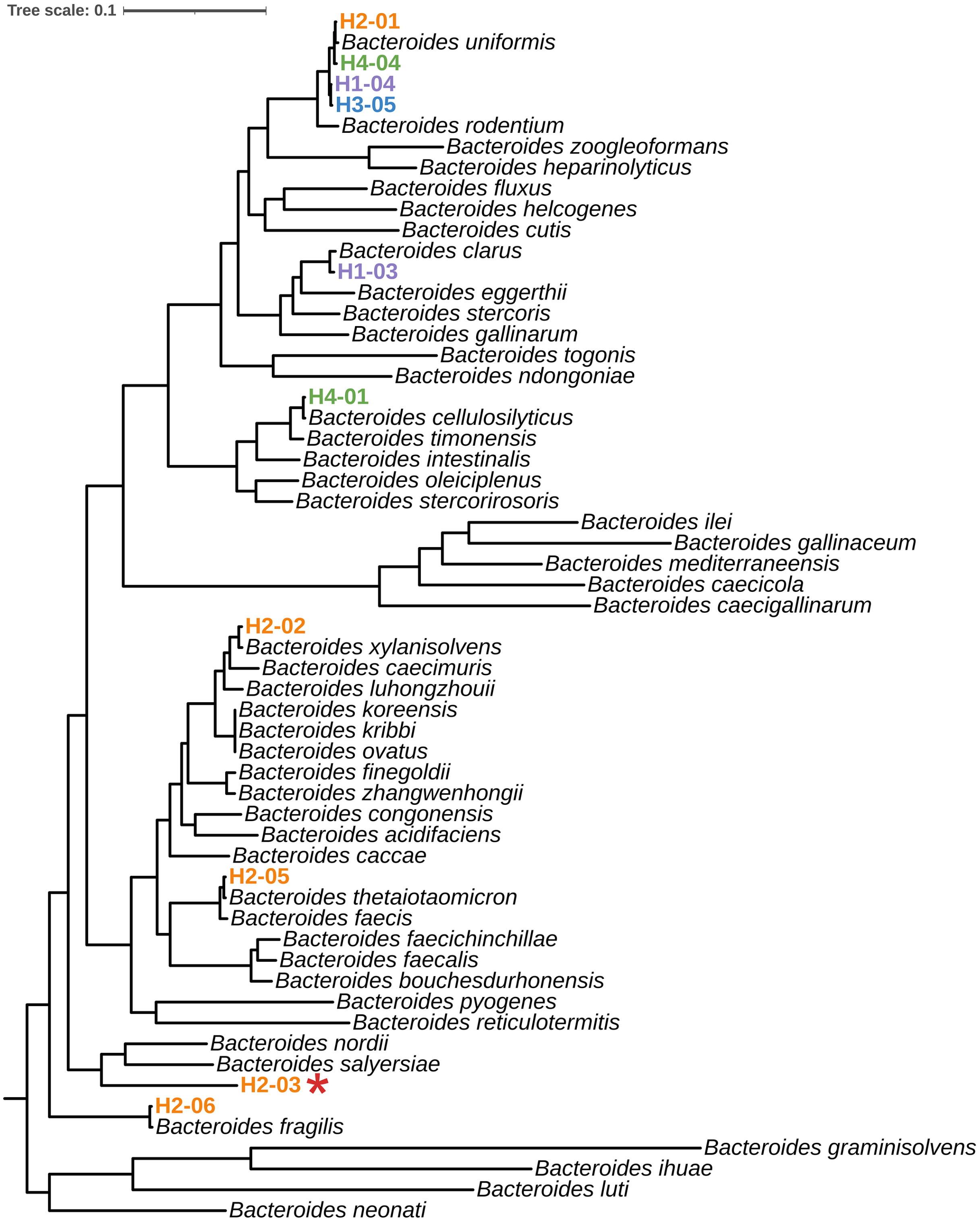


Figure 5.5. Core-genome alignment tree of *Bacteroides* reference genomes and cultured isolates.

A core-genome alignment was generated between all *Bacteroides* species reference genomes (downloaded from NCBI) and the genomes of the sequenced *Bacteroides* isolates (coloured based on the faecal sample they were cultured from: H1 = purple, H2 = orange, H3 = blue, H4 = green). H2-03 is highlighted with a red asterisk as it does not share a clade with a reference genome.

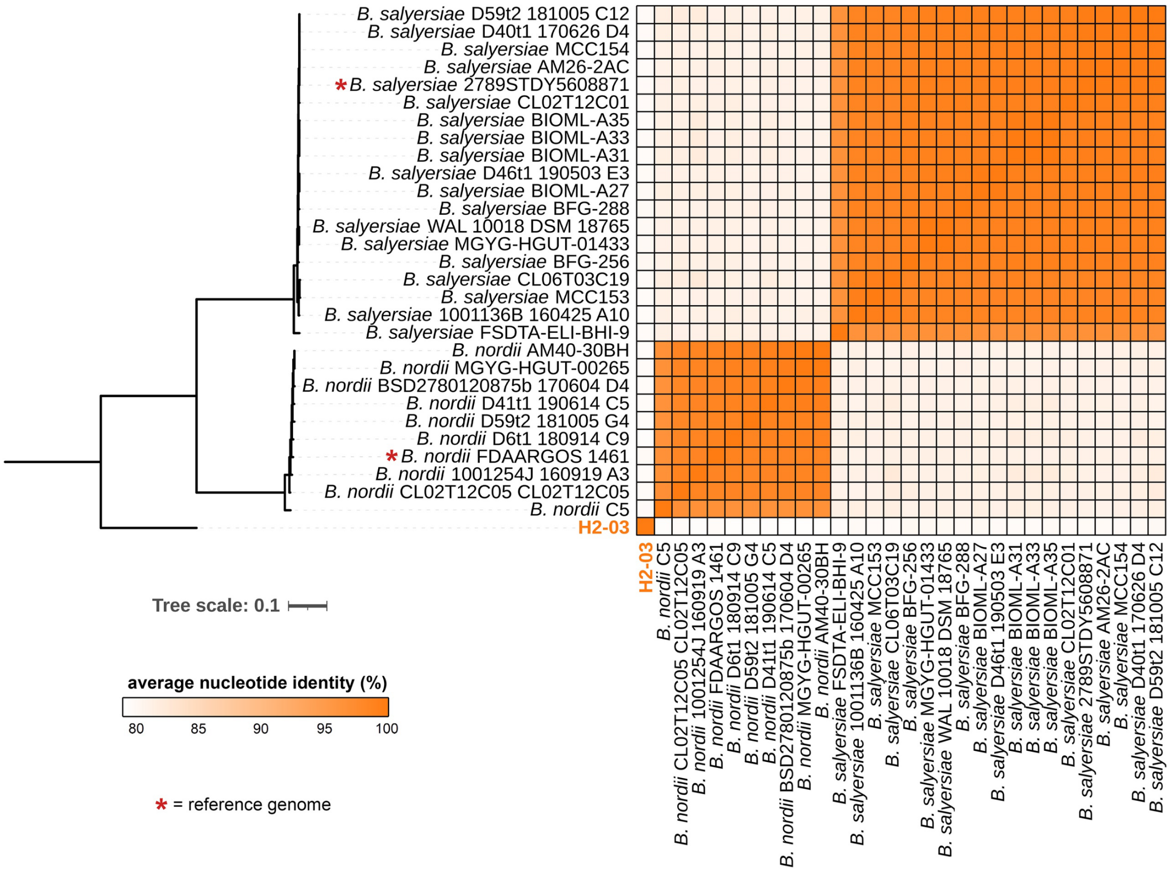


Figure 5.6. Comparison of H2-03 with *Bacteroides nordii* and *Bacteroides salyersiae*.

A core-genome alignment tree of isolate H2-03 and all *B. nordii* and *B. salyersiae* genomes (downloaded from NCBI) is shown on the left. The heatmap shows the average nucleotide identity (%) of all genomes vs all genomes.

5.2.4.2 Plasmids carried by ARG-host isolates

Most (13/20) of the sequenced isolates contained complete, circular plasmids in their assemblies (Figure 5.3, Table 5.3). The largest plasmid (159.4 kbp) in this collection of isolates, pJabba, an F-type plasmid in H2-04 *E. coli*, contained 6/7 of the ARGs in this MDR isolate. There were ARGs present in two other plasmids, pArrtoo and pDeetoo, which both contained tetracycline resistance genes (Table 5.3, Figure 5.3). Plasmids pArrtoo and pDeetoo were carried by isolate H2-03, the strain representing a novel species of *Bacteroides*. The closest match to pArrtoo in the NCBI nt database was *Phocaeicola vulgatus* strain CL10T00C06 plasmid pBCPT_CL10 with 71% coverage and 100% ID. This was the same closest match as pMaxRebo in H3-05 *Bacteroides uniformis*, which did not contain any ARGs and had 100% coverage to pBCPT_CL10 (Table 5.3). Plasmid pArrtoo has an identical backbone to pBCPT_CL10 and pMaxRebo with an insertion of CTn341, which contains *tet*(Q) (Husain *et al.*, 2017), and is the first example of evolution for this plasmid (Figure 5.7).

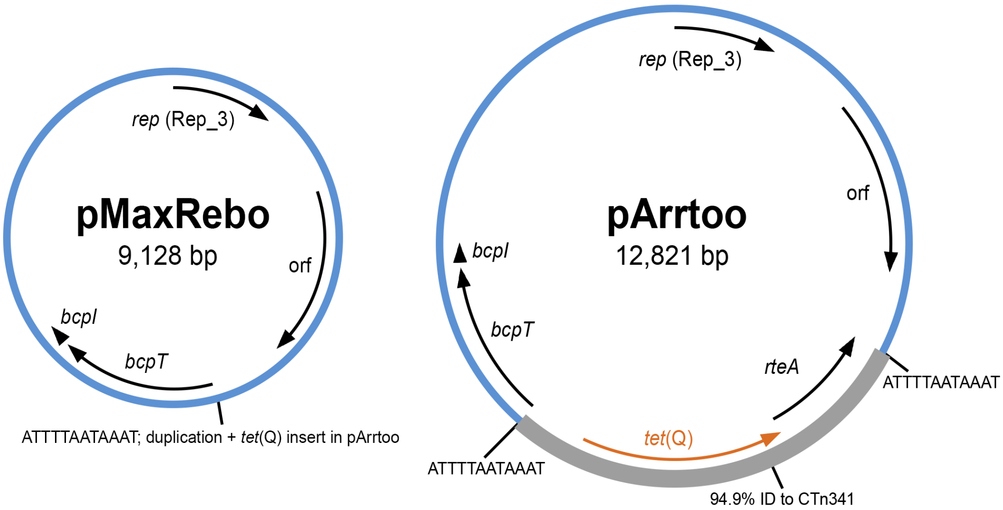


Figure 5.7. Comparison of pMaxRebo and pArrtoo.
Each plasmid is represented by a circle (labelled with name and length in base pairs (bp)). Blue represents a shared (identical) backbone, and the thicker grey section of pArrtoo indicates the inserted region containing the tetracycline resistance gene *tet*(Q). The sequences and positions of the target site duplication generated by the insertion are indicated. Arrows inside the circle show open reading frames (orfs).

Table 5.3. Plasmids present in isolates

Isolate	Plasmid name	Depth	Length (bp)	Closest BLAST match			
				BLAST match name	Accession	COV (%)	ID (%)
H1-01	pGamorrean	26.83x	8,017	*Streptococcus infantis* plasmid pSI01	JX275965	44	84.49
H1-02	pFortuna	0.35x	28,109	Unidentified plasmid FAKO02_3061	CP021610	74	98.75
H1-03	pRancor	1.76x	35,557	*Bacteroides thetaiotaomicron* F9-2 plasmid p1-F9-2	AP022661	99	99.73
	pMalakili	47.49x	4,148	*Bacteroides ovatus* strain 3725 D1 iv plasmid unnamed3	CP041398	100	100
H2-01	pBoba	1.69x	55,666	Unidentified plasmid FA1-2_000250F	CP021620	77	94.62
	pBossk	1.15x	55,340	*Bacteroides thetaiotaomicron* F9-2 plasmid p1-F9-2	AP022661	79	99.94
	pBeedo	32.90x	5,594	Unidentified plasmid FAKO05_2273	CP021595	100	99.79
	pAmanaman	35.34x	4,148	*Bacteroides ovatus* strain 3725 D1 iv plasmid unnamed3	CP041398	100	99.98
H2-02	pNinedenine	8.31x	5,594	Uncultured prokaryote from Rat gut metagenome	LN854347	100	99.95
	p3PO	12.17x	4,148	*Bacteroides ovatus* strain 3725 D1 iv plasmid unnamed3	CP041398	100	99.98
H2-03	pArrtoo*	23.02x	12,821	*Phocaeicola vulgatus* strain CL10T00C06 plasmid pBCPT_CL10	CP096968	71	100
	pDeetoo†	71.29x	8,170	*Bacteroides thetaiotaomicron* F9-2 plasmid p2-F9-2	AP022662	100	99.99
H2-04	pJabba‡	1.15x	159,447	*Escherichia coli* str. UMN026 plasmid p1ESCUM	CU928148	71	99.83
	pHutt§	6.39x	33,825	*Escherichia coli* str. UMN026 plasmid p2ESCUM	CU928149	99	99.16
H2-05	pEphantMon	13.13x	5,594	*Bacteroides fragilis* strain DCMOUH0042B plasmid pBFO42_2	CP036552	100	99.92
	pYakFace	31.13x	4,148	*Bacteroides ovatus* strain 3725 D1 iv plasmid unnamed3	CP041398	100	100
H2-06	pMando	7.15x	8,944	*Bacteroides uniformis* strain BUN24 plasmid pBUN24	EU818711	100	99.92
	pGrogu	7.53x	4,148	*Bacteroides ovatus* strain 3725 D1 iv plasmid unnamed3	CP041398	100	99.95
H2-07	pOola	4.07x	89,066	*Parabacteroides distasonis* strain CavFT-hAR46 chromosome	CP040468	41	92.46
H3-05	pMaxRebo	15.19x	9,128	Phocaeicola vulgatus strain CL10T00C06 plasmid pBCPT_CL10	CP096968	100	99.92
	pYowza	28.75x	6,748	*Bacteroides thetaiotaomicron* strain BFG-510 plasmid unnamed1	CP103117	99	99.3
	pSySnootles	149.43x	5,595	*Bacteroides fragilis* strain DCMSKEJBY0001B plasmid pBFS01_3	CP036549	100	99.87
	pDroopy	81.82x	4,138	*Bacteroides fragilis* strain DCMOUH0018B plasmid pBFO18_2	CP036544	100	99.95
H4-01	pBomarr	21.54x	4,148	*Bacteroides ovatus* strain 3725 D1 iv plasmid unnamed3	CP041398	100	100
H4-03	pSalBCrumb	1.83x	3,164	*Collinsella aerofaciens* ATCC 25986 strain JCM 10188 plasmid putative_pCaero2	CP048435	71	97.15

bp = base pairs; COV = coverage; ID = identity; depth is relative to chromosome.

Antimicrobial resistance genes present in plasmids: **tet*(Q)_3; †*tet*(X)_2, *erm*(F)_3; ‡*tet*(B)_2, *dfrA14*_5, *mph*(A)_2, *sul2*_2, *strAB*, *bla*$_{\text{TEM-1B}}$_1

Plasmid type: ‡F-type (FII:FIA:FIB); §X1

5.2.4.3 Genomic context of ARGs

The genomic context of the ARGs was investigated for all isolates. Only two isolates had ARGs present in a plasmid (9 plasmid ARGs in total present in 3 different plasmids), with the rest of the ARGs present in the chromosomal contig(s) of the whole-genome sequence (Figure 5.3). After using ABRicate with the ImmeDB and ICEberg databases, 21/45 of the chromosomal ARGs were associated with putative MGEs (Table 5.4). Out of these, 11 were present in ICEs, 4 in genomic islands, 3 in integrative and mobilisable elements (IME), and 2 in a transposon. Two isolates (H1-01 and H1-03) had ARGs on ICEs named in the ICEberg database. The ARGs *mef*(A), *msr*(D) and *tet*(M) in H1-01 *Streptococcus parasanguinis* were present in ICE*Spn*Tw19F14-1, a Tn*916*-like element. In H1-03 (*Bacteroides clarus*), *tet*(Q) was present in ICE*Bfr*YCH46-1, an ICE in the family of CTnDOT, a widespread conjugative transposon in *Bacteroides* (Waters and Salyers, 2013).

In sample H2, the tetracycline resistance gene *tet*(Q) was present in five different putative ICEs in four different species of *Bacteroides* (Table 5.4), showing how widespread and mobilisable this gene is in *Bacteroides*. Likewise, across the samples, *cfxA* was always present in putative genomic islands in species of *Bacteroides*, including three identical genomic islands in three different *Bacteroides* spp. in samples H1 and H2 (Table 5.4).

Table 5.4. Isolate antimicrobial resistance genes present in mobile genetic elements

Isolate	ARG	Type of MGE	MGE name	Closest NCBI accession / ImmeDB name	COV (%)	ID (%)
H1-01	*mef*(A)_2, *msr*(D)_2, *tet*(M)_2	ICE	ICE*Spn*Tw19F14-1 (Tn*916*-like)	CP000921	89.27	99.81
H1-03	*tet*(Q)_1	ICE	ICE*Bfr*YCH46-1 (CTnDOT family)	AP006841	100	99.99
H1-04	*cfxA3*_1	GEI	Putative GEI	NZ_GL622500.1:627449-639111	82.62	99.99
H2-01	*tet*(Q)_2	ICE	Putative ICE	NZ_CYYB01000002.1:626664-672571	75.85	99.97
H2-02	*tet*(Q)_4	ICE	Putative ICE	NZ_JGEG01000072.1:351658-400761	90.15	94.08
	tet(Q)_2	ICE	Putative ICE	NZ_AKBX01000005.1:406375-452975	81.55	95.97
H2-03	*tet*(Q)_3	Plasmid	pArrtoo	CP096968.1	71	100
	tet(X)_2, *erm*(F)_3	Plasmid	pDeetoo	AP022662.1	100	99.99
H2-04	*tet*(B)_2, *dfrA14*_5, *mph*(A)_2, *sul2*_2, *strAB*, *bla*$_{\text{TEM-1B}}$_1	Plasmid	pJabba	CU928148	71	99.83
H2-05	*cfxA3*_1	GEI	Putative GEI	NZ_GL622500.1:627449-639111	82.62	99.99
H2-06	*tet*(Q)_1	ICE	Putative ICE	NZ_JH976526.1:371311-427124	64.4	99.99
	*cfxA3*_1	GEI	Putative GEI	NZ_GL622500.1:627449-639111	82.62	99.99
H2-07	*cfxA3*_1	GEI	Putative GEI	NZ_JH636043.1:78434-88495	99.95	100
H3-02	*tet*(O)_3	IME	Putative IME	NC_021018.1:814351-825376	77.13	99.09
H3-03	*erm*(B)_18	ICE	Putative ICE	NZ_JH976466.1:1185361-1252986	61.92	99.99
	tet(Q)_1	ICE	Putative ICE	NZ_JGED01000015.1:51460-96974	99.94	97.34
H3-04	*erm*(B)_18	IME	Putative IME	NZ_NMTS02000042.1:145-10919	60.48	99.97
	tet(O)_1	IME	Putative IME	NZ_GL870812.1:22713-33714	84.87	100
H3-05	*cfxA5*_1	GEI	Putative GEI	NZ_GL622500.1:627449-639111	78.95	99.99
	tet(X)_2, *erm*(F)_3	Transposon	Putative transposon	NZ_LGTH01000001.1:1413421-1418525	77.59	99.92
H4-01	*tet*(Q)_1	ICE	Putative ICE	NZ_JH724296.1:340419-389167	76.07	99.98

ARG = antimicrobial resistance gene; MGE = mobile genetic element; COV = coverage; ID = identity; ICE = integrative and conjugative element; GEI = genomic island; IME = integrative and mobilisable element; genomic islands highlighted in grey are identical to each other

Overall, the genomic context of these ARGs showed that even though most were present in the chromosome, the majority of ARGs here were potentially mobilisable. To test this, conjugation assays were performed to determine whether *tet*(M) in H1-01 *S. parasanguinis* (present in ICE*Spn*Tw19F14-1) or *erm*(B) in H3-04 *Enterocloster aldenensis* (present in a putative IME) could be transferred to *Enterococcus faecium*. No transconjugants grew in either of the assays, however this may have been due to poor growth of the donor strains compared to the *Enterococcus* recipient strain which grew significantly faster and there was insufficient time to optimise the conjugation assay.

5.2.5 Comparison of cultured isolates to Hi-C ARG-host linkage results

The results of the Hi-C ARG-host linkage results from Chapter 4 were compared to the results of the WGS here. Firstly, the ARG-host linkage results from Chapter 4 were examined for each ARG present in each sequenced isolate (Figure 5.8, column 2 and 4). For the majority of isolates, (17/20) the ARGs were not linked by Hi-C to a bin with the same classification as the whole-genome sequence of the isolate. Sample H1, H2, and H4 had one isolate each (H1-02, H2-06, and H4-01) where the ARGs were linked by Hi-C to contigs in a bin with the same classification as the isolate (Figure 5.8, column 4).

All contigs that were linked by Hi-C to the ARGs were then aligned against the whole-genome sequence of the isolates, and this showed that contigs linked to ARGs with Hi-C aligned to the genome for 4/4, 1/7, 4/5, and 3/4 isolates in H1, H2, H3, and H4, respectively (Figure 5.8, column 5). The results of this indicated that some of the Hi-C results were consistent with culturing data. For example, in sample H1, 11.7-13.1% of Hi-C links to *tet*(W) linked the ARG to contigs in a *Collinsella* bin

(Figure 5.8a, column 4), and 11.7-13.5% of all *tet*(W)-linked contigs aligned to the genome of isolate H1-02 (*Collinsella* with *tet*(W) in its chromosome) (Figure 5.8a, column 5). Similarly, 23.1% and 12.7% of Hi-C links to *tet*(Q) and *erm*(F)-*tet*(X), respectively, in sample H4 were linking the ARGs to a *Bacteroides cellulosilyticus* bin (Figure 5.8d, column 4), and 30.5% of all contigs linking to *tet*(Q) and 13.4% linking to *erm*(F)-*tet*(X) aligned to the genome of H4-01 (*Bacteroides cellulosilyticus* with *tet*(Q), *erm*(F), and *tet*(X) in its chromosome) (Figure 5.8d, column 5). The metagenomic assembly had two contigs containing *tet*(Q), and whilst the contig containing *tet*(Q)_3 had no Hi-C links to contigs in the *B. cellulosilyticus* bin, 42.3% of contigs linking to it did align to the H4-01 genome (Figure 5.8d), meaning Hi-C had successfully linked the ARG to the contigs from the correct host, but those linked contigs had not been clustered into the *B. cellulosilyticus* bin.

This issue, where the Hi-C results had failed to link ARGs to a bin with the same classification of the sequenced isolates, was present for most of the sequenced isolates. For example, none of the Hi-C ARG-host links for *cfxA3* were linking to a *Bacteroides uniformis* bin, however 61.8% of contigs linking to *cfxA3* aligned to the H1-04 (*B. uniformis*) genome (Figure 5.8a). Likewise, none of the ARGs present in H1-01 *S. parasanguinis* were linked by Hi-C to a *Streptococcus parasanguinis* bin. However, 14.7%, 25.8%, and 14.9% of contigs linked to *tet*(M), *tetA*(46), and *msr*(D), respectively, did align to the genome of H1-01 (Figure 5.8a).

There were examples of Hi-C completely failing to link the ARGs to any contigs that aligned to the genome of the sequenced isolates. In sample H1-01, although contigs linking to three of the ARGs aligned to the genome, *mef*(A) had no links to contigs aligning to H1-01, and *lsa*(C) was not present in the H1 metagenomic assembly so was not linked to any host by Hi-C (Figure 5.8a). This issue was particularly prominent in sample H2, where 6/7 of the isolates had no ARGs with Hi-C links to contigs aligning to the genome. For H2-04, all of the ARGs present were also present in the H2 metagenomic assembly, however Hi-C failed to link any contigs to these ARGs (Figure 5.8b).

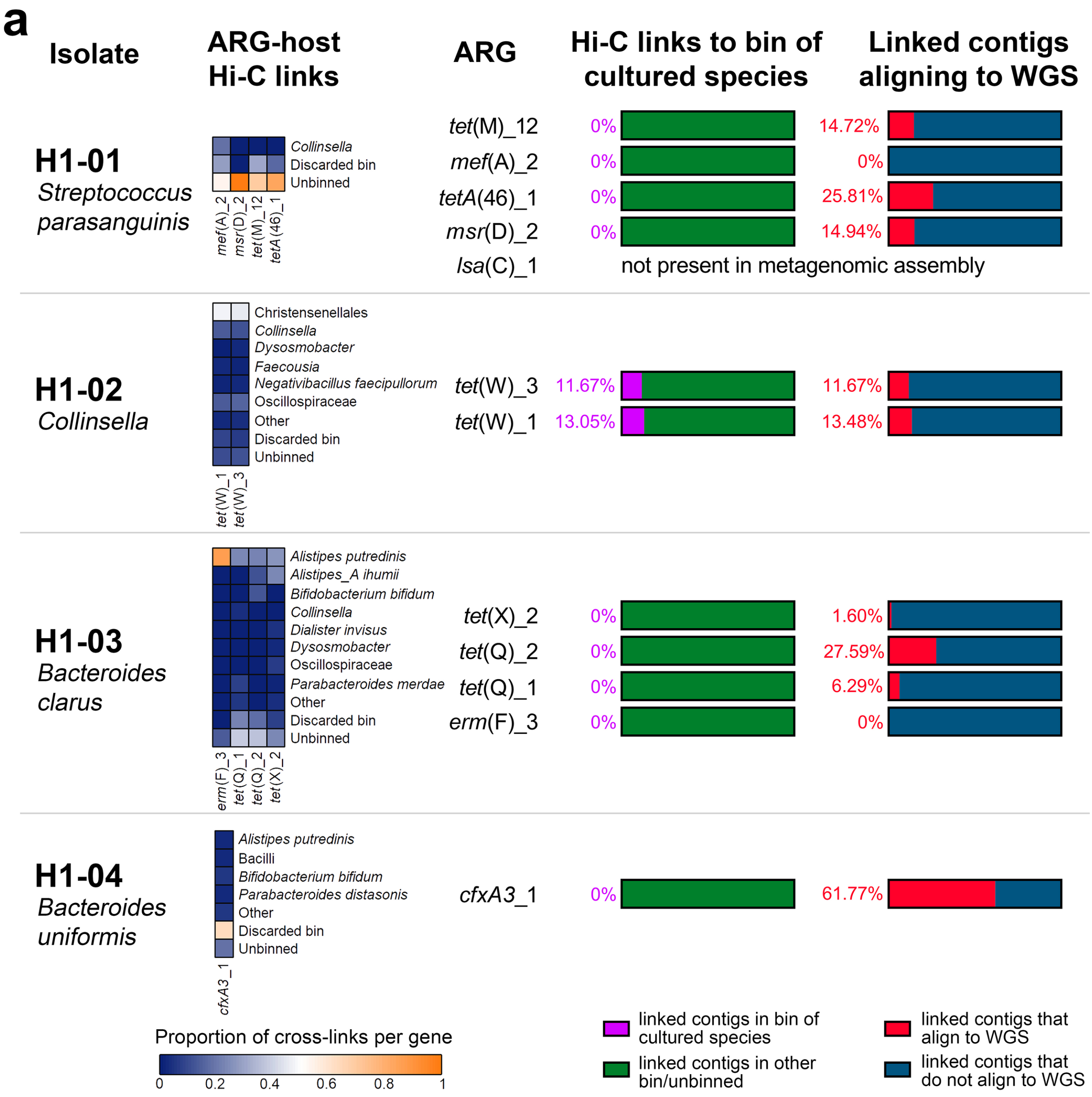


Figure 5.8 continued on next page.

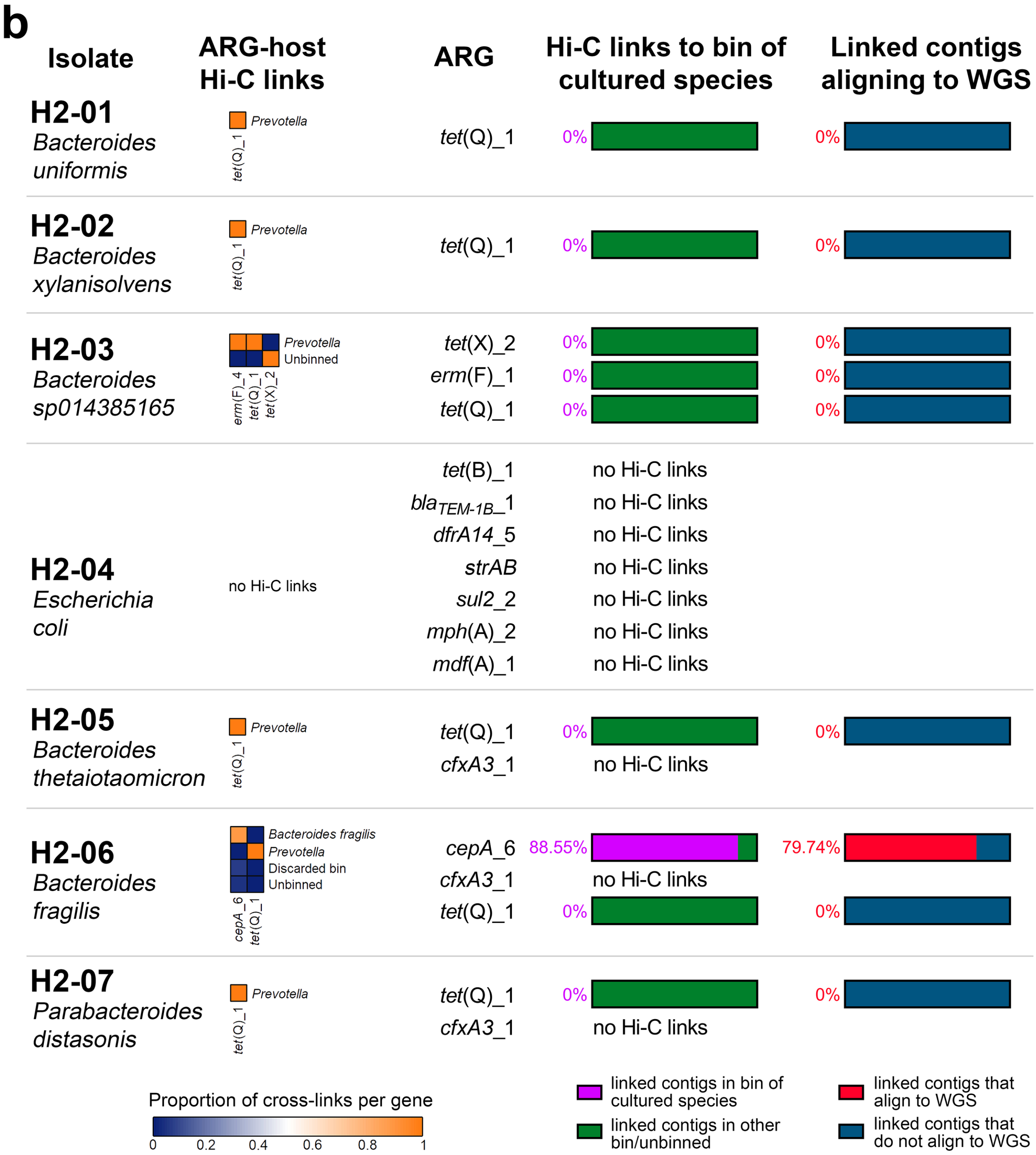


Figure 5.8 continued on next page.

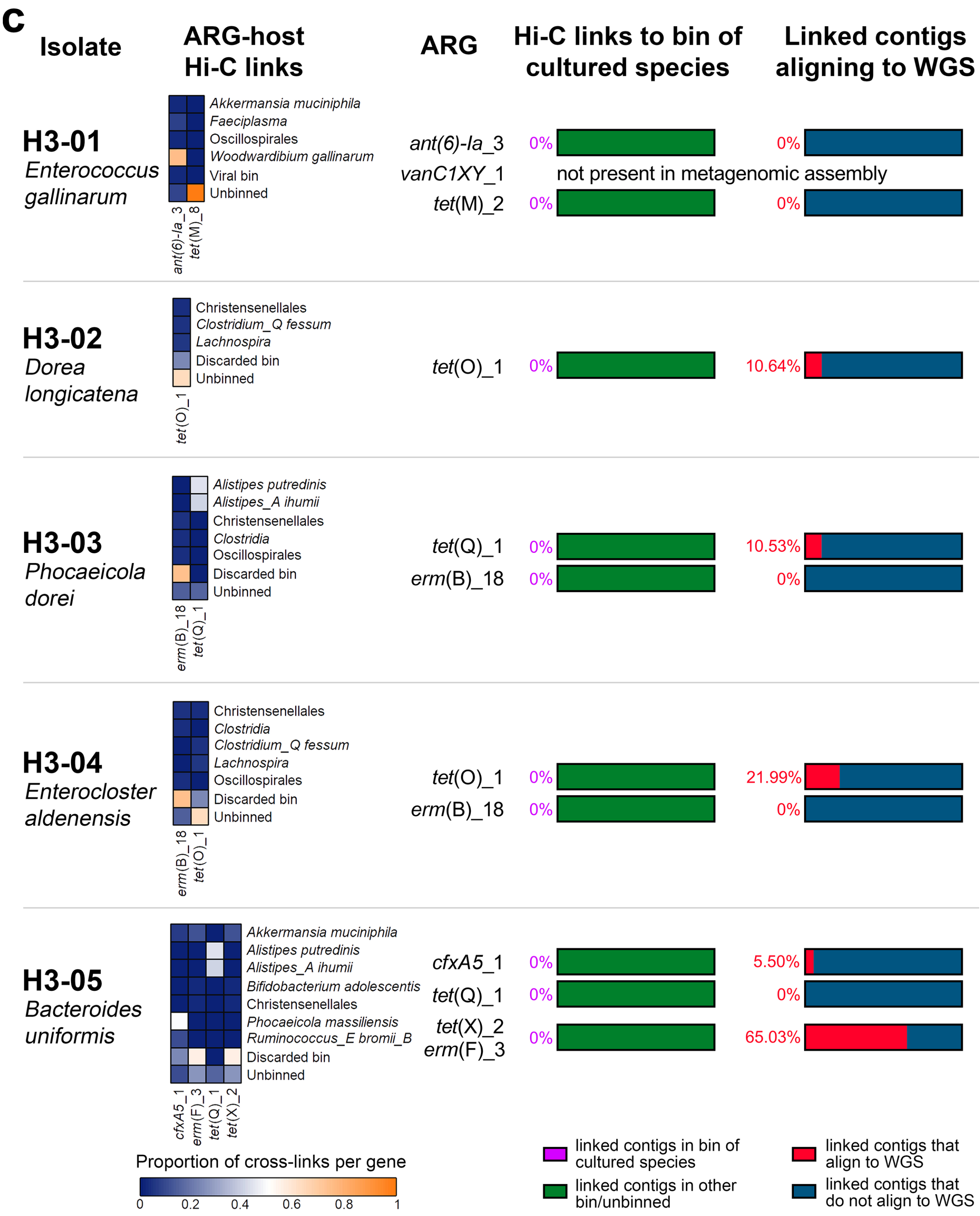


Figure 5.8 continued on next page.

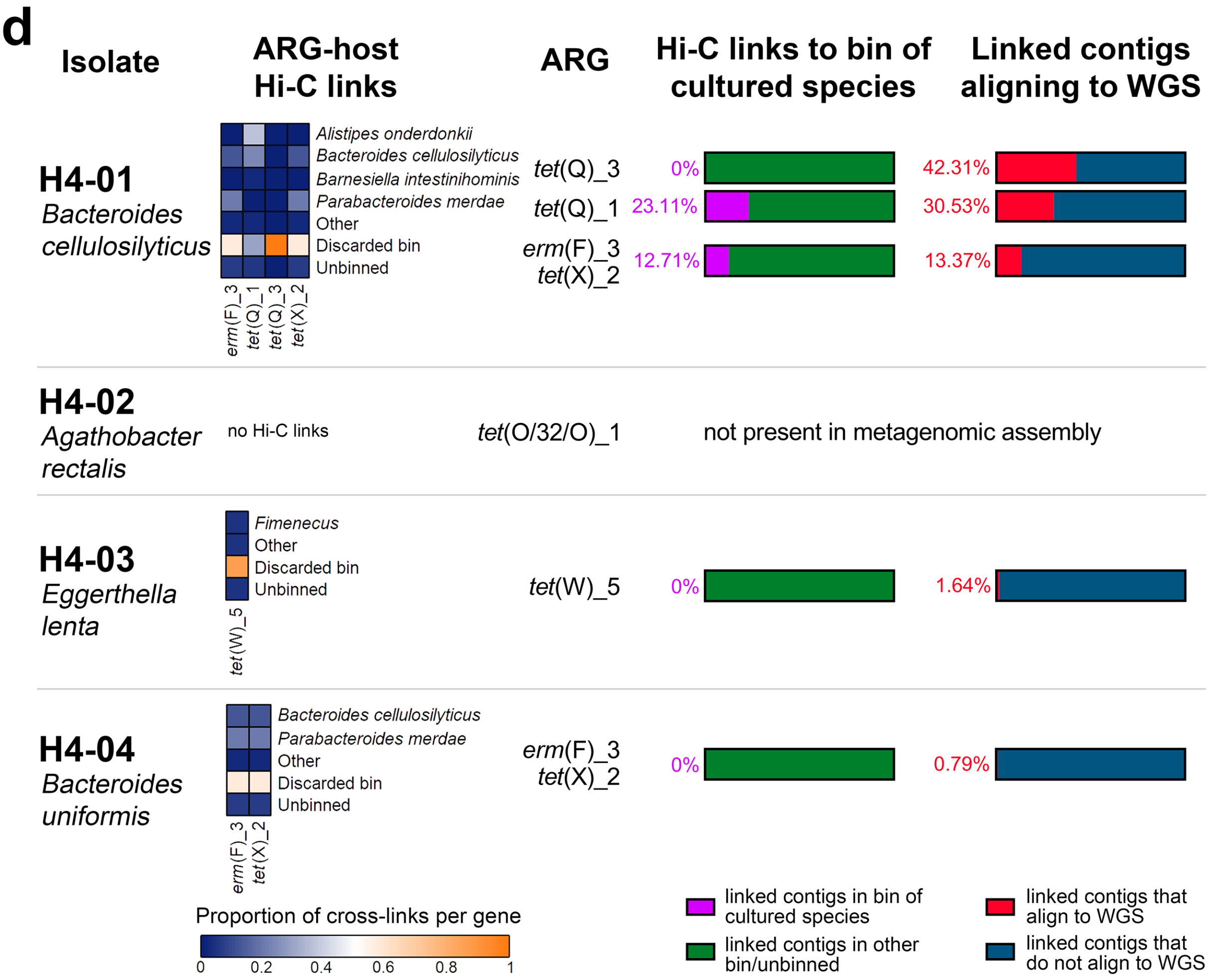


Figure 5.8. Comparison of Hi-C ARG-host linkage and sequenced isolates. Hi-C antimicrobial resistance gene (ARG)-host linkage was compared to the sequenced isolates for each sample: **a)** H1; **b)** H2; **c)** H3; **d)** H4. The name and taxonomic classification of each isolate is shown in the first column. The second column shows the Hi-C results for the ARGs present in the isolate genome that were linked to a host in Chapter 4. The fourth column shows each ARG present in the isolate genome. Column five shows the proportion of intercontig reads that linked the ARG to contigs present in a bin matching the taxonomic classification of the sequenced isolate. Column six show the proportion of all contigs linked by Hi-C to the ARG that also align to the isolate whole genome sequence (WGS) with >85% coverage and identity.

Where Hi-C had linked the ARGs to contigs that aligned to the sequenced genome of the isolates, but those contigs were not present in a bin with the same classification of the respective isolate, the contigs were most often present in a discarded bin or were unbinned (Figure 5.9). This further highlights that the H-LARGe v2 workflow developed in Chapter 4 is limited by the ability to cluster the majority of contigs into high-quality bins.

There were examples where the ARG-linked contigs that aligned to the genome were present in a bin with identical classifications to the isolate. For isolate H1-02, 97.5% of the contigs linked to *tet*(W) that also aligned to the H1-02 genome were present in a *Collinsella* bin. Similarly, 93.9% of aligned ARG-linked contigs for H2-06 (*Bacteroides fragilis*) were present in a *Bacteroides fragilis* bin, with the remaining contigs being unbinned. For H4-01 (*Bacteroides cellulosilyticus*), 89.3% of ARG-linked contigs that aligned to the genome were in a *B. cellulosilyticus* bin, 3.5% were in a bin classified as *Parabacteroides merdae*, in the same order (Bacteroidales) as *B. cellulosilyticus*, and the rest (7.2%) were in a discarded bin (Figure 5.9). Genome-aligned ARG-linked contigs for H3-03 (*Phocaeicola dorei*) were all present in a bin (*Alistipes putredinis*) classified to the same order (Bacteroidales) as H3-03. The same was true for 69.6% of aligned ARG-linked contigs for H4-04 (*B. uniformis*), which were again present in a *Parabacteroides merdae* bin. The remaining 30.4% of aligned ARG-linked contigs for H4-04 were present in a *Phocaeicola vulgatus*, which is in the same family (Bacteroidaceae) as *B. uniformis* (Figure 5.9).

Overall, the comparisons showed that Hi-C had often correctly linked ARGs to contigs originating from the correct bacterial host, but further indicated that linking ARGs to their host using Hi-C alone is limited by the ability to bin and classify contigs.

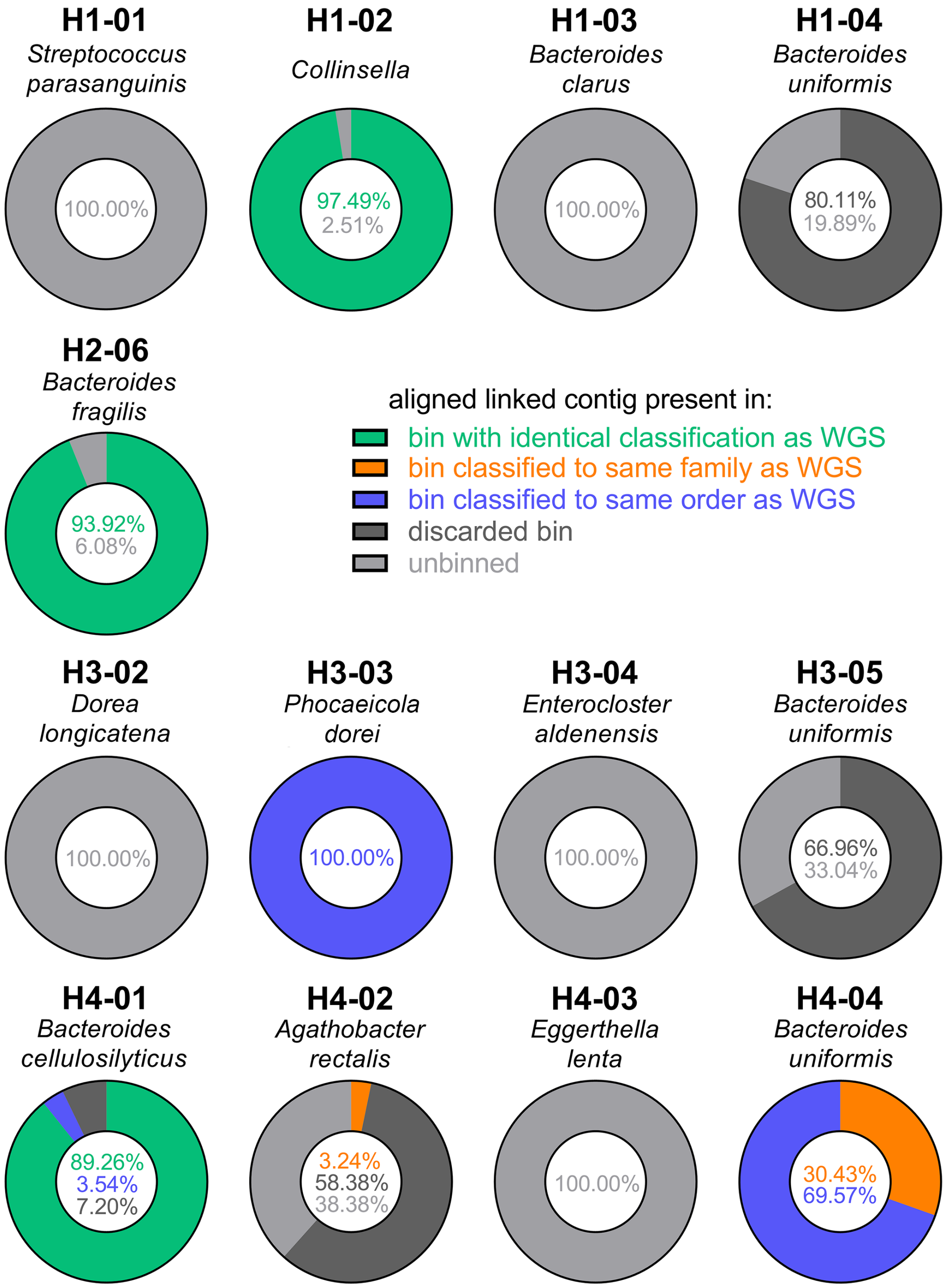


Figure 5.9. Classifications of bins that correctly linked contigs were present in. Each doughnut chart represents a different isolate (name and classification shown above). Contigs in the metagenomic assemblies that were linked by Hi-C to antimicrobial resistance genes present in the isolate genomes were aligned to the whole-genome sequences (WGS) of the respective isolate. If aligned with >85% coverage and identity, the bin that the contig was clustered into was checked. The doughnut charts show the proportion of contigs that were in a bin classified as the same species (green), family (orange), or order (blue) as the isolate, or present in a discarded bin (dark grey).

5.2.5.1 The impact of genome coverage on Hi-C linkage

The coverage of the genomes by the Hi-C reads, shotgun reads, and contigs in the metagenomic assemblies of the samples that the isolates were cultured from were assessed for each isolate. Nearly all isolates had around 100% coverage by the metagenomic shotgun reads from the samples they were cultured from (Figure 5.10a). Two isolates in sample H3, H3-01 *Enterococcus gallinarum* and H3-04 *Enterocloster aldenensis*, had less than 50% coverage by shotgun reads (29.4% and 44.9%, respectively). This may indicate that these isolates were present in sample H3 at low abundance. These H3 isolates also had low coverage by contigs in the metagenomic assembly (1.3% and 3.7%, respectively), meaning that the majority of the whole-genome sequence for each isolate was not represented by contigs in the metagenomic assembly. They also had 0% coverage by Hi-C reads from the H3 dataset, meaning that there were no Hi-C reads that mapped to the genomes (Figure 5.10a). Isolate H2-03 also had 0% coverage by Hi-C reads and only 16.4% coverage by assembled contigs, despite having 90.3% coverage by shotgun reads. All of the sequenced isolates in sample H2 had low coverage by Hi-C reads (37.2±21.7% average ± standard deviation) and contigs (73.7±24.8%) relative to the coverage by shotgun reads (97.7±3.2%) (Figure 5.10a). This may explain why the H-LARGe workflow failed to link many of the ARGs in this sample to their bacterial host (Figure 5.8b). Indeed, isolate H2-06 had the highest coverage by Hi-C reads (73.0%) and contigs (98.8%) in sample H2 (Figure 5.10a), and it was the only isolate in H2 where Hi-C had linked an ARG to contigs that aligned to the sequence genome (Figure 5.8b).

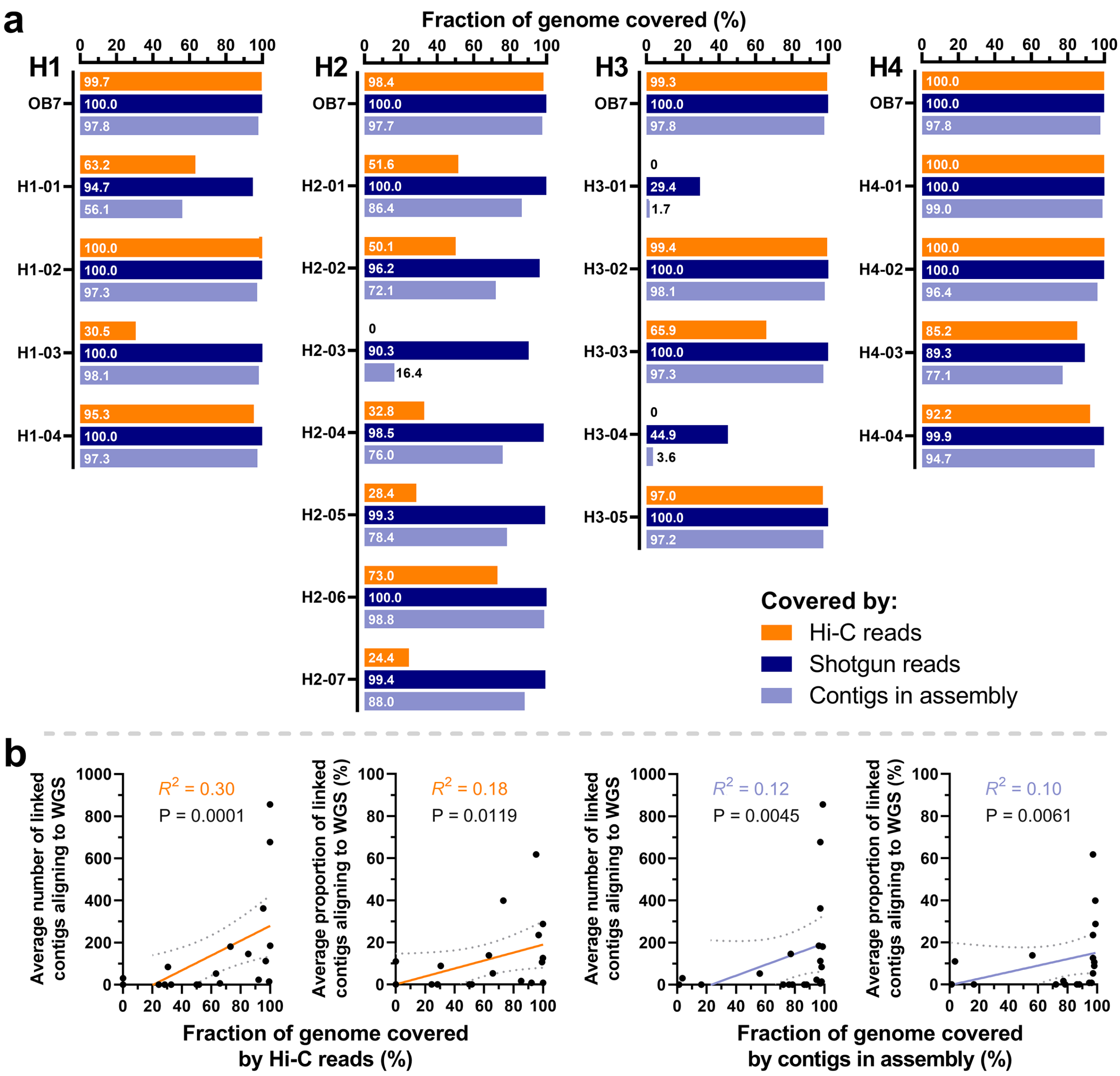


Figure 5.10. The effect of read and contig coverage on Hi-C results. Hi-C and shotgun metagenomic reads from the respective samples were mapped to the isolate whole-genome sequences to find the fraction of the genome covered by reads. Contigs from the respective shotgun metagenomic assembly were also aligned to the isolate genome to find the fraction covered by contigs. **a)** shows the fraction of each genome covered by Hi-C reads (orange), shotgun reads (dark blue), and contigs (light blue). The genome OB7 is the genome sequence of the *Acinetobacter pittii* OB7 spike-in added to the stool samples prior to Hi-C. **b)** shows the correlation between the fraction of the genome covered by Hi-C reads (orange) or contigs (light blue), and the total number/proportion of contigs linked by Hi-C to the isolate ARGs that aligned to the isolate whole genome sequence (WGS) with >85% coverage and identity. Slopes were calculated via linear regression analysis showing a statistically significant correlation (R^2 and approximate P values shown). Grey dotted lines indicate the 95% confidence interval.

In fact, there was a significant, albeit weak, correlation between the fraction of the genome covered by the Hi-C reads or assembled contigs and both the total number of and proportion of ARG-linked contigs that aligned to the sequenced genome of the isolates (Figure 5.10b). This showed that the higher the fraction of the genome covered by Hi-C reads, the more likely that ARGs carried by the isolate were linked to their host, with the same being true for coverage by contigs in the metagenomic assembly.

However, Hi-C read or contig coverage did not fully explain why Hi-C had not linked some ARGs to the sequenced isolates. For example, isolate H4-02 had 100% coverage by both Hi-C and shotgun reads, and 98.4% coverage by the metagenomic contigs (Figure 5.10a), however *tet*(O/32/O) was not present in the metagenomic assembly so could not be linked by the H-LARGe workflow. For another H4 isolate, H4-04 *Bacteroides uniformis*, only 0.79% of contigs linked to *erm*(F)-*tet*(X) aligned to the H4-04 genome, despite having over 90% coverage by Hi-C reads and metagenomic contigs (Figure 5.10a). In this case, H-LARGe may have failed to link these ARGs to H4-04 as these genes were widespread in the sample and H4-04 had lower abundance in the sample than the other hosts of these ARGs. The *erm*(F)-*tet*(X) contig was linked to several hosts in sample H4, and 13.4% of the linked contigs aligned to the genome of H4-01 *B. cellulosilyticus* which had a relative abundance of 2.7% in the sample. In comparison, H4-04 *B. uniformis* had a nearly 10-fold lower relative abundance (0.3%), which may explain why only 0.79% of the contigs linked to the ARGs aligned to the H4-04 genome. This indicates that for widespread ARGs that have multiple bacterial hosts in a sample, links to those with lower abundance may be missed, although it should be noted that the H-LARGe workflow had linked these ARGs to H4-04, albeit at a low proportion of total links.

5.3 Discussion

Previous Hi-C studies of the gut microbiome have focused exclusively on culture-independent analysis. The work in this chapter is the first to couple culturing with metagenomic Hi-C data to characterise the human gut resistome. The results validated some of the Hi-C data by confirming that the correct ARGs were linked to their host using the H-LARGe workflow in Chapter 4. However, the findings also revealed that there is limited overlap between Hi-C and culture-based approaches and highlighted the complementarity of the two methods to fully characterise the gut resistome.

5.3.1 Culturing of ARG-host isolates

To isolate the hosts of ARGs, the stool samples were added to modified Gifu Anaerobic Medium. This is a commonly used, non-selective medium that several studies have shown is able to grow the highest overall and most diverse isolates from the gut microbiome compared to other available media (Rettedal, Gumpert, and Sommer, 2014; Tramontano *et al.*, 2018), and mGAM has been proven to support the growth of many dominant species of human gut microbes (Gotoh *et al.*, 2017).

Using mGAM broth, I was able to enrich the hosts of 25/101 ARGs across four faecal samples. From there, individual colonies were screened for ARGs, resulting in 219/353 colonies being found to be PCR-positive for an ARG. To avoid sequencing duplicate isolates, the 16S rRNA gene of selected colonies was amplified using the 27F/1492R universal primers. These primers allow amplification of nearly the entire length of the 16S rRNA gene, and are some of the most widely used primers for species-level identification (Frank *et al.*, 2008). The amplified 16S rRNA genes were then sequenced and identified using the NCBI 16S ribosomal RNA sequences database. Although sharing genus-level taxonomic classification, many colonies with the same ARG-profile and morphology were classified as different species, such as several colonies in H2 being classified as different species of the *Escherichia* genus. This was likely due to the low-resolution offered by 16S rRNA gene sequencing (Janda and Abbott, 2007). Another example of discordance between the 16S rRNA gene classification and whole-genome sequence classification was the tetracycline-resistant isolate H4-02, classified as *Eubacterium rectale* during 16S rRNA gene classification and *Agathobacter rectalis* following WGS. This is because the species *Eubacterium rectale* was recently reclassified as *Agathobacter rectalis* (Rosero *et al.*, 2016), although this reclassification is controversial because it has been argued that there is a lack of evidence supporting it (Sheridan *et al.*, 2016; Karcher *et al.*, 2020).

Historically, 16S rRNA gene sequences that shared >97% identity were considered to be from the same species, and those that shared >95% identity were considered part of the same genus (Schloss and Handelsman, 2005). However, there are many species, including common gut bacteria such as those in the Enterobacteriaceae family, that have high similarity 16S rRNA gene sequences to each other (>99%) (Janda and Abbott, 2007). Recent studies have proposed that these historical cut-offs are no longer valid, and more specific examination of certain sub-regions, differing depending on the taxa being studied, is needed for better species discrimination (Johnson *et al.*, 2019). This was clear when assigning a taxon to the colonies based on the 16S rRNA gene sequences, as when aligned to the NCBI ribosomal RNA sequences database using BLAST, the species-level classification of the top results would often vary despite the 16S rRNA gene sequence sharing high sequence identity to all of them. However, 16S rRNA gene sequencing was only used here for screening of the colonies prior to WGS, therefore this was not a major issue. Based on this, colonies were selected for WGS based on the ARG-profile according to the colony PCRs, colony morphology, and genus-level assignment from 16S rRNA gene sequencing to avoid sequencing many duplicate isolates.

5.3.1.1 Whole-genome sequencing of ARG-hosts

Following WGS of selected isolates, it was revealed that some isolates had been sequenced in duplicate. The *E. coli* isolate in H2 was sequenced twice as it was cultured during enrichment for different classes of antibiotics, and in H4 *B. cellulosilyticus* was sequenced twice for the same reason. However, there were several examples of isolates unexpectedly sequenced in duplicate, or had unexpected classifications after WGS. Two isolates from sample H1, expected to be *B. vulgatus* from the 16S rRNA gene classification, were classified as *S. parasanguinis* and *Collinsella* after WGS. These isolates had different ARG-profiles compared to the expected ARGs from the colony PCRs, so it was not just a case of incorrect 16S rRNA gene classification. The *S. parasanguinis* isolate was identical to H1-01 (*S. parasanguinis*), so was likely an accidental duplicate of this isolate. The *Collinsella* isolate was confirmed to be from sample H1 from the coverage of metagenomic reads (100% of genome covered by both Hi-C and shotgun reads from H1). It is most likely that the colony used for colony PCR, including 16S rRNA gene amplification, was a mix of both *B. vulgatus* and *Collinsella*, and when re-streaked for WGS, a single colony of just the *Collinsella* isolate was picked. This is also likely the case for two other isolates in sample H3 of *D. longicatena* and *E. gallinarum*, both expected to be *B. uniformis*. Mixed colonies could also explain occasions where the colony was PCR-positive for an ARG that the whole-genome sequence did not contain.

Cross-contamination during re-streaking is also possible. Because no Bunsen burner can be used in an anaerobic chamber to create a sterile environment, stricter measures and organisation need to be taken for aseptic technique (Edwards, Suárez, and McBride, 2013). However, even with these measures, cross-contamination can be difficult to avoid (Mooiman *et al.*, 2021). In future, 16S rRNA gene amplification and sequencing of the extracted genomic DNA before WGS could be used to avoid accidentally sequencing isolates in duplicate.

Duplicate sequences were discarded, which left 20 isolates classified as a diverse range of gut bacteria. Half of the isolates were classified as a species of *Bacteroides*, with two more being species that until recently were part of the *Bacteroides* genus, H2-07 *Parabacteroides distasonis* (Sakamoto and Benno, 2006) and H3-03 *Phocaeicola dorei* (García-López *et al.*, 2019). This was unsurprising as *Bacteroides* are highly abundant in the healthy human gut microbiota (Feng *et al.*, 2018). On top of the high abundance, several ARGs are prevalent in species of this genus, including β-lactamase genes such as *cfxA* (Veloo *et al.*, 2019), as well as tetracycline and macrolide resistance genes (Bartha *et al.*, 2011; Kierzkowska, Majewska, and Mlynarczyk, 2020; Sóki *et al.*, 2020; Lamberte and van Schaik, 2022), all of which were enriched for during culturing.

Other isolates sequenced included some less studied species of the gut microbiome such as *Dorea longicatena* (H3-02) carrying the tetracycline resistance gene *tet*(O). *D. longicatena* is a dominant, yet understudied, species of the healthy human gut microbiota (Qin *et al.*, 2010). Interestingly, a strain of *D. longicatena* has been recently shown to be able to transfer a plasmid harbouring a tetracycline resistance gene to *Klebsiella oxytoca* by conjugation (Forster *et al.*, 2022). This interphylum HGT of an

ARG-plasmid indicates the potential for gut commensals such as *D. longicatena* to act as a reservoir for ARGs that could be transferred to even distantly related opportunistic pathogens, including those in the Enterobacteriaceae family.

Several of the sequenced isolates were common opportunistic pathogens in the gut microbiota. The macrolide and tetracycline resistant isolate, H3-04, was classified as *Enterocloster aldenensis*, formally known as *Clostridium aldenense* (Haas and Blanchard, 2020), a gut commensal that has been associated with intra-abdominal infection and bacteraemia (Warren *et al.*, 2006; Williams *et al.*, 2010). Isolate H4-03 was classified as *Eggerthella lenta*, an opportunistic pathogen commonly found in the gut microbiota that can cause life-threatening bacteraemia (Wong, Aoki, and Rubinstein, 2014; Gardiner *et al.*, 2015). Strains of *E. lenta* have been found to harbour vancomycin resistance genes (Stinear *et al.*, 2001). However, the isolate sequenced in this chapter only contained a single ARG, *tet*(W), conferring resistance to tetracycline, which is not used to treat blood-stream infections caused by *Eggerthella lenta* (Gardiner *et al.*, 2015). The *E. coli* ST69 isolate, H2-04, carried a plasmid, pJabba, containing ARGs conferring resistance to seven different antibiotic classes. This level of multidrug-resistance is concerning as ST69 is a lineage that recently ranked second in a list of the top 20 most common ExPEC STs (Manges *et al.*, 2019).

5.3.1.2 A novel species of *Bacteroides* was cultured

Two isolates were only classified to genus-level by GTDB-Tk. Further phylogenetic analysis revealed that H1-02 could be classified as the species *Collinsella aerofaciens*, although it is borderline and may represent a novel species. The other isolate only classified to genus-level, H2-03, is likely a new species of *Bacteroides* as the closest

ANI to a validly named species was 80.12% with a strain of *B. salyersiae*. Whole-genome ANI is a robust method for determining novel species, with a general consensus of >95% ANI being the cut-off score for the same species (Richter and Rosselló-Móra, 2009; Figueras *et al.*, 2014; C. Jain *et al.*, 2018). H2-03 carried two plasmids, pArrtoo and pDeetoo, both of which contained ARGs. It had similar (>99%) ANI with two MAGs in the GenBank database, although these were unnamed and uncultured. This study marks the first time this unnamed species of *Bacteroides* has been cultured and whole-genome sequenced. A future study should fully characterise this novel species of *Bacteroides*.

5.3.2 Many of the ARGs were associated with mobile elements

Two isolates, H2-03 *Bacteroides* sp. and H2-04 *E. coli*, carried ARGs present in plasmids, with ARGs in the rest of the isolates being present in the chromosome. However, despite not being present in plasmids, 21/45 of the chromosomal ARGs were able to be associated with putative MGEs. Three ARGs, *mef*(A), *msr*(D), and *tet*(M) were present in ICE*Spn*Tw19F14-1, an ICE in the Tn*916*-like family in H1-01 *S. parasanguinis*. The Tn*916* family is one of the most common and best-characterised family of conjugative transposons in Gram-positive bacteria, which play a major role in the wide-spread dissemination of tetracycline resistance (Rice, 1998; Roberts and Mullany, 2011). Tn*916*-like elements are responsible for widespread tetracycline resistance in *Streptococcus pneumoniae* due to the presence of *tet*(M), and often also contain the macrolide efflux pump genes *mef*(A) and *msr*(D) (Nikolaou *et al.*, 2020), as seen in H1-01. These conjugative transposons have been previously observed in *S. parasanguinis* isolated from dental plaques (Roberts *et al.*, 2001). In the *Bacteroides clarus* isolate, H1-03, *tet*(Q) was associated with ICE*Bfr*YCH46-1, an ICE in the

CTnDOT family, a CTn family harbouring *tet*(Q) that is highly prevalent in *Bacteroides* and related genera (Waters and Salyers, 2013; Veloo *et al.*, 2019).

These findings showed that, although not present in plasmids, nearly half of the ARGs were associated with mobile elements and thus had potential to horizontally transfer within the gut microbiota, indicating that these gut commensals can act as a reservoir of ARGs. Conjugation assays were attempted for transfer of *tet*(M) in *S. parasanguinis* and *erm*(B) in *E. aldenensis* to a strain of *Enterococcus faecium*, although no transconjugants grew in either of the assays. However, growth of both donor strains was poor compared to the *Enterococcus* recipient, so this may have impacted the results and these assays should be optimised.

5.3.3 Hi-C vs culture

Importantly, the WGS results could be compared to the Hi-C results from Chapter 4. The findings validated some of the Hi-C results as some of the hosts that Hi-C and the H-LARGe v2 had linked to ARGs were able to be isolated and sequenced, confirming that they did carry the ARG. This included the *Collinsella* isolate in H1 which was linked to *tet*(W), and the *Bacteroides cellulosilyticus* isolate in H4 which carried *tet*(Q), *erm*(F), and *tet*(X), all of which had been correctly linked to bins of the same classification in Chapter 4. In addition, aligning the respective ARG-linked contigs to the isolate genomes revealed that ARGs in half of the isolates had successfully been linked to contigs in the metagenomic assemblies originating from the sequenced host. However, these ARG-linked contigs were often unbinned, so the hosts of the ARGs had not been revealed using the H-LARGe v2 workflow. This further highlights the limitations caused by the host classification stages of the H-LARGe workflow and reinforces the need of improved binning as discussed in Chapter 4. Nevertheless, the results confirmed that

Hi-C, in combination with the H-LARGe workflow, is able to correctly link ARGs to contigs from their host in human faecal samples. This is the first study to have validated any Hi-C ARG-host associations using culturing.

These results are promising and show that with further improvement to the binning and host classification stage, Hi-C could be a powerful tool for linking ARGs to their host to classify the human gut resistome. However, there was limited overlap between the culturing and Hi-C results, as the ARG-linked contigs did not align to the whole-genome sequence for half of the isolates. To investigate this, the coverage of the genomes by Hi-C and shotgun reads from the respective faecal sample, as well as coverage by contigs in the respective metagenomic assembly, was calculated. The coverage data revealed, unsurprisingly, that the higher the coverage of the genome by Hi-C reads, the more likely it was that Hi-C had linked ARGs to host contigs for that isolate. This is expected, as having less Hi-C reads originating from the isolate clearly means there is less chance of having intercontig reads linking the ARGs to contigs from the isolate. The coverage of reads is also proportional to the relative abundance of the species in the sample, and species with low abundance can be missed by Hi-C (Press *et al.*, 2017). Some of the isolates had low coverage by the Hi-C reads despite having high coverage by the shotgun reads, which may indicate that there are some biases in which species ProxiMeta Hi-C performs best in when used on a faecal sample, warranting further investigation in this direction. Even when the abundance of the isolate in the sample was not low, it is possible that a species with much higher abundance could share the same ARG, which would mask the Hi-C results. This was likely the case for H4-03 and H4-04, where 1.64% and 0.79% of ARG-linked contigs aligned to the isolate genomes, respectively, indicating that Hi-C had successfully linked the ARGs to the isolates, but the vast majority of the links to those ARGs were to another species.

Culturing of the isolates also provided insights into some ARGs that Hi-C had failed to link. The best example of this was the *E. coli* ST69 H2-04 isolate, which had 7 ARGs, including 6 in plasmid pJabba. None of these 7 ARGs were linked to a host by Hi-C in Chapter 4, despite all being present in the metagenomic assembly. It is unclear why these ARGs were unable to be linked to a host using the H-LARGe workflow. However, by culturing the stool samples, I was able to isolate the strain carrying these genes in a plasmid and perform whole-genome sequencing, highlighting the benefit of coupling Hi-C and culture-based methods to link more ARGs to their hosts and broaden the characterisation of the human gut resistome.

5.3.4 Limitations and future studies

Only using a single growth condition to culture the hosts of ARGs may have limited the diversity and number of species isolated. Whilst mGAM has been shown to be capable of culturing a diverse range of gut bacteria (Gotoh *et al.*, 2017), using a single antibiotic concentration with a single media for each enrichment may have resulted in fewer organisms being isolated and selected for faster growing organisms. For example, during enrichment, if a few fast-growing organisms can grow in the presence of the antibiotic, other, slower-growing, resistant species would be outcompeted. Indeed, this was probably the case during the enrichments, as the hosts of only a few ARGs were isolated for each antibiotic. This is most obvious in the macrolide enrichments, where only one or two ARGs were enriched for, with the other ARGs actually decreasing in relative abundance compared to the controls that grew without antibiotic. A more systematic approach using different media, conditions, and antibiotic concentrations may have allowed more isolates to be cultured and sequenced, and improved the overlap with the Hi-C data. Systematic culturomic methods have been successful in culturing over 1,000 species from human faecal samples by utilising over 212 different culture conditions (Lagier *et al.*, 2016), and a similar method has been used to culture hundreds of chicken caecal microbiota species using 174 culture conditions (Crhanova *et al.*, 2019). These approaches were, however, not feasible within the time constraints of this project. Fastidious bacteria have also been successfully cultured from human faecal samples using a large-scale, single-medium approach (Browne *et al.*, 2016). This approach used YCFA medium to grow bacteria from faecal samples with and without prior treatment with ethanol to kill vegetative cells and select for spore-forming species. However, an important part of the approach was that the stool samples were placed in anaerobic conditions within 1 hour of passing (Browne *et al.*, 2016). The stool samples H1, H2, H3, and H4 used here had been stored in aerobic conditions at -80°C for several years, and this also likely resulted in many obligate anaerobic species dying. Ideally, future studies should use fresh stool samples, and more systematic approaches to culturing could be considered, coupled with Hi-C, to fully characterise the human gut resistome. This is not to say that the culturing was unsuccessful here, however, as 20 diverse isolates were cultured and sequenced using a single medium. The range of isolates consisted of both Gram-positive and Gram-negative obligate anaerobes, and even included a novel, previously uncultured, species of *Bacteroides*.

Other than sequencing more isolates, the overlap between the Hi-C and culturing results could be improved with better host classification during the H-LARGe workflow. As discussed in Chapter 4, improvement of the binning process by implementing long-read metagenomics could vastly improve the proportion of contigs clustered into bins and allow more ARG-linked hosts to be classified.

5.3.5 Conclusions

The work in this chapter demonstrated that the hosts of ARGs linked by Hi-C could be successfully isolated and used to validate the Hi-C results. However, the findings also revealed that the overlap between Hi-C and culture-dependant methods is limited, with some cultured isolates having no ARG-host links from the Hi-C data, and many ARG-linked hosts being missed during culturing. The results indicate that the techniques are complementary rather than entirely overlapping. Improvement of both techniques could improve the overlap, but, importantly, the data highlight the complementarity of Hi-C and culture-based approaches and show there is a role for both techniques to fully characterise the gut resistome.

CHAPTER 6

GENERAL DISCUSSION

Antimicrobial resistance is one of the greatest threats facing humanity. The human gut microbiota harbours many ARGs, collectively termed the human gut resistome, as well as opportunistic pathogens. Recent decades have seen a rise in infections caused by MDR opportunistic pathogens originating from the human gut microbiota. There is thus a need to characterise the human gut resistome to determine which bacterial species carry and transfer ARGs in the gut. Various techniques to study the transfer of ARGs in the gut have recently been developed, including 3C-based techniques to link bacterial genes to phylogenetic markers. This study aimed to use 3C-based techniques to link ARGs to their microbial hosts to explore the extent to which commensal bacteria can act as a reservoir for ARGs.

The work in **Chapter 3** described the development of the H-LARGe workflow, a novel bioinformatic workflow using 3C/Hi-C data to link ARGs to their hosts in gut microbiota samples. This included a reanalysis of existing 3C/Hi-C datasets and my own implementation of meta3C on a human faecal sample. During development of the analysis workflow, I found that many 3C/Hi-C datasets contain problematic background noise from spurious intercontig reads that can confound ARG-host associations. This noise existed in all analysed 3C/Hi-C datasets, although it had more of an impact on results if there was less efficient cross-linking during the experimental stages of 3C/Hi-C, which particularly affected the K_HiC dataset and, to a lesser extent, my own G_3C dataset. By analysing shotgun metagenomic reads with the same workflow as the 3C/Hi-C reads, I demonstrated that spurious intercontig reads are inherent for short-read metagenomic datasets, as approximately 2% of read pairs in a shotgun metagenomic dataset are intercontig, that is, each read of the pair maps to a different contig in the assembly. The G_3C dataset generated for Chapter 3 included two spike-in strains added prior to performing meta3C. This allowed me to investigate where intercontig reads were mapping to within the published whole-genome sequences of the spike-ins, revealing that intercontig reads were more likely to map to IS element regions in the genomes. The intercontig reads mapping to the spike-in genomes were also more likely to map near the ends of a contig compared to non-intercontig read pairs. This was true for all 3C/Hi-C datasets, where the majority of all spurious intercontig reads from shotgun metagenomic datasets mapped near the ends of contigs. These findings demonstrated that spurious intercontig reads were caused by fragmentation in the metagenomic assemblies. Therefore, during the H-LARGe workflow, intercontig reads are filtered to remove those mapping within 500 nt of the ends of a contig or mapping to an IS element. To further reduce the chance of noise impacting the results, contigs are only considered linked to each other if they are linked by at least 5 unique intercontig reads. These noise reduction steps are vital to reduce the potential for false host-ARG associations during analysis of 3C data and has been overlooked by previously published 3C/Hi-C studies.

After filtering of the intercontig read pairs, host-ARG associations were made using the H-LARGe workflow, which was able to link 87/123 ARGs to their microbial hosts across the datasets. The ARG-host associations revealed that ARGs were widespread in gut commensals, especially tetracycline and aminoglycoside resistance genes. Importantly, the ARGs of the spike-in strains in the G_3C sample were linked to the correct host, demonstrating that the H-LARGe workflow was able to successfully link ARGs their hosts using 3C data. However, the results were limited by the ability to accurately taxonomically classify the hosts using Kraken2 as there were several confounding ARG-host associations. The reanalysis of published 3C/Hi-C datasets also revealed that studies using ProxiMeta Hi-C achieved a considerably higher proportion of intercontig reads.

Following the development of the H-LARGe workflow, and the insights from reanalysis of published 3C/Hi-C datasets, I implemented ProxiMeta Hi-C on four human faecal samples in **Chapter 4**. The H-LARGe workflow was also further optimised to improve the taxonomic classification of the ARG-hosts by implementing binning. ProxiMeta Hi-C achieved a higher proportion of intercontig reads (>20%) than the published datasets reanalysed in Chapter 3, and thus is the desirable protocol for future 3C-based studies on the gut microbiome. Binning was implemented into version 2 of the H-LARGe workflow for better host classification. The improved workflow was used to link 87/119 ARGs, as well as 64/77 plasmids, to contigs from their hosts across the four samples. However, the classification of these hosts was limited by the ability to bin the majority of contigs as many of the ARG-linked contigs were unbinned. Nevertheless, the results indicated that ARGs were widespread in gut commensals, particularly those from the phylum Firmicutes. Several clinically important ARGs, like the multiresistance gene *cfr*(C), were widespread in commensals in the class Clostridia, which is particularly concerning as this gene has been identified as present in plasmids in clinical isolates of the opportunistic pathogen *C. difficile* (Chatedaki *et al.*, 2019), and *cfr* genes confer linezolid resistance in *E. faecium* (Deshpande *et al.*, 2015).

In **Chapter 5**, the hosts of ARGs were cultured from the same faecal samples used for Hi-C in Chapter 4. In total, 20 unique, ARG-carrying isolates were cultured and whole-genome sequenced, allowing investigations into the genomic context of ARGs and comparison to the Hi-C results. The sequenced isolates were from a diverse range of gut commensals and included two potential novel species from the genera *Collinsella* and *Bacteroides* that may require further taxonomic study. The majority (30/54) of ARGs across all isolates were associated with MGEs, including 9 ARGs carried on 3 plasmids, and 21 chromosomal ARGs associated with putative ICEs, IMEs, genomic islands, and transposons. Comparisons to the Hi-C results revealed limited overlap between the cultured isolates and the Hi-C results from Chapter 4. Nonetheless, some Hi-C results were able to be validated by the WGS results, as several ARG-hosts indicated by the H-LARGe workflow were cultured and isolated, confirming the carriage of the linked ARGs. Overall, the work in Chapter 5 highlighted the complementarity of Hi-C and culture-dependent approaches.

6.1 Future directions

The H-LARGe workflow is able to link ARGs to their hosts using 3C/Hi-C data. However, one of the weaknesses of the workflow was host classification which is limited by the success of the binning process to cluster the majority of the contigs into high-quality bins. Future work should refine this stage of the workflow. Other binning algorithms could be implemented, such as bin3C (Demaere and Darling, 2019), HiCBin (Du and Sun, 2022), hicSPAdes (Ivanova *et al.*, 2022), or HAM-ART (Kalmar *et al.*, 2022). Additionally, metagenomic long-read sequencing could improve the binning process. Several studies have demonstrated that metagenomic long-read sequencing can improve binning (Xie *et al.*, 2020; Cuscó *et al.*, 2021; Jin *et al.*, 2022), including in combination with Hi-C data (Stewart *et al.*, 2018; Bickhart *et al.*, 2022; Cuscó *et al.*, 2022), although no studies so far have used this combination to explore the human gut microbiota. With improved binning from long-read metagenomic data, the H-LARGe workflow should be capable of resolving more hosts of ARGs in the gut microbiota.

In addition, other variations of Hi-C, such as CHi-C, could be considered for future studies of the human gut resistome. A recently developed targeted sequence capture platform, named ResCap, is comprised of probes for nearly 8,000 ARGs and has been employed for analysing gene abundance and diversity in faecal samples (Lanza *et al.*, 2018). Combining CHi-C with the ResCap platform could allow ligated DNA fragments containing a diverse range of ARGs to be enriched and amplified prior to sequencing. This would allow less costly, shallower sequencing of the CHi-C library whilst acquiring more intercontig reads linking ARGs to their host, potentially leading to a greater number of ARG-host associations during the H-LARGe workflow.

The ARG-host associations found here indicated that commensals in the gut microbiota are a reservoir for ARGs. This was true for both Gram-negative commensals, such as those in the phylum Bacteroidetes, and Gram-positive bacteria, such as those in the phylum Firmicutes. Whilst tetracycline and β-lactam resistance genes were widespread in *Bacteroides* spp., inter-phyla transfer of these resistance genes to opportunistic pathogens is most likely a rare event (Ellabaan *et al.*, 2021). However, the results from this work indicate that Gram-positive gut commensals may be an important reservoir of clinically relevant ARGs. In particular, future research should focus on Clostridia and other Firmicutes in the gut to see the extent to which they are transferring ARGs to opportunistic pathogens like species of *Clostridioides* and *Enterococcus*. The findings in Chapter 4 indicated that genes conferring resistance to linezolid are associated with gut commensals from the Clostridia class, and as AMR rises in Gram-positive opportunistic pathogens, resistance to this last-resort drug becomes increasingly concerning (Bender *et al.*, 2018).

6.2 Conclusion

The work in this thesis has demonstrated a novel bioinformatic workflow, H-LARGe, to link ARGs to their microbial hosts in microbiome data using 3C-based data. Refinement of this workflow should continue to further improve the host assignment. I also demonstrated the advantages and complementarity of coupling 3C-based studies with culturing. The findings from implementing the H-LARGe workflow with Hi-C data from human faecal samples has shown that commensals are an important reservoir of ARGs in the human gut microbiota. This is particularly important for Gram-positive gut microbes, and future work studying the human gut resistome should focus on these. Overall, this work has added to the understanding of the role that commensal species in the human gut microbiota play in the emergence of MDR opportunistic pathogens.

REFERENCES

Abbo, L., Shukla, B. S., Giles, A., Aragon, L., Jimenez, A., Camargo, J. F., Simkins, J., Sposato, K., Tran, T. T., Diaz, L., Reyes, J., Rios, R., Carvajal, L. P., Cardozo, J., Ruiz, M., Rosello, G., Cardona, A. P., Martinez, O., Guerra, G., Beduschi, T., Vianna, R., and Arias, C. A. (2019) 'Linezolid- and Vancomycin-resistant *Enterococcus faecium* in Solid Organ Transplant Recipients: Infection Control and Antimicrobial Stewardship Using Whole Genome Sequencing', *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 69(2), pp. 259–265. doi: 10.1093/CID/CIY903.

Abraham, E. P. and Chain, E. (1940) 'An Enzyme from Bacteria able to Destroy Penicillin', *Nature 1940 146:3713*, 146(3713), pp. 837–837. doi: 10.1038/146837a0.

Acuña-Amador, L., Primot, A., Cadieu, E., Roulet, A., and Barloy-Hubler, F. (2018) 'Genomic repeats, misassembly and reannotation: A case study with long-read resequencing of *Porphyromonas gingivalis* reference strains', *BMC Genomics*, 19(1). doi: 10.1186/s12864-017-4429-4.

Adamecz, Z., Nielsen, K. L., Kirkby, N. S., and Frimodt-Møller, N. (2021) 'Aminoglycoside resistance genes in *Enterococcus faecium*: mismatch with phenotype', *Journal of Antimicrobial Chemotherapy*, 76(8), pp. 2215–2217. doi: 10.1093/JAC/DKAB137.

Adobe (2018) *Adobe Photoshop CC* (Version 20.0.0) [Computer program]. Available at: https://www.adobe.com/uk/products/photoshop.html.

Ahmed, M. O. and Baptiste, K. E. (2018) 'Vancomycin-Resistant Enterococci: A Review of Antimicrobial Resistance Mechanisms and Perspectives of Human and Animal Health', *Microbial Drug Resistance*, 24(5), pp. 590–606. doi: 10.1089/MDR.2017.0147/ASSET/IMAGES/LARGE/FIGURE1.JPEG.

Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A. L. V., Cheng, A. A., Liu, S., Min, S. Y., Miroshnichenko, A., Tran, H. K., Werfalli, R. E., Nasir, J. A., Oloni, M., Speicher, D. J., Florescu, A., Singh, B., Faltyn, M., Hernandez-Koutoucheva, A., Sharma, A. N., Bordeleau, E., Pawlowski, A. C., Zubyk, H. L., Dooley, D., Griffiths, E., Maguire, F., Winsor, G. L., Beiko, R. G., Brinkman, F. S. L., Hsiao, W. W. L., Domselaar, G. V., and McArthur, A. G. (2020) 'CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database', *Nucleic acids research*, 48(D1), pp. D517–D525. doi: 10.1093/NAR/GKZ935.

Alekseeva, M. G., Zakharevich, N. V., Ratkin, A. V., and Danilenko, V. N. (2022) 'Human Intestinal Microbiome—A Reservoir of Aminoglycoside-*N*-Acetyltransferases—Drug Resistance Genes', *Russian Journal of Genetics*, 58(9), pp. 1072–1078. doi: 10.1134/S1022795422090022/FIGURES/1.

Alexander, J., Bollmann, A., Seitz, W., and Schwartz, T. (2015) ‘Microbiological characterization of aquatic microbiomes targeting taxonomical marker genes and antibiotic resistance genes of opportunistic bacteria’, *Science of the Total Environment*, 512–513, pp. 316–325. doi: 10.1016/j.scitotenv.2015.01.046.

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., Pollard, K. S., Sakharova, E., Parks, D. H., Hugenholtz, P., Segata, N., Kyrpides, N. C., and Finn, R. D. (2020) ‘A unified catalog of 204,938 reference genomes from the human gut microbiome’, *Nature Biotechnology 2020 39:1*, 39(1), pp. 105–114. doi: 10.1038/s41587-020-0603-3.

Alou, M. T., Naud, S., Khelaifia, S., Bonnet, M., Lagier, J. C., and Raoult, D. (2021) ‘State of the Art in the Culture of the Human Microbiota: New Interests and Strategies’, *Clinical Microbiology Reviews*, 34(1), pp. 1–21. doi: 10.1128/CMR.00129-19.

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020) ‘Opportunities and challenges in long-read sequencing data analysis’, *Genome Biology 2020 21:1*, 21(1), pp. 1–16. doi: 10.1186/S13059-020-1935-5.

Aminov, R. I. (2010) ‘A Brief History of the Antibiotic Era: Lessons Learned and Challenges for the Future’, *Frontiers in Microbiology*, 1, p. 134. doi: 10.3389/fmicb.2010.00134.

Aminov, R. I. (2011) ‘Horizontal Gene Exchange in Environmental Microbiota’, *Frontiers in Microbiology*, 2(JULY). doi: 10.3389/FMICB.2011.00158.

Amir, I., Bouvet, P., Legeay, C., Gophna, U., and Weinberger, A. (2014) ‘*Eisenbergiella tayi* gen. nov., sp. nov., isolated from human blood’, *International Journal of Systematic and Evolutionary Microbiology*, 64(PART 3), pp. 907–914. doi: 10.1099/IJS.0.057331-0/CITE/REFWORKS.

Ammam, F., Meziane-cherif, D., Mengin-Lecreulx, D., Blanot, D., Patin, D., Boneca, I. G., Courvalin, P., Lambert, T., and Candela, T. (2013) ‘The functional *vanG$_{Cd}$* cluster of *Clostridium difficile* does not confer vancomycin resistance’, *Molecular Microbiology*, 89(4), pp. 612–625. doi: 10.1111/MMI.12299.

Andersen, H., Connolly, N., Bangar, H., Staat, M., Mortensen, J., Deburger, B., and Haslam, D. B. (2016) ‘Use of Shotgun Metagenome Sequencing To Detect Fecal Colonization with Multidrug-Resistant Bacteria in Children’, *Journal of Clinical Microbiology*, 54(7), pp. 1804–1813. doi: 10.1128/JCM.02638-15.

Angelakis, E., Bachar, D., Henrissat, B., Armougom, F., Audoly, G., Lagier, J. C., Robert, C., and Raoult, D. (2016) ‘Glycans affect DNA extraction and induce substantial differences in gut metagenomic studies’, *Scientific Reports 2016 6:1*, 6(1), pp. 1–8. doi: 10.1038/srep26276.

Anthony, W. E., Burnham, C. A. D., Dantas, G., and Kwon, J. H. (2021) ‘The Gut Microbiome as a Reservoir for Antimicrobial Resistance’, *The Journal of Infectious Diseases*, 223(Supplement_3), pp. S209–S213. doi: 10.1093/INFDIS/JIAA497.

Arcilla, M. S., van Hattem, J. M., Haverkate, M. R., Bootsma, M. C. J., van Genderen, P. J. J., Goorhuis, A., Grobusch, M. P., Lashof, A. M. O., Molhoek, N., Schultsz, C., Stobberingh, E. E., Verbrugh, H. A., de Jong, M. D., Melles, D. C., and Penders, J.

(2017) ‘Import and spread of extended-spectrum β-lactamase-producing Enterobacteriaceae by international travellers (COMBAT study): a prospective, multicentre cohort study’, *The Lancet Infectious Diseases*, 17(1), pp. 78–85. doi: 10.1016/S1473-3099(16)30319-X.

Arredondo-Alonso, S., Top, J., Corander, J., Willems, R. J. L., and Schürch, A. C. (2021) ‘Mode and dynamics of *vanA*-type vancomycin resistance dissemination in Dutch hospitals’, *Genome Medicine*, 13(1), pp. 1–18. doi: 10.1186/S13073-020-00825-3/FIGURES/5.

Arumugam, M., Raes, J., Pelletier, E., Paslier, D. Le, Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J. M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H. B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E. G., Wang, J., Guarner, F., Pedersen, O., de Vos, W. M., Brunak, S., Doré, J., Weissenbach, J., Ehrlich, S. D., Bork, P., Antolín, M., Artiguenave, F., Blottiere, H. M., Almeida, M., Brechot, C., Cara, C., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Foerstner, K. U., Friss, C., Guchte, M. van de, Guedon, E., Haimet, F., Huber, W., Hylckama-Vlieg, J. van, Jamet, A., Juste, C., Kaci, G., Knol, J., Kristiansen, K., Lakhdari, O., Layec, S., Roux, K. Le, Maguin, E., Mérieux, A., Minardi, R. M., M’rini, C., Muller, J., Oozeer, R., Parkhill, J., Renault, P., Rescigno, M., Sanchez, N., Sunagawa, S., Torrejon, A., Turner, K., Vandemeulebrouck, G., Varela, E., Winogradsky, Y., and Zeller, G. (2011) ‘Enterotypes of the human gut microbiome’, *Nature*, 473(7346), p. 174. doi: 10.1038/NATURE09944.

Arzese, A. R., Tomasetig, L., and Botta, G. A. (2000) ‘Detection of *tetQ* and *ermF* antibiotic resistance genes in *Prevotella* and *Porphyromonas* isolates from clinical specimens and resident microbiota of humans’, *The Journal of antimicrobial chemotherapy*, 45(5), pp. 577–582. doi: 10.1093/JAC/45.5.577.

Attebery, H. R. and Finegold, S. M. (1969) ‘Combined Screw-Cap and Rubber-Stopper Closure for Hungate Tubes (Pre-reduced Anaerobically Sterilized Roll Tubes and Liquid Media)’, *Applied Microbiology*, 18(4), pp. 558–561. doi: 10.1128/AM.18.4.558-561.1969.

Ballard, S. A., Grabsch, E. A., Johnson, P. D. R., and Grayson, M. L. (2005) ‘Comparison of Three PCR Primer Sets for Identification of *vanB* Gene Carriage in Feces and Correlation with Carriage of Vancomycin-Resistant Enterococci: Interference by *vanB*-Containing Anaerobic Bacilli’, *Antimicrobial Agents and Chemotherapy*, 49(1), p. 77. doi: 10.1128/AAC.49.1.77-81.2005.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., and Horvath, P. (2007) ‘CRISPR provides acquired resistance against viruses in prokaryotes’, *Science*, 315(5819), pp. 1709–1712. doi: 10.1126/SCIENCE.1138140/SUPPL_FILE/BARRANGOU.SOM.PDF.

Bartha, N. A., Sóki, J., Urbán, E., and Nagy, E. (2011) ‘Investigation of the prevalence of *tetQ*, *tetX* and *tetX1* genes in *Bacteroides* strains with elevated tigecycline minimum inhibitory concentrations’, *International Journal of Antimicrobial Agents*, 38(6), pp. 522–525. doi: 10.1016/j.ijantimicag.2011.07.010.

Baudry, L., Foutel-Rodier, T., Thierry, A., Koszul, R., and Marbouty, M. (2019) ‘MetaTOR: A Computational Pipeline to Recover High-Quality Metagenomic Bins From Mammalian Gut Proximity-Ligation (meta3C) Libraries’, *Frontiers in Genetics*, 10(JUL), p. 753. doi: 10.3389/fgene.2019.00753.

Baunwall, S. M. D., Lee, M. M., Eriksen, M. K., Mullish, B. H., Marchesi, J. R., Dahlerup, J. F., and Hvas, C. L. (2020) ‘Faecal microbiota transplantation for recurrent *Clostridioides difficile* infection: An updated systematic review and meta-analysis’, *EClinicalMedicine*, 29–30. doi: 10.1016/j.eclinm.2020.100642.

Beaulaurier, J., Zhu, S., Deikus, G., Mogno, I., Zhang, X. S., Davis-Richardson, A., Canepa, R., Triplett, E. W., Faith, J. J., Sebra, R., Schadt, E. E., and Fang, G. (2017) ‘Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation’, *Nature Biotechnology 2017 36:1*, 36(1), pp. 61–69. doi: 10.1038/nbt.4037.

Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A. M., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E. A., and Segata, N. (2021) ‘Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3’, *eLife*, 10. doi: 10.7554/ELIFE.65088.

Beitel, C. W., Froenicke, L., Lang, J. M., Korf, I. F., Michelmore, R. W., Eisen, J. A., and Darling, A. E. (2014) ‘Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products.’, *PeerJ*, 2, p. e415. doi: 10.7717/peerj.415.

Belton, J. M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012) ‘Hi-C: A comprehensive technique to capture the conformation of genomes’, *Methods*, 58(3), pp. 268–276. doi: 10.1016/j.ymeth.2012.05.001.

Bender, J. K., Fiedler, S., Klare, I., and Werner, G. (2015) ‘Complete Genome Sequence of the Gut Commensal and Laboratory Strain *Enterococcus faecium* 64/3’, *Genome Announcements*, 3(6). doi: 10.1128/GENOMEA.01275-15.

Bender, J. K., Cattoir, V., Hegstad, K., Sadowy, E., Coque, T. M., Westh, H., Hammerum, A. M., Schaffer, K., Burns, K., Murchan, S., Novais, C., Freitas, A. R., Peixe, L., Del Grosso, M., Pantosti, A., and Werner, G. (2018) ‘Update on prevalence and mechanisms of resistance to linezolid, tigecycline and daptomycin in enterococci in Europe: Towards a common nomenclature’, *Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy*, 40, pp. 25–39. doi: 10.1016/J.DRUP.2018.10.002.

Berglund, F., Österlund, T., Boulund, F., Marathe, N. P., Larsson, D. G. J., and Kristiansson, E. (2019) ‘Identification and reconstruction of novel antibiotic resistance genes from metagenomes’, *Microbiome*, 7(1), pp. 1–14. doi: 10.1186/S40168-019-0670-1/FIGURES/5.

Bickhart, D. M., Watson, M., Koren, S., Panke-Buisse, K., Cersosimo, L. M., Press, M. O., Van Tassell, C. P., Van Kessel, J. A. S., Haley, B. J., Kim, S. W., Heiner, C., Suen, G., Bakshy, K., Liachko, I., Sullivan, S. T., Myer, P. R., Ghurye, J., Pop, M., Weimer, P. J., Phillippy, A. M., and Smith, T. P. L. (2019) ‘Assignment of virus and antimicrobial resistance genes to microbial hosts in a complex microbial community by combined long-read assembly and proximity ligation', *Genome Biology*, 20(1), p. 153. doi: 10.1186/s13059-019-1760-x.

Bickhart, D. M., Kolmogorov, M., Tseng, E., Portik, D. M., Korobeynikov, A., Tolstoganov, I., Uritskiy, G., Liachko, I., Sullivan, S. T., Shin, S. B., Zorea, A., Andreu, V. P., Panke-Buisse, K., Medema, M. H., Mizrahi, I., Pevzner, P. A., and Smith, T. P. L. (2022) 'Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities', *Nature Biotechnology 2022 40:5*, 40(5), pp. 711–719. doi: 10.1038/s41587-021-01130-z.

van der Bij, A. K. and Pitout, J. D. D. (2012) 'The role of international travel in the worldwide spread of multiresistant Enterobacteriaceae', *Journal of Antimicrobial Chemotherapy*, 67(9), pp. 2090–2100. doi: 10.1093/jac/dks214.

Bilinski, J., Dziurzynski, M., Grzesiowski, P., Podsiadly, E., Stelmaszczyk-Emmel, A., Dzieciatkowski, T., Lis, K., Tyszka, M., Ozieranski, K., Dziewit, Ł., and Basak, G. W. (2022) 'Fresh Versus Frozen Stool for Fecal Microbiota Transplantation—Assessment by Multimethod Approach Combining Culturing, Flow Cytometry, and Next-Generation Sequencing', *Frontiers in Microbiology*, 0, p. 2106. doi: 10.3389/FMICB.2022.872735.

Binta, B. and Patel, M. (2016) 'Detection of *cfxA2*, *cfxA3*, and *cfxA6* genes in beta-lactamase producing oral anaerobes', *Journal of Applied Oral Science*, 24(2), pp. 142–147. doi: 10.1590/1678-775720150469.

Blackwell, G. A. and Hall, R. M. (2017) 'The *tet39* Determinant and the *msrE-mphE* Genes in *Acinetobacter* Plasmids Are Each Part of Discrete Modules Flanked by Inversely Oriented p*dif* (XerC-XerD) Sites', *Antimicrobial Agents and Chemotherapy*, 61(8). doi: 10.1128/AAC.00780-17.

Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O., and Piddock, L. J. V. (2015) 'Molecular mechanisms of antibiotic resistance', *Nature Reviews Microbiology*, 13(1), pp. 42–51. doi: 10.1038/nrmicro3380.

Blin, K. (2022) *NCBI Genome Downloading Scripts* (Version 0.3.1) [Computer program]. Available at: https://github.com/kblin/ncbi-genome-download.

Blow, M. J., Clark, T. A., Daum, C. G., Deutschbauer, A. M., Fomenkov, A., Fries, R., Froula, J., Kang, D. D., Malmstrom, R. R., Morgan, R. D., Posfai, J., Singh, K., Visel, A., Wetmore, K., Zhao, Z., Rubin, E. M., Korlach, J., Pennacchio, L. A., and Roberts, R. J. (2016) 'The Epigenomic Landscape of Prokaryotes', *PLOS Genetics*, 12(2), p. e1005854. doi: 10.1371/JOURNAL.PGEN.1005854.

Böcker, U., Nebe, T., Herweck, F., Holt, L., Panja, A., Jobin, C., Rossol, S., Sartor, R. B., and Singer, M. V. (2003) 'Butyrate modulates intestinal epithelial cell-mediated neutrophil migration', *Clinical and Experimental Immunology*, 131(1), pp. 53–60. doi: 10.1046/J.1365-2249.2003.02056.X.

Bortolaia, V., Kaas, R. S., Ruppe, E., Roberts, M. C., Schwarz, S., Cattoir, V., Philippon, A., Allesoe, R. L., Rebelo, A. R., Florensa, A. F., Fagelhauer, L., Chakraborty, T., Neumann, B., Werner, G., Bender, J. K., Stingl, K., Nguyen, M., Coppens, J., Xavier, B. B., Malhotra-Kumar, S., Westh, H., Pinholt, M., Anjum, M. F., Duggett, N. A., Kempf, I., Nykäsenoja, S., Olkkola, S., Wieczorek, K., Amaro, A., Clemente, L., Mossong, J., Losch, S., Ragimbeau, C., Lund, O., and Aarestrup, F. M. (2020) 'ResFinder 4.0 for predictions of phenotypes from genotypes', *Journal of Antimicrobial Chemotherapy*, 75(12), pp. 3491–3500. doi: 10.1093/JAC/DKAA345.

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., Schulz, F., Jarett, J., Rivers, A. R., Eloe-Fadrosh, E. A., Tringe, S. G., Ivanova, N. N., Copeland, A., Clum, A., Becraft, E. D., Malmstrom, R. R., Birren, B., Podar, M., Bork, P., Weinstock, G. M., Garrity, G. M., Dodsworth, J. A., Yooseph, S., Sutton, G., Glöckner, F. O., Gilbert, J. A., Nelson, W. C., Hallam, S. J., Jungbluth, S. P., Ettema, T. J. G., Tighe, S., Konstantinidis, K. T., Liu, W. T., Baker, B. J., Rattei, T., Eisen, J. A., Hedlund, B., McMahon, K. D., Fierer, N., Knight, R., Finn, R., Cochrane, G., Karsch-Mizrachi, I., Tyson, G. W., Rinke, C., Lapidus, A., Meyer, F., Yilmaz, P., Parks, D. H., Eren, A. M., Schriml, L., Banfield, J. F., Hugenholtz, P., and Woyke, T. (2017) 'Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea', *Nature Biotechnology 2017 35:8*, 35(8), pp. 725–731. doi: 10.1038/nbt.3893.

Bowler, L. D., Zhang, Q. Y., Riou, J. Y., and Spratt, B. G. (1994) 'Interspecies recombination between the *penA* genes of *Neisseria meningitidis* and commensal *Neisseria* species during the emergence of penicillin resistance in *N. meningitidis*: natural events and laboratory simulation', *Journal of Bacteriology*, 176(2), pp. 333–337. doi: 10.1128/JB.176.2.333-337.1994.

Breathnach, A. S., Cubbon, M. D., Karunaharan, R. N., Pope, C. F., and Planche, T. D. (2012) 'Multidrug-resistant *Pseudomonas aeruginosa* outbreaks in two hospitals: association with contaminated hospital waste-water systems', *Journal of Hospital Infection*, 82(1), pp. 19–24. doi: 10.1016/J.JHIN.2012.06.007.

Browne, H. P., Forster, S. C., Anonye, B. O., Kumar, N., Neville, B. A., Stares, M. D., Goulding, D., and Lawley, T. D. (2016) 'Culturing of "unculturable" human microbiota reveals novel taxa and extensive sporulation', *Nature*, 533(7604), p. 543. doi: 10.1038/NATURE17645.

Buc, E., Dubois, D., Sauvanet, P., Raisch, J., Delmas, J., Darfeuille-Michaud, A., Pezet, D., and Bonnet, R. (2013) 'High Prevalence of Mucosa-Associated *E. coli* Producing Cyclomodulin and Genotoxin in Colon Cancer', *PLoS ONE*, 8(2), p. 56964. doi: 10.1371/JOURNAL.PONE.0056964.

Buelow, E., Gonzalez, T. B., Versluis, D., Oostdijk, E. A. N., Ogilvie, L. A., van Mourik, M. S. M., Oosterink, E., van Passel, M. W. J., Smidt, H., D'Andrea, M. M., de Been, M., Jones, B. V., Willems, R. J. L., Bonten, M. J. M., and van Schaik, W. (2014) 'Effects of selective digestive decontamination (SDD) on the gut resistome', *Journal of Antimicrobial Chemotherapy*, 69(8), pp. 2215–2223. doi: 10.1093/jac/dku092.

Buelow, E., Bello González, T. d. j., Fuentes, S., de Steenhuijsen Piters, W. A. A., Lahti, L., Bayjanov, J. R., Majoor, E. A. M., Braat, J. C., van Mourik, M. S. M., Oostdijk, E. A. N., Willems, R. J. L., Bonten, M. J. M., van Passel, M. W. J., Smidt, H., and van Schaik, W. (2017) 'Comparative gut microbiota and resistome profiling of intensive care patients receiving selective digestive tract decontamination and healthy subjects', *Microbiome*, 5(1), p. 88. doi: 10.1186/s40168-017-0309-z.

Burton, J. N., Liachko, I., Dunham, M. J., and Shendure, J. (2014) ‘Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps.’, *G3 (Bethesda, Md.)*, 4(7), pp. 1339–46. doi: 10.1534/g3.114.011825.

Cabezón, E., de la Cruz, F., and Arechaga, I. (2017) ‘Conjugation Inhibitors and Their Potential Use to Prevent Dissemination of Antibiotic Resistance Genes in Bacteria.’, *Frontiers in microbiology*, 8, p. 2329. doi: 10.3389/fmicb.2017.02329.

Camacho, C., Madden, T., Tao, T., Agarwala, R., and Morgulis, A. (2008) *BLAST Command Line Applications User Manual [Internet]*, *Bethesda (MD): National Center for Biotechnology Information (US)*. Available at: https://www.ncbi.nlm.nih.gov/books/NBK279690/ (Accessed: 27 April 2020).

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009) ‘BLAST+: Architecture and applications’, *BMC Bioinformatics*, 10(1), pp. 1–9. doi: 10.1186/1471-2105-10-421/FIGURES/4.

Candela, T., Marvaud, J. C., Nguyen, T. K., and Lambert, T. (2017) ‘A *cfr*-like gene *cfr*(C) conferring linezolid resistance is common in *Clostridium difficile*’, *International Journal of Antimicrobial Agents*, 50(3), pp. 496–500. doi: 10.1016/J.IJANTIMICAG.2017.03.013.

Cao, Y., Shen, J., and Ran, Z. H. (2014) ‘Association between *Faecalibacterium prausnitzii* Reduction and Inflammatory Bowel Disease: A Meta-Analysis and Systematic Review of the Literature’, *Gastroenterology research and practice*, 2014. doi: 10.1155/2014/872725.

Carattoli, A., Zankari, E., Garciá-Fernández, A., Larsen, M. V., Lund, O., Villa, L., Aarestrup, F. M., and Hasman, H. (2014) ‘In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing’, *Antimicrobial Agents and Chemotherapy*, 58(7), pp. 3895–3903. doi: 10.1128/AAC.02412-14.

Carding, S., Verbeke, K., Vipond, D. T., Corfe, B. M., and Owen, L. J. (2015) ‘Dysbiosis of the gut microbiota in disease’, *Microbial Ecology in Health and Disease*, 26(0). doi: 10.3402/MEHD.V26.26191.

Cassini, A., Högberg, L. D., Plachouras, D., Quattrocchi, A., Hoxha, A., Simonsen, G. S., Colomb-Cotinat, M., Kretzschmar, M. E., Devleesschauwer, B., Cecchini, M., Ouakrim, D. A., Oliveira, T. C., Struelens, M. J., Suetens, C., Monnet, D. L., Burden of AMR Collaborative Group, R., Mertens, K., Struyf, T., Catry, B., Latour, K., Ivanov, I. N., Dobreva, E. G., Andraševic, A. T., Soprek, S., Budimir, A., Paphitou, N., Žemlicková, H., Olsen, S. S., Sönksen, U. W., Märtin, P., Ivanova, M., Lyytikäinen, O., Jalava, J., Coignard, B., Eckmanns, T., Sin, M. A., Haller, S., Daikos, G. L., Gikas, A., Tsiodras, S., Kontopidou, F., Tóth, Á., Hajdu, Á., Guólaugsson, Ó., Kristinsson, K. G., Murchan, S., Burns, K., Pezzotti, P., Gagliotti, C., Dumpis, U., Liuimiene, A., Perrin, M., Borg, M. A., Greeff, S. C. de, Monen, J. C., Koek, M. B., Elstrøm, P., Zabicka, D., Deptula, A., Hryniewicz, W., Caniça, M., Nogueira, P. J., Fernandes, P. A., Manageiro, V., Popescu, G. A., Serban, R. I., Schréterová, E., Litvová, S., Štefkovicová, M., Kolman, J., Klavs, I., Korošec, A., Aracil, B., Asensio, A., Pérez-Vázquez, M., Billström, H., Larsson, S., Reilly, J. S., Johnson, A., and Hopkins, S. (2019) ‘Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015: a population-level modelling analysis.', *The Lancet. Infectious diseases*, 19(1), pp. 56–66. doi: 10.1016/S1473-3099(18)30605-4.

Castanheira, M., Simner, P. J., and Bradford, P. A. (2021) 'Extended-spectrum β-lactamases: an update on their characteristics, epidemiology and detection', *JAC-Antimicrobial Resistance*, 3(3). doi: 10.1093/JACAMR/DLAB092.

Centers for Disease Control and Prevention (2013) *Antibiotic Resistance Threats in the United States, 2013*. Available at: https://www.cdc.gov/drugresistance/threat-report-2013/pdf/ar-threats-2013-508.pdf (Accessed: 14 November 2022).

Chain, E., Florey, H. W., Gardner, A. D., Heatley, N. G., Jennings, M. A., Orr-Ewing, J., and Sanders, A. G. (1940) 'PENICILLIN AS A CHEMOTHERAPEUTIC AGENT', *The Lancet*, 236(6104), pp. 226–228. doi: 10.5555/URI:PII:S0140673601087281.

Chatedaki, C., Voulgaridi, I., Kachrimanidou, M., Hrabak, J., Papagiannitsis, C. C., and Petinaki, E. (2019) 'Antimicrobial susceptibility and mechanisms of resistance of Greek *Clostridium difficile* clinical isolates', *Journal of Global Antimicrobial Resistance*, 16, pp. 53–58. doi: 10.1016/J.JGAR.2018.09.009.

Chatterjee, S., Mondal, A., Mitra, S., and Basu, S. (2017) '*Acinetobacter baumannii* transfers the *bla*$_{\mathrm{NDM-1}}$ gene via outer membrane vesicles', *Journal of Antimicrobial Chemotherapy*, 72(8), pp. 2201–2207. doi: 10.1093/JAC/DKX131.

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2022) 'GTDB-Tk v2: memory friendly classification with the Genome Taxonomy Database', *Bioinformatics (Oxford, England)*. Edited by K. Borgwardt. doi: 10.1093/BIOINFORMATICS/BTAC672.

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2020) 'GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database', *Bioinformatics*, 36(6), pp. 1925–1927. doi: 10.1093/BIOINFORMATICS/BTZ848.

Che, Y., Xia, Y., Liu, L., Li, A. D., Yang, Y., and Zhang, T. (2019) 'Mobile antibiotic resistome in wastewater treatment plants revealed by Nanopore metagenomic sequencing', *Microbiome*, 7(1), pp. 1–13. doi: 10.1186/S40168-019-0663-0/FIGURES/5.

Chen, J., Quiles-Puchalt, N., Chiang, Y. N., Bacigalupe, R., Fillol-Salom, A., Chee, M. S. J., Fitzgerald, J. R., and Penadés, J. R. (2018) 'Genome hypermobility by lateral transduction', *Science*, 362(6411), pp. 207–212. doi: 10.1126/SCIENCE.AAT5867/SUPPL_FILE/AAT5867_TABLE_S4.XLSX.

Chiang, Y. N., Penadés, J. R., and Chen, J. (2019) 'Genetic transduction by phages and chromosomal islands: The new and noncanonical', *PLOS Pathogens*, 15(8), p. e1007878. doi: 10.1371/JOURNAL.PPAT.1007878.

Clavel, T., Horz, H. P., Segata, N., and Vehreschild, M. (2022) 'Next steps after 15 stimulating years of human gut microbiome research', *Microbial Biotechnology*, 15(1), pp. 164–175. doi: 10.1111/1751-7915.13970.

Clokie, M. R. J., Millard, A. D., Letarov, A. V., and Heaphy, S. (2011) 'Phages in nature', *Bacteriophage*, 1(1), p. 31. doi: 10.4161/BACT.1.1.14942.

Cobo, F., Aliaga, L., Expósito-Ruiz, M., and Navarro-Marí, J. M. (2020) 'Anaerobic bacteraemia: A score predicting mortality', *Anaerobe*, 64, p. 102219. doi: 10.1016/J.ANAEROBE.2020.102219.

Colavecchio, A., Cadieux, B., Lo, A., and Goodridge, L. D. (2017) 'Bacteriophages contribute to the spread of antibiotic resistance genes among foodborne pathogens of the Enterobacteriaceae family - A review', *Frontiers in Microbiology*, 8(JUN), p. 1108. doi: 10.3389/FMICB.2017.01108/BIBTEX.

Connor, T. R., Loman, N. J., Thompson, S., Smith, A., Southgate, J., Poplawski, R., Bull, M. J., Richardson, E., Ismail, M., Thompson, S. E., Kitchen, C., Guest, M., Bakke, M., Sheppard, S. K., and Pallen, M. J. (2016) 'CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community', *Microbial Genomics*, 2(9), p. e000086. doi: 10.1099/MGEN.0.000086.

Coque, T. M., Baquero, F., and Canton, R. (2008) 'Increasing prevalence of ESBL-producing Enterobacteriaceae in Europe', *Eurosurveillance*, 13(47), p. 19044. doi: 10.2807/ESE.13.47.19044-EN/CITE/PLAINTEXT.

Crémazy, F. G., Rashid, F. Z. M., Haycocks, J. R., Lamberte, L. E., Grainger, D. C., and Dame, R. T. (2018) 'Determination of the 3D genome organization of bacteria using Hi-C', in *Methods in Molecular Biology*, pp. 3–18. doi: 10.1007/978-1-4939-8675-0_1.

Crhanova, M., Karasova, D., Juricova, H., Matiasovicova, J., Jahodarova, E., Kubasova, T., Seidlerova, Z., Cizek, A., and Rychlik, I. (2019) 'Systematic Culturomics Shows that Half of Chicken Caecal Microbiota Members can be Grown in Vitro Except for Two Lineages of Clostridiales and a Single Lineage of Bacteroidetes', *Microorganisms*, 7(11). doi: 10.3390/MICROORGANISMS7110496.

Crits-Christoph, A., Hallowell, H. A., Koutouvalis, K., and Suez, J. (2022) 'Good microbes, bad genes? The dissemination of antimicrobial resistance in the human microbiome', *Gut Microbes*, 14(1), p. 2055944. doi: 10.1080/19490976.2022.2055944.

Cuscó, A., Pérez, D., Viñes, J., Fàbregas, N., and Francino, O. (2021) 'Long-read metagenomics retrieves complete single-contig bacterial genomes from canine feces', *BMC Genomics*, 22(1), pp. 1–15. doi: 10.1186/S12864-021-07607-0/FIGURES/5.

Cuscó, A., Pérez, D., Viñes, J., Fàbregas, N., and Francino, O. (2022) 'Novel canine high-quality metagenome-assembled genomes, prophages and host-associated plasmids provided by long-read metagenomics together with Hi-C proximity ligation', *Microbial Genomics*, 8(3), p. 802. doi: 10.1099/MGEN.0.000802.

Dadgostar, P. (2019) 'Antimicrobial Resistance: Implications and Costs', *Infection and Drug Resistance*, 12, p. 3903. doi: 10.2147/IDR.S234610.

Darby, E. M., Trampari, E., Siasat, P., Gaya, M. S., Alav, I., Webber, M. A., and Blair, J. M. A. (2022) 'Molecular mechanisms of antibiotic resistance revisited', *Nature Reviews Microbiology 2022*, pp. 1–16. doi: 10.1038/s41579-022-00820-y.

Darfeuille-Michaud, A., Boudeau, J., Bulois, P., Neut, C., Glasser, A. L., Barnich, N., Bringer, M. A., Swidsinski, A., Beaugerie, L., and Colombel, J. F. (2004) 'High prevalence of adherent-invasive *Escherichia coli* associated with ileal mucosa in

Crohn's disease', *Gastroenterology*, 127(2), pp. 412–421. doi: 10.1053/j.gastro.2004.04.061.

Darling, A. and Liu, M. (2015) 'Metagenomic Chromosome Conformation Capture (3C): Techniques, applications, and challenges', *F1000Research*, 4. doi: 10.12688/f1000research.7281.1.

Davies, M., Galazzo, G., van Hattem, J. M., Arcilla, M. S., Melles, D. C., de Jong, M. D., Schultsz, C., Wolffs, P., McNally, A., Schaik, W. van, and Penders, J. (2022) 'Enterobacteriaceae and Bacteroidaceae provide resistance to travel-associated intestinal colonization by multi-drug resistant *Escherichia coli*', *Gut microbes*, 14(1), p. 2060676. doi: 10.1080/19490976.2022.2060676.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002) 'Capturing chromosome conformation.', *Science (New York, N.Y.)*, 295(5558), pp. 1306–11. doi: 10.1126/science.1067799.

Dekker, J. (2006) 'The three "C" s of chromosome conformation capture: controls, controls, controls', *Nature Methods*, 3(1), pp. 17–21. doi: 10.1038/nmeth823.

Dell'annunziata, F., Folliero, V., Giugliano, R., De Filippis, A., Santarcangelo, C., Izzo, V., Daglia, M., Galdiero, M., Arciola, C. R., and Franci, G. (2021) 'Gene Transfer Potential of Outer Membrane Vesicles of Gram-Negative Bacteria', *International Journal of Molecular Sciences*, 22(11). doi: 10.3390/IJMS22115985.

DeMaere, M. Z., Liu, M. Y. Z., Lin, E., Djordjevic, S. P., Charles, I. G., Worden, P., Burke, C. M., Monahan, L. G., Gardiner, M., Borody, T. J., and Darling, A. E. (2020) 'Metagenomic Hi-C of a Healthy Human Fecal Microbiome Transplant Donor', *Microbiology Resource Announcements*, 9(6), pp. e01523-19. doi: 10.1128/mra.01523-19.

Demaere, M. Z. and Darling, A. E. (2019) 'Bin3C: Exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes', *Genome Biology*, 20(1), p. 46. doi: 10.1186/s13059-019-1643-1.

DeMaere, M. Z. and Darling, A. E. (2018) 'Sim3C: Simulation of Hi-C and Meta3C proximity ligation sequencing technologies', *GigaScience*, 7(2), pp. 1–12. doi: 10.1093/gigascience/gix103.

DeMaere, M. Z. and Darling, A. E. (2021) 'qc3C: Reference-free quality control for Hi-C sequencing data', *PLOS Computational Biology*, 17(10), p. e1008839. doi: 10.1371/JOURNAL.PCBI.1008839.

Deshpande, L. M., Ashcraft, D. S., Kahn, H. P., Pankey, G., Jones, R. N., Farrell, D. J., and Mendes, R. E. (2015) 'Detection of a New *cfr*-Like Gene, *cfr*(B), in *Enterococcus faecium* Isolates Recovered from Human Specimens in the United States as Part of the SENTRY Antimicrobial Surveillance Program', *Antimicrobial Agents and Chemotherapy*, 59(10), pp. 6256–6261. doi: 10.1128/AAC.01473-15/SUPPL_FILE/ZAC010154428SO1.PDF.

Diab, M., Salem, D., El-Shenawy, A., El-Far, A., Abdelghany, A., Awad, A. R., El Defrawy, I., and Shemis, M. (2019) ‘Detection of high level aminoglycoside resistance genes among clinical isolates of *Enterococcus* species’, *Egyptian Journal of Medical Human Genetics*, 20(1), pp. 1–6. doi: 10.1186/S43042-019-0032-3/TABLES/4.

Diebold, P. J., New, F. N., Hovan, M., Satlin, M. J., and Brito, I. L. (2021) ‘Linking plasmid-based beta-lactamases to their bacterial hosts using single-cell fusion pcr’, *eLife*, 10. doi: 10.7554/ELIFE.66834.

Dijkshoorn, L., van Aken, E., Shunburne, L., van der Reijden, T. J. K., Bernards, A. T., Nemec, A., and Towner, K. J. (2005) ‘Prevalence of *Acinetobacter baumannii* and other *Acinetobacter* spp. in faecal samples from non-hospitalised individuals’, *Clinical Microbiology and Infection*, 11(4), pp. 329–332. doi: 10.1111/J.1469-0691.2005.01093.X.

DiPippo, J. L., Nesbø, C. L., Dahle, H., Doolittle, W. F., Birkland, N. K., and Noll, K. M. (2009) ‘*Kosmotoga olearia* gen. nov., sp. nov., a thermophilic, anaerobic heterotroph isolated from an oil production fluid’, *International Journal of Systematic and Evolutionary Microbiology*, 59(12), pp. 2991–3000. doi: 10.1099/ijs.0.008045-0.

Djukovic, A., Garzón, M. J., Canlet, C., Cabral, V., Lalaoui, R., García-Garcerá, M., Rechenberger, J., Tremblay-Franco, M., Peñaranda, I., Puchades-Carrasco, L., Pineda-Lucena, A., González-Barberá, E. M., Salavert, M., López-Hontangas, J. L., Sanz, M., Sanz, J., Kuster, B., Rolain, J. M., Debrauwer, L., Xavier, K. B., Xavier, J. B., and Ubeda, C. (2022) ‘*Lactobacillus* supports Clostridiales to restrict gut colonization by multidrug-resistant Enterobacteriaceae’, *Nature Communications 2022 13:1*, 13(1), pp. 1–18. doi: 10.1038/s41467-022-33313-w.

Domingo, M. C., Huletsky, A., Boissinot, M., Bernard, K. A., Picard, F. J., and Bergeron, M. G. (2008) ‘*Ruminococcus gauvreauii* sp. nov., a glycopeptide-resistant species isolated from a human faecal specimen’, *International Journal of Systematic and Evolutionary Microbiology*, 58(6), pp. 1393–1397. doi: 10.1099/ijs.0.65259-0.

Domingues, S. and Nielsen, K. M. (2017) ‘Membrane vesicles and horizontal gene transfer in prokaryotes’, *Current Opinion in Microbiology*, 38, pp. 16–21. doi: 10.1016/J.MIB.2017.03.012.

Du, Y., Laperriere, S. M., Fuhrman, J., and Sun, F. (2022) ‘Normalizing Metagenomic Hi-C Data and Detecting Spurious Contacts Using Zero-Inflated Negative Binomial Regression’, *Journal of computational biology : a journal of computational molecular cell biology*, 29(2), pp. 106–120. doi: 10.1089/CMB.2021.0439.

Du, Y. and Sun, F. (2022) ‘HiCBin: binning metagenomic contigs and recovering metagenome-assembled genomes using Hi-C contact maps’, *Genome Biology*, 23(1), p. 63. doi: 10.1186/S13059-022-02626-W.

Dubos, R. J. and Schaedler, R. W. (1960) ‘The effect of the intestinal flora on the growth rate of mice, and on their susceptibility to experimental infections’, *Journal of Experimental Medicine*, 111(3), pp. 407–417. doi: 10.1084/JEM.111.3.407.

Ducarmon, Q. R., Zwittink, R. D., Hornung, B. V. H., Schaik, W. van, Young, V. B., and Kuijper, E. J. (2019) 'Gut Microbiota and Colonization Resistance against Bacterial Enteric Infection', *Microbiology and Molecular Biology Reviews : MMBR*, 83(3). doi: 10.1128/MMBR.00007-19.

Duggett, N. A. (2016) *High-throughput sequencing of the chicken gut microbiome*.

Durack, J. and Lynch, S. V. (2019) 'The gut microbiome: Relationships with disease and opportunities for therapy', *Journal of Experimental Medicine*, 216(1), pp. 20–40. doi: 10.1084/jem.20180448.

Duranti, S., Lugli, G. A., Mancabelli, L., Turroni, F., Milani, C., Mangifesta, M., Ferrario, C., Anzalone, R., Viappiani, A., van Sinderen, D., and Ventura, M. (2017) 'Prevalence of Antibiotic Resistance Genes among Human Gut-Derived Bifidobacteria', *Applied and environmental microbiology*, 83(3). doi: 10.1128/AEM.02894-16.

Ebmeyer, S., Kristiansson, E., and Larsson, D. G. J. (2021) 'A framework for identifying the recent origins of mobile antibiotic resistance genes', *Communications Biology*, 4(1). doi: 10.1038/S42003-020-01545-5.

ECDC (2009) *Joint report with EMEA: The bacterial challenge: Time to react*. Available at: http://ecdc.europa.eu/en/publications/Publications/0909_TER_The_Bacterial_Challenge_Time_to_React.pdf (Accessed: 30 March 2019).

Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S. R., Nelson, K. E., and Relman, D. A. (2005) 'Diversity of the human intestinal microbial flora', *Science*, 308(5728), pp. 1635–1638. doi: 10.1126/science.1110591.

Edwards, A. N., Suárez, J. M., and McBride, S. M. (2013) 'Culturing and Maintaining *Clostridium difficile* in an Anaerobic Environment', *Journal of Visualized Experiments : JoVE*, (79), p. 50787. doi: 10.3791/50787.

Egan, S. A., Shore, A. C., O'Connell, B., Brennan, G. I., and Coleman, D. C. (2020) 'Linezolid resistance in *Enterococcus faecium* and *Enterococcus faecalis* from hospitalized patients in Ireland: high prevalence of the MDR genes *optrA* and *poxtA* in isolates with diverse genetic backgrounds', *Journal of Antimicrobial Chemotherapy*, 75(7), pp. 1704–1711. doi: 10.1093/JAC/DKAA075.

Ellabaan, M. M. H., Munck, C., Porse, A., Imamovic, L., and Sommer, M. O. A. (2021) 'Forecasting the dissemination of antibiotic resistance genes across bacterial genomes', *Nature communications*, 12(1). doi: 10.1038/S41467-021-22757-1.

Exner, M., Bhattacharya, S., Christiansen, B., Gebel, J., Goroncy-Bermes, P., Hartemann, P., Heeg, P., Ilschner, C., Kramer, A., Larson, E., Merkens, W., Mielke, M., Oltmanns, P., Ross, B., Rotter, M., Schmithausen, R. M., Sonntag, H.-G., and Trautmann, M. (2017) 'Antibiotic resistance: What is so special about multidrug-resistant Gram-negative bacteria?', *GMS hygiene and infection control*, 12, p. Doc05. doi: 10.3205/dgkh000290.

Feng, J., Li, B., Jiang, X., Yang, Y., Wells, G. F., Zhang, T., and Li, X. (2018) 'Antibiotic resistome in a large-scale healthy human gut microbiota deciphered by metagenomic and network analyses', *Environmental Microbiology*, 20(1), pp. 355–368. doi: 10.1111/1462-2920.14009.

Figueras, M. J., Beaz-Hidalgo, R., Hossain, M. J., and Liles, M. R. (2014) 'Taxonomic Affiliation of New Genomes Should Be Verified Using Average Nucleotide Identity and Multilocus Phylogenetic Analysis', *Genome Announcements*, 2(6). doi: 10.1128/GENOMEA.00927-14.

Fleming, A. (1929) 'On the Antibacterial Action of Cultures of a *Penicillium*, with Special Reference to their Use in the Isolation of B. influenzæ', *British journal of experimental pathology*, 10(3), p. 226. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2048009/ (Accessed: 31 October 2022).

Forslund, K., Sunagawa, S., Kultima, J. R., Mende, D. R., Arumugam, M., Typas, A., and Bork, P. (2013) 'Country-specific antibiotic use practices impact the human gut resistome', *Genome Research*, 23(7), pp. 1163–1169. doi: 10.1101/gr.155465.113.

Forster, S. C., Liu, J., Kumar, N., Gulliver, E. L., Gould, J. A., Escobar-Zepeda, A., Mkandawire, T., Pike, L. J., Shao, Y., Stares, M. D., Browne, H. P., Neville, B. A., and Lawley, T. D. (2022) 'Strain-level characterization of broad host range mobile genetic elements transferring antibiotic resistance from the human microbiome', *Nature Communications*, 13(1). doi: 10.1038/S41467-022-29096-9.

Foutel-Rodier, T., Thierry, A., Koszul, R., and Marbouty, M. (2018) 'Generation of a Metagenomics Proximity Ligation 3C Library of a Mammalian Gut Microbiota', in *Methods in Enzymology*, pp. 183–195. doi: 10.1016/bs.mie.2018.08.001.

Frank, D. N., St. Amand, A. L., Feldman, R. A., Boedeker, E. C., Harpaz, N., and Pace, N. R. (2007) 'Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases', *Proceedings of the National Academy of Sciences of the United States of America*, 104(34), p. 13780. doi: 10.1073/PNAS.0706625104.

Frank, J. A., Reich, C. I., Sharma, S., Weisbaum, J. S., Wilson, B. A., and Olsen, G. J. (2008) 'Critical Evaluation of Two Primers Commonly Used for Amplification of Bacterial 16S rRNA Genes', *Applied and Environmental Microbiology*, 74(8), p. 2461. doi: 10.1128/AEM.02272-07.

Frost, L. S., Leplae, R., Summers, A. O., and Toussaint, A. (2005) 'Mobile genetic elements: the agents of open source evolution', *Nature Reviews Microbiology 2005 3:9*, 3(9), pp. 722–732. doi: 10.1038/nrmicro1235.

Fukuda, S., Toh, H., Hase, K., Oshima, K., Nakanishi, Y., Yoshimura, K., Tobe, T., Clarke, J. M., Topping, D. L., Suzuki, T., Taylor, T. D., Itoh, K., Kikuchi, J., Morita, H., Hattori, M., and Ohno, H. (2011) 'Bifidobacteria can protect from enteropathogenic infection through production of acetate', *Nature 2011 469:7331*, 469(7331), pp. 543–547. doi: 10.1038/nature09646.

García-López, M., Meier-Kolthoff, J. P., Tindall, B. J., Gronow, S., Woyke, T., Kyrpides, N. C., Hahnke, R. L., and Göker, M. (2019) 'Analysis of 1,000 Type-Strain Genomes Improves Taxonomic Classification of Bacteroidetes', *Frontiers in Microbiology*, 10, p. 2083. doi: 10.3389/FMICB.2019.02083/BIBTEX.

García, N., Gutiérrez, G., Lorenzo, M., García, J. E., Píriz, S., and Quesada, A. (2008) 'Genetic determinants for *cfxA* expression in *Bacteroides* strains isolated from human infections', *Journal of Antimicrobial Chemotherapy*, 62(5), pp. 942–947. doi: 10.1093/JAC/DKN347.

Gardiner, B. J., Tai, A. Y., Kotsanas, D., Francis, M. J., Roberts, S. A., Ballard, S. A., Junckerstorff, R. K., and Kormana, T. M. (2015) 'Clinical and Microbiological Characteristics of *Eggerthella lenta* Bacteremia', *Journal of Clinical Microbiology*, 53(2), p. 626. doi: 10.1128/JCM.02926-14.

Garmendia, L., Hernandez, A., Sanchez, M. B., and Martinez, J. L. (2012) 'Metagenomics and antibiotics', *Clinical Microbiology and Infection*, 18(SUPPL. 4), pp. 27–31. doi: 10.1111/J.1469-0691.2012.03868.X.

Gaynes, R. (2017) 'The Discovery of Penicillin—New Insights After More Than 75 Years of Clinical Use', *Emerging Infectious Diseases*, 23(5), p. 849. doi: 10.3201/EID2305.161556.

Gill, S. R., Pop, M., DeBoy, R. T., Eckburg, P. B., Turnbaugh, P. J., Samuel, B. S., Gordon, J. I., Relman, D. A., Fraser-Liggett, C. M., and Nelson, K. E. (2006) 'Metagenomic analysis of the human distal gut microbiome', *Science*, 312(5778), pp. 1355–1359. doi: 10.1126/SCIENCE.1124234/SUPPL_FILE/GILL.SOM.PDF.

Gilroy, R., Ravi, A., Getino, M., Pursley, I., Horton, D. L., Alikhan, N. F., Baker, D., Gharbi, K., Hall, N., Watson, M., Adriaenssens, E. M., Foster-Nyarko, E., Jarju, S., Secka, A., Antonio, M., Oren, A., Chaudhuri, R. R., Ragione, R. La, Hildebrand, F., and Pallen, M. J. (2021) 'Extensive microbial diversity within the chicken gut microbiome revealed by metagenomics and culture', *PeerJ*, 9. doi: 10.7717/PEERJ.10941/SUPP-2.

Gloor, G. B., Hummelen, R., Macklaim, J. M., Dickson, R. J., Fernandes, A. D., MacPhee, R., and Reid, G. (2010) 'Microbiome Profiling by Illumina Sequencing of Combinatorial Sequence-Tagged PCR Products', *PLoS ONE*. Edited by F. R. DeLeo, 5(10), p. e15406. doi: 10.1371/journal.pone.0015406.

Gonçalves, O. S., Campos, K. F., de Assis, J. C. S., Fernandes, A. S., Souza, T. S., Rodrigues, L. G. D. C., de Queiroz, M. V., and Santana, M. F. (2020) 'Transposable elements contribute to the genome plasticity of *Ralstonia solanacearum* species complex', *Microbial Genomics*, 6(5), pp. 1–12. doi: 10.1099/mgen.0.000374.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016) 'Coming of age: ten years of next-generation sequencing technologies', *Nature Reviews Genetics 2016 17:6*, 17(6), pp. 333–351. doi: 10.1038/nrg.2016.49.

Goren, M. G., Carmeli, Y., Schwaber, M. J., Chmelnitsky, I., Schechner, V., and Navon-Venezia, S. (2010) 'Transfer of Carbapenem-Resistant Plasmid from *Klebsiella pneumoniae* ST258 to *Escherichia coli* in Patient', *Emerging Infectious Diseases*, 16(6), p. 1014. doi: 10.3201/EID1606.091671.

Gorrie, C. L., Mirc Eta, M., Wick, R. R., Edwards, D. J., Thomson, N. R., Strugnell, R. A., Pratt, N. F., Garlick, J. S., Watson, K. M., Pilcher, D. V., McGloughlin, S. A., Spelman, D. W., Jenney, A. W. J., and Holt, K. E. (2017) 'Gastrointestinal Carriage Is a Major Reservoir of *Klebsiella pneumoniae* Infection in Intensive Care Patients', *Clinical Infectious Diseases*, 65(2), pp. 208–215. doi: 10.1093/CID/CIX270.

Gotoh, A., Nara, M., Sugiyama, Y., Sakanaka, M., Yachi, H., Kitakata, A., Nakagawa, A., Minami, H., Okuda, S., Katoh, T., Katayama, T., and Kurihara, S. (2017) 'Use of Gifu Anaerobic Medium for culturing 32 dominant species of human gut microbes and its evaluation based on short-chain fatty acids fermentation profiles', *Bioscience, Biotechnology, and Biochemistry*, 81(10), pp. 2009–2017. doi: 10.1080/09168451.2017.1359486.

Gould, K. (2016) 'Antibiotics: from prehistory to the present day', *Journal of Antimicrobial Chemotherapy*, 71(3), pp. 572–575. doi: 10.1093/JAC/DKV484.

Graham, M., Ballard, S. A., Grabsch, E. A., Johnson, P. D. R., and Grayson, M. L. (2008) 'High Rates of Fecal Carriage of Nonenterococcal vanB in both Children and Adults', *Antimicrobial Agents and Chemotherapy*, 52(3), p. 1195. doi: 10.1128/AAC.00531-07.

GraphPad Software (2022) *GraphPad Prism* (Version 9.4.1) [Computer program]. Available at: www.graphpad.com.

Gross, R. H. (1990) 'The Gene Construction Kit: a new computer program for manipulating and presenting DNA constructs', *BioTechniques*, 8(6), pp. 684–689. Available at: https://pubmed.ncbi.nlm.nih.gov/2357385/ (Accessed: 20 October 2022).

Guédon, G., Libante, V., Coluzzi, C., Payot, S., and Leblond-Bourget, N. (2017) 'The Obscure World of Integrative and Mobilizable Elements, Highly Widespread Elements that Pirate Bacterial Conjugative Systems', *Genes*, 8(11). doi: 10.3390/GENES8110337.

Gueimonde, M., Salminen, S., and Isolauri, E. (2006) 'Presence of specific antibiotic (*tet*) resistance genes in infant faecal microbiota', *FEMS Immunology & Medical Microbiology*, 48(1), pp. 21–25. doi: 10.1111/j.1574-695X.2006.00112.x.

Guiney, D. G. and Davis, C. E. (1978) 'Identification of a conjugative R plasmid in *Bacteroides ochraceus* capable of transfer to *Escherichia coli*', *Nature*, 274(5667), pp. 181–182. doi: 10.1038/274181A0.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013) 'QUAST: quality assessment tool for genome assemblies', *Bioinformatics*, 29(8), pp. 1072–1075. doi: 10.1093/BIOINFORMATICS/BTT086.

Guzman Prieto, A. M., van Schaik, W., Rogers, M. R. C., Coque, T. M., Baquero, F., Corander, J., and Willems, R. J. L. (2016) 'Global Emergence and Dissemination of Enterococci as Nosocomial Pathogens: Attack of the Clones?', *Frontiers in microbiology*, 7, p. 788. doi: 10.3389/fmicb.2016.00788.

Haas, K. N. and Blanchard, J. L. (2020) 'Reclassification of the *Clostridium clostridioforme* and *Clostridium sphenoides* clades as *Enterocloster* gen. nov. and *Lacrimispora* gen. nov., including reclassification of 15 taxa', *International Journal of Systematic and Evolutionary Microbiology*, 70(1), pp. 23–34. doi: 10.1099/IJSEM.0.003698/CITE/REFWORKS.

Haas, L. (1999) 'Papyrus of Ebers and Smith', *Journal of Neurology, Neurosurgery & Psychiatry*, 67(5), pp. 578–578. doi: 10.1136/JNNP.67.5.578.

Hansen, L. H. and Vester, B. (2015) 'A *cfr*-Like Gene from *Clostridium difficile* Confers Multiple Antibiotic Resistance by the Same Mechanism as the *cfr* Gene', *Antimicrobial Agents and Chemotherapy*, 59(9), pp. 5841–5843. doi: 10.1128/AAC.01274-15/ASSET/FD3C4A4A-BAB1-4547-A101-F89C241AADE8/ASSETS/GRAPHIC/ZAC0091543540002.JPEG.

Harrison, E. and Brockhurst, M. A. (2012) 'Plasmid-mediated horizontal gene transfer is a coevolutionary process', *Trends in Microbiology*, 20(6), pp. 262–267. doi: 10.1016/J.TIM.2012.04.003.

Hashimoto, Y., Hisatsune, J., Suzuki, M., Kurushima, J., Nomura, T., Hirakawa, H., Kojima, N., Ono, Y., Hasegawa, Y., Tanimoto, K., Sugai, M., and Tomita, H. (2022) 'Elucidation of host diversity of the VanD-carrying genomic islands in enterococci and anaerobes', *JAC-antimicrobial resistance*, 4(1). doi: 10.1093/JACAMR/DLAB189.

Hensgens, M. P. M., Goorhuis, A., Dekkers, O. M., and Kuijper, E. J. (2012) 'Time interval of increased risk for *Clostridium difficile* infection after exposure to antibiotics', *Journal of Antimicrobial Chemotherapy*, 67(3), pp. 742–748. doi: 10.1093/JAC/DKR508.

Hitch, T. C. A., Afrizal, A., Riedel, T., Kioukis, A., Haller, D., Lagkouvardos, I., Overmann, J., and Clavel, T. (2021) 'Recent advances in culture-based gut microbiome research', *International Journal of Medical Microbiology*, 311(3), p. 151485. doi: 10.1016/J.IJMM.2021.151485.

Ho, J., Yeoh, Y. K., Barua, N., Chen, Z., Lui, G., Wong, S. H., Yang, X., Chan, M. C. W., Chan, P. K. S., Hawkey, P. M., and Ip, M. (2020) 'Systematic review of human gut resistome studies revealed variable definitions and approaches', *Gut Microbes*, 12(1). doi: 10.1080/19490976.2019.1700755/SUPPL_FILE/KGMI_A_1700755_SM7281.DOCX.

Hobro, A. J. and Smith, N. I. (2017) 'An evaluation of fixation methods: Spatial and compositional cellular changes observed by Raman imaging', *Vibrational Spectroscopy*, 91, pp. 31–45. doi: 10.1016/j.vibspec.2016.10.012.

Howden, B. P., Holt, K. E., Lam, M. M. C., Seemann, T., Ballard, S., Coombs, G. W., Tong, S. Y. C., Lindsay Grayson, M., Johnson, P. D. R., and Stinear, T. P. (2013) ‘Genomic Insights to Control the Emergence of Vancomycin-Resistant Enterococci’, *mBio*, 4(4), pp. 412–425. doi: 10.1128/MBIO.00412-13.

Hu, Y., Yang, X., Qin, J., Lu, N., Cheng, G., Wu, N., Pan, Y., Li, J., Zhu, L., Wang, X., Meng, Z., Zhao, F., Liu, D., Ma, J., Qin, N., Xiang, C., Xiao, Y., Li, L., Yang, H., Wang, Jian, Yang, R., Gao, G. F., Wang, Jun, and Zhu, B. (2013) ‘Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota’, *Nature Communications*, 4(1), p. 2151. doi: 10.1038/ncomms3151.

Huddleston, J. R. (2014) ‘Horizontal gene transfer in the human gastrointestinal tract: potential spread of antibiotic resistance genes.’, *Infection and drug resistance*, 7, pp. 167–76. doi: 10.2147/IDR.S48820.

Hultman, J., Tamminen, M., Pärnänen, K., Cairns, J., Karkman, A., and Virta, M. (2018) ‘Host range of antibiotic resistance genes in wastewater treatment plant influent and effluent’, *FEMS Microbiology Ecology*, 94(4). doi: 10.1093/femsec/fiy038.

Husain, F., Veeranagouda, Y., Boente, R., Tang, K., Mulato, G., and Wexler, H. M. (2014) ‘The Ellis Island Effect’, *http://dx.doi.org/10.4161/mge.29801*, 4(4), p. e29801. doi: 10.4161/MGE.29801.

Husain, F., Tang, K., Veeranagouda, Y., Boente, R., Patrick, S., Blakely, G., and Wexler, H. M. (2017) ‘Novel large-scale chromosomal transfer in *Bacteroides fragilis* contributes to its pan-genome and rapid environmental adaptation’, *Microbial Genomics*, 3(11), p. e000136. doi: 10.1099/MGEN.0.000136/CITE/REFWORKS.

Hutchings, M., Truman, A., and Wilkinson, B. (2019) ‘Antibiotics: past, present and future’, *Current Opinion in Microbiology*, 51, pp. 72–80. doi: 10.1016/J.MIB.2019.10.008.

Imelfort, M., Parks, D., Woodcroft, B. J., Dennis, P., Hugenholtz, P., and Tyson, G. W. (2014) ‘GroopM: An automated tool for the recovery of population genomes from related metagenomes’, *PeerJ*, 2014(1). doi: 10.7717/PEERJ.603/SUPP-2.

Ivanova, V., Chernevskaya, E., Vasiluev, P., Ivanov, A., Tolstoganov, I., Shafranskaya, D., Ulyantsev, V., Korobeynikov, A., Razin, S. V., Beloborodova, N., Ulianov, S. V., and Tyakht, A. (2022) ‘Hi-C Metagenomics in the ICU: Exploring Clinically Relevant Features of Gut Microbiome in Chronically Critically Ill Patients’, *Frontiers in Microbiology*, 12, p. 3935. doi: 10.3389/FMICB.2021.770323/BIBTEX.

Jacobsen, L., Wilcks, A., Hammer, K., Huys, G., Gevers, D., and Andersen, S. R. (2007) ‘Horizontal transfer of *tet*(M) and *erm*(B) resistance plasmids from food strains of *Lactobacillus plantarum* to *Enterococcus faecalis* JH2-2 in the gastrointestinal tract of gnotobiotic rats’, *FEMS Microbiology Ecology*, 59(1), pp. 158–166. doi: 10.1111/J.1574-6941.2006.00212.X.

Jain, A. and Srivastava, P. (2013) ‘Broad host range plasmids’, *FEMS Microbiology Letters*, 348(2), pp. 87–96. doi: 10.1111/1574-6968.12241.

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018) ‘High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries’, *Nature Communications 2018 9:1*, 9(1), pp. 1–8. doi: 10.1038/s41467-018-07641-9.

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O’Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., Snutch, T. P., Tee, L., Paten, B., Phillippy, A. M., Simpson, J. T., Loman, N. J., and Loose, M. (2018) ‘Nanopore sequencing and assembly of a human genome with ultra-long reads’, *Nature biotechnology*, 36(4), pp. 338–345. doi: 10.1038/NBT.4060.

Jakobsson, H. E., Jernberg, C., Andersson, A. F., Sjölund-Karlsson, M., Jansson, J. K., and Engstrand, L. (2010) ‘Short-Term Antibiotic Treatment Has Differing Long-Term Impacts on the Human Throat and Gut Microbiome’, *PLoS ONE*, 5(3), p. 9836. doi: 10.1371/JOURNAL.PONE.0009836.

Janda, J. M. and Abbott, S. L. (2007) ‘16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls’, *Journal of Clinical Microbiology*, 45(9), pp. 2761–2764. doi: 10.1128/JCM.01228-07.

Jandhyala, S. M., Talukdar, R., Subramanyam, C., Vuyyuru, H., Sasikala, M., and Nageshwar Reddy, D. (2015) ‘Role of the normal gut microbiota.’, *World journal of gastroenterology*, 21(29), pp. 8787–803. doi: 10.3748/wjg.v21.i29.8787.

Janoir, C., Podglajen, I., Kitzis, M. D., Poyart, C., and Gutmann, L. (1999) ‘In Vitro Exchange of Fluoroquinolone Resistance Determinants between *Streptococcus pneumoniae* and Viridans Streptococci and Genomic Organization of the *parE-parC* Region in *S. mitis*’, *The Journal of Infectious Diseases*, 180(2), pp. 555–558. doi: 10.1086/314888.

Janssen, A. B., Bartholomew, T. L., Marciszewska, N. P., Bonten, M. J. M., Willems, R. J. L., Bengoechea, J. A., and Van Schaik, W. (2020) ‘Nonclonal Emergence of Colistin Resistance Associated with Mutations in the BasRS Two-Component System in *Escherichia coli* Bloodstream Isolates’, *mSphere*, 5(2), pp. e00143-20. doi: 10.1128/mSphere.00143-20.

Jarvis, R. A. and Patrick, E. A. (1973) ‘Clustering Using a Similarity Measure Based on Shared Near Neighbors’, *IEEE Transactions on Computers*, C–22(11), pp. 1025–1034. doi: 10.1109/T-C.1973.223640.

Jasemi, S., Emaneini, M., Ahmadinejad, Z., Fazeli, M. S., Sechi, L. A., Sadeghpour Heravi, F., and Feizabadi, M. M. (2021) ‘Antibiotic resistance pattern of *Bacteroides fragilis* isolated from clinical and colorectal specimens’, *Annals of clinical microbiology and antimicrobials*, 20(1). doi: 10.1186/S12941-021-00435-W.

Jiang, X., Hall, A. B., Xavier, R. J., and Alm, E. J. (2019) ‘Comprehensive analysis of chromosomal mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools’, *PloS one*, 14(12). doi: 10.1371/JOURNAL.PONE.0223680.

Jin, H., You, L., Zhao, F., Li, S., Ma, T., Kwok, L. Y., Xu, H., and Sun, Z. (2022) ‘Hybrid, ultra-deep metagenomic sequencing enables genomic and functional characterization of low-abundance species in the human gut microbiome’, *Gut Microbes*, 14(1). doi: 10.1080/19490976.2021.2021790.

Johnson, J. S., Spakowicz, D. J., Hong, B. Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., and Weinstock, G. M. (2019) ‘Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis’, *Nature Communications 2019 10:1*, 10(1), pp. 1–11. doi: 10.1038/s41467-019-13036-1.

Johnston, C., Martin, B., Fichant, G., Polard, P., and Claverys, J.-P. (2014) ‘Bacterial transformation: distribution, shared mechanisms and divergent control’, *Nature Reviews Microbiology*, 12(3), pp. 181–196. doi: 10.1038/nrmicro3199.

Jolley, K. A. and Maiden, M. C. J. (2010) ‘BIGSdb: Scalable analysis of bacterial genome variation at the population level’, *BMC Bioinformatics*, 11(1), pp. 1–11. doi: 10.1186/1471-2105-11-595/FIGURES/4.

Jones, B. V., Sun, F., and Marchesi, J. R. (2010) ‘Comparative metagenomic analysis of plasmid encoded functions in the human gut microbiome’, *BMC Genomics*, 11(1), p. 46. doi: 10.1186/1471-2164-11-46.

Kalmar, L., Gupta, S., Kean, I. R. L., Ba, X., Hadjirin, N., Lay, E. M., de Vries, S. P. W., Bateman, M., Bartlet, H., Hernandez-Garcia, J., Tucker, A. W., Restif, O., Stevens, M. P., Wood, J. L. N., Maskell, D. J., Grant, A. J., and Holmes, M. A. (2022) ‘HAM-ART: An optimised culture-free Hi-C metagenomics pipeline for tracking antimicrobial resistance genes in complex microbial communities’, *PLOS Genetics*, 18(3), p. e1009776. doi: 10.1371/JOURNAL.PGEN.1009776.

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019) ‘MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies’, *PeerJ*, 2019(7). doi: 10.7717/PEERJ.7359/SUPP-3.

Kao, S. J., You, I., Clewell, D. B., Donabedian, S. M., Zervos, M. J., Petrin, J., Shaw, K. J., and Chow, J. W. (2000) ‘Detection of the High-Level Aminoglycoside Resistance Gene *aph(2")-Ib* in *Enterococcus faecium*’, *Antimicrobial Agents and Chemotherapy*, 44(10), p. 2876. doi: 10.1128/AAC.44.10.2876-2879.2000.

Karcher, N., Pasolli, E., Asnicar, F., Huang, K. D., Tett, A., Manara, S., Armanini, F., Bain, D., Duncan, S. H., Louis, P., Zolfo, M., Manghi, P., Valles-Colomer, M., Raffaetà, R., Rota-Stabelli, O., Collado, M. C., Zeller, G., Falush, D., Maixner, F., Walker, A. W., Huttenhower, C., and Segata, N. (2020) ‘Analysis of 1321 *Eubacterium rectale* genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations’, *Genome Biology*, 21(1), pp. 1–27. doi: 10.1186/S13059-020-02042-Y/FIGURES/6.

Kechagias, K. S., Chorepsima, S., Triarides, N. A., and Falagas, M. E. (2020) ‘Tigecycline for the treatment of patients with *Clostridium difficile* infection: an update of the clinical evidence’, *European Journal of Clinical Microbiology and Infectious Diseases*, 39(6), pp. 1053–1058. doi: 10.1007/S10096-019-03756-Z/FIGURES/2.

Kent, A. G., Vill, A. C., Shi, Q., Satlin, M. J., and Brito, I. L. (2020) 'Widespread transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial Hi-C', *Nature Communications*, 11(1), pp. 1–9. doi: 10.1038/s41467-020-18164-7.

Kierzkowska, M., Majewska, A., and Mlynarczyk, G. (2020) 'Trends and Impact in Antimicrobial Resistance Among *Bacteroides* and *Parabacteroides* Species in 2007–2012 Compared to 2013–2017', *https://home.liebertpub.com/mdr*, 26(12), pp. 1452–1457. doi: 10.1089/MDR.2019.0462.

Kizny Gordon, A. E., Mathers, A. J., Cheong, E. Y. L., Gottlieb, T., Kotay, S., Walker, A. S., Peto, T. E. A., Crook, D. W., and Stoesser, N. (2017) 'The Hospital Water Environment as a Reservoir for Carbapenem-Resistant Organisms Causing Hospital-Acquired Infections—A Systematic Review of the Literature', *Clinical Infectious Diseases*, 64(10), pp. 1435–1444. doi: 10.1093/cid/cix132.

Koessler, D. R., Knisley, D. J., Knisley, J., and Haynes, T. (2010) 'Assembly complexity of prokaryotic genomes using short reads.', *BMC Bioinformatics*, 11(SUPPL. 6), pp. 21–21. doi: 10.1186/1471-2105-11-21.

Kolde, R. (2019) *pheatmap: Pretty Heatmaps* (R package version 1.0.12) [Computer program]. Available at: https://cran.r-project.org/web/packages/pheatmap/.

Krauth, S. J., Coulibaly, J. T., Knopp, S., Traoré, M., N'Goran, E. K., and Utzinger, J. (2012) 'An In-Depth Analysis of a Piece of Shit: Distribution of *Schistosoma mansoni* and Hookworm Eggs in Human Stool', *PLoS Neglected Tropical Diseases*, 6(12), p. e1969. doi: 10.1371/journal.pntd.0001969.

Lagier, J. C., Armougom, F., Million, M., Hugon, P., Pagnier, I., Robert, C., Bittar, F., Fournous, G., Gimenez, G., Maraninchi, M., Trape, J. F., Koonin, E. V., La Scola, B., and Raoult, D. (2012) 'Microbial culturomics: paradigm shift in the human gut microbiome study', *Clinical Microbiology and Infection*, 18(12), pp. 1185–1193. doi: 10.1111/1469-0691.12023.

Lagier, J. C., Khelaifia, S., Alou, M. T., Ndongo, S., Dione, N., Hugon, P., Caputo, A., Cadoret, F., Traore, S. I., Seck, E. H., Dubourg, G., Durand, G., Mourembou, G., Guilhot, E., Togo, A., Bellali, S., Bachar, D., Cassir, N., Bittar, F., Delerce, J., Mailhe, M., Ricaboni, D., Bilen, M., Dangui Nieko, N. P. M., Dia Badiane, N. M., Valles, C., Mouelhi, D., Diop, K., Million, M., Musso, D., Abrahão, J., Azhar, E. I., Bibi, F., Yasir, M., Diallo, A., Sokhna, C., Djossou, F., Vitton, V., Robert, C., Rolain, J. M., La Scola, B., Fournier, P. E., Levasseur, A., and Raoult, D. (2016) 'Culture of previously uncultured members of the human gut microbiota by culturomics', *Nature Microbiology 2016 1:12*, 1(12), pp. 1–8. doi: 10.1038/nmicrobiol.2016.203.

Lagier, J. C., Dubourg, G., Million, M., Cadoret, F., Bilen, M., Fenollar, F., Levasseur, A., Rolain, J. M., Fournier, P. E., and Raoult, D. (2018) 'Culturing the human microbiota and culturomics', *Nature Reviews Microbiology 2018 16:9*, 16(9), pp. 540–550. doi: 10.1038/s41579-018-0041-0.

Lamberte, L. E. and van Schaik, W. (2022) 'Antibiotic resistance in the commensal human gut microbiota', *Current Opinion in Microbiology*, 68, p. 102150. doi: 10.1016/J.MIB.2022.102150.

Lane, D. J. (1991) ‘16S/23S rRNA sequencing’, in Stackebrandt, E. and Goodfellow, M. (eds) *Nucleic acid techniques in bacterial systematics*, pp. 115–175.

Langmead, B. and Salzberg, S. L. (2012) ‘Fast gapped-read alignment with Bowtie 2’, *Nature Methods*, 9(4), pp. 357–359. doi: 10.1038/nmeth.1923.

Lanza, V. F., Baquero, F., Martínez, J. L., Ramos-Ruíz, R., González-Zorn, B., Andremont, A., Sánchez-Valenzuela, A., Ehrlich, S. D., Kennedy, S., Ruppé, E., van Schaik, W., Willems, R. J., de la Cruz, F., and Coque, T. M. (2018) ‘In-depth resistome analysis by targeted metagenomics’, *Microbiome*, 6(1). doi: 10.1186/S40168-017-0387-Y.

Lapidus, A. L. and Korobeynikov, A. I. (2021) ‘Metagenomic Data Assembly – The Way of Decoding Unknown Microorganisms’, *Frontiers in Microbiology*, 12, p. 653. doi: 10.3389/FMICB.2021.613791/BIBTEX.

Lau, J. T., Whelan, F. J., Herath, I., Lee, C. H., Collins, S. M., Bercik, P., and Surette, M. G. (2016) ‘Capturing the diversity of the human gut microbiota through culture-enriched molecular profiling’, *Genome Medicine*, 8(1), pp. 1–10. doi: 10.1186/S13073-016-0327-7/FIGURES/3.

Launay, A., Ballard, S. A., Johnson, P. D. R., Grayson, M. L., and Lambert, T. (2006) ‘Transfer of vancomycin resistance transposon Tn*1549* from *Clostridium symbiosum* to *Enterococcus* spp. in the gut of gnotobiotic mice’, *Antimicrobial agents and chemotherapy*, 50(3), pp. 1054–1062. doi: 10.1128/AAC.50.3.1054-1062.2006.

Layton, B. A., Walters, S. P., Lam, L. H., and Boehm, A. B. (2010) ‘*Enterococcus* species distribution among human and animal hosts using multiplex PCR’, *Journal of Applied Microbiology*, 109(2), pp. 539–547. doi: 10.1111/j.1365-2672.2010.04675.x.

Lazaris, A., Coleman, D. C., Kearns, A. M., Pichon, B., Kinnevey, P. M., Earls, M. R., Boyle, B., O’Connell, B., Brennan, G. I., and Shore, A. C. (2017) ‘Novel multiresistance *cfr* plasmids in linezolid-resistant methicillin-resistant *Staphylococcus* epidermidis and vancomycin-resistant *Enterococcus faecium* (VRE) from a hospital outbreak: co-location of *cfr and* optrA in VRE’, *The Journal of antimicrobial chemotherapy*, 72(12), pp. 3252–3257. doi: 10.1093/JAC/DKX292.

Lee, C. R., Lee, J. H., Park, K. S., Kim, Y. B., Jeong, B. C., and Lee, S. H. (2016) ‘Global dissemination of carbapenemase-producing *Klebsiella pneumoniae*: Epidemiology, genetic context, treatment options, and detection methods’, *Frontiers in Microbiology*, 7(JUN), p. 895. doi: 10.3389/FMICB.2016.00895/BIBTEX.

Lester, C. H., Frimodt-Moller, N., and Hammerum, A. M. (2004) ‘Conjugal transfer of aminoglycoside and macrolide resistance between *Enterococcus faecium* isolates in the intestine of streptomycin-treated mice’, *FEMS microbiology letters*, 235(2), pp. 385–391. doi: 10.1016/J.FEMSLE.2004.04.050.

Letunic, I. and Bork, P. (2021) ‘Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation’, *Nucleic Acids Research*, 49(W1), pp. W293–W296. doi: 10.1093/NAR/GKAB301.

Ley, R. E., Bäckhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., and Gordon, J. I. (2005) ‘Obesity alters gut microbial ecology’, *Proceedings of the National Academy* *of Sciences of the United States of America*, 102(31), pp. 11070–11075. doi: 10.1073/PNAS.0504978102/SUPPL_FILE/04978FIG5.PDF.

Li, D., Luo, R., Liu, C. M., Leung, C. M., Ting, H. F., Sadakane, K., Yamashita, H., and Lam, T. W. (2016) ‘MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices’, *Methods*, 102, pp. 3–11. doi: 10.1016/j.ymeth.2016.02.020.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009) ‘The Sequence Alignment/Map format and SAMtools’, *Bioinformatics*, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

Li, H. and Durbin, R. (2009) ‘Fast and accurate short read alignment with Burrows-Wheeler transform’, *Bioinformatics*, 25(14), pp. 1754–1760. doi: 10.1093/bioinformatics/btp324.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009) ‘Comprehensive mapping of long-range interactions reveals folding principles of the human genome.’, *Science (New York, N.Y.)*, 326(5950), pp. 289–93. doi: 10.1126/science.1181369.

Linkevicius, M., Sandegren, L., and Andersson, D. I. (2015) ‘Potential of Tetracycline Resistance Proteins To Evolve Tigecycline Resistance’, *Antimicrobial agents and chemotherapy*, 60(2), pp. 789–796. doi: 10.1128/AAC.02465-15.

Liu, M., Li, X., Xie, Y., Bi, D., Sun, J., Li, J., Tai, C., Deng, Z., and Ou, H. Y. (2019) ‘ICEberg 2.0: an updated database of bacterial integrative and conjugative elements’, *Nucleic Acids Research*, 47(D1), pp. D660–D665. doi: 10.1093/NAR/GKY1123.

Mallawaarachchi, V., Wickramarachchi, A., and Lin, Y. (2020) ‘GraphBin: refined binning of metagenomic contigs using assembly graphs’, *Bioinformatics (Oxford, England)*, 36(11), pp. 3307–3313. doi: 10.1093/BIOINFORMATICS/BTAA180.

Mandal, R. S., Saha, S., and Das, S. (2015) ‘Metagenomic Surveys of Gut Microbiota’, *Genomics, Proteomics & Bioinformatics*, 13(3), pp. 148–158. doi: 10.1016/J.GPB.2015.02.005.

Manges, A. R., Geum, H. M., Guo, A., Edens, T. J., Fibke, C. D., and Pitout, J. D. D. (2019) ‘Global Extraintestinal Pathogenic *Escherichia coli* (ExPEC) Lineages’, *Clinical Microbiology Reviews*, 32(3). doi: 10.1128/CMR.00135-18.

Manyi-Loh, C., Mamphweli, S., Meyer, E., and Okoh, A. (2018) ‘Antibiotic Use in Agriculture and Its Consequential Resistance in Environmental Sources: Potential Public Health Implications’, *Molecules : A Journal of Synthetic Chemistry and Natural Product Chemistry*, 23(4). doi: 10.3390/MOLECULES23040795.

Marbouty, M., Cournac, A., Flot, J.-F., Marie-Nelly, H., Mozziconacci, J., and Koszul, R. (2014) ‘Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms’, *eLife*, 3, p. e03318. doi: 10.7554/ELIFE.03318.

Marbouty, M., Baudry, L., Cournac, A., and Koszul, R. (2017) ‘Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay.’, *Science advances*, 3(2), p. e1602105. doi: 10.1126/sciadv.1602105.

Marbouty, M., Thierry, A., Millot, G. A., and Koszul, R. (2021) ‘MetaHiC phage-bacteria infection network reveals active cycling phages of the healthy human gut’, *eLife*, 10, pp. 1–51. doi: 10.7554/eLife.60608.

Marbouty, M. and Koszul, R. (2017) ‘Generation and analysis of chromosomal contact maps of bacteria’, in *Methods in Molecular Biology*, pp. 75–84. doi: 10.1007/978-1-4939-7098-8_7.

Marchesi, J. R. and Ravel, J. (2015) ‘The vocabulary of microbiome research: a proposal’, *Microbiome 2015 3:1*, 3(1), pp. 1–3. doi: 10.1186/S40168-015-0094-5.

Marín, M., Martín, A., Alcalá, L., Cercenado, E., Iglesias, C., Reigadas, E., and Bouzaa, E. (2015) ‘*Clostridium difficile* isolates with high linezolid MICs harbor the multiresistance gene *cfr*’, *Antimicrobial agents and chemotherapy*, 59(1), pp. 586–589. doi: 10.1128/AAC.04082-14.

Markwart, R., Willrich, N., Haller, S., Noll, I., Koppe, U., Werner, G., Eckmanns, T., and Reuss, A. (2019) ‘The rise in vancomycin-resistant *Enterococcus faecium* in Germany: data from the German Antimicrobial Resistance Surveillance (ARS)’, *Antimicrobial Resistance and Infection Control*, 8(1), pp. 1–11. doi: 10.1186/S13756-019-0594-3/FIGURES/4.

Martin, J., Phan, H. T. T., Findlay, J., Stoesser, N., Pankhurst, L., Navickaite, I., De Maio, N., Eyre, D. W., Toogood, G., Orsi, N. M., Kirby, A., Young, N., Turton, J. F., Hill, R. L. R., Hopkins, K. L., Woodford, N., Peto, T. E. A., Walker, A. S., Crook, D. W., and Wilcox, M. H. (2017) ‘Covert dissemination of carbapenemase-producing *Klebsiella pneumoniae* (KPC) in a successfully controlled outbreak: long- and short-read whole-genome sequencing demonstrate multiple genetic modes of transmission’, *Journal of Antimicrobial Chemotherapy*, 72(11), pp. 3025–3034. doi: 10.1093/jac/dkx264.

Martin, M. (2011) ‘Cutadapt removes adapter sequences from high-throughput sequencing reads’, *EMBnet.journal*, 17(1), p. 10. doi: 10.14806/ej.17.1.200.

Martín, R., Miquel, S., Ulmer, J., Kechaou, N., Langella, P., and Bermúdez-Humarán, L. G. (2013) ‘Role of commensal and probiotic bacteria in human health: A focus on inflammatory bowel disease’, *Microbial Cell Factories*, 12(1), pp. 1–11. doi: 10.1186/1475-2859-12-71/FIGURES/2.

McCombie, W. R., McPherson, J. D., and Mardis, E. R. (2019) ‘Next-Generation Sequencing Technologies’, *Cold Spring Harbor Perspectives in Medicine*, 9(11). doi: 10.1101/CSHPERSPECT.A036798.

McInnes, R. S., McCallum, G. E., Lamberte, L. E., and van Schaik, W. (2020) ‘Horizontal transfer of antibiotic resistance genes in the human gut microbiome’, *Current Opinion in Microbiology*, 53, pp. 35–43. doi: 10.1016/j.mib.2020.02.002.

Meek, R. W., Vyas, H., and Piddock, L. J. V. (2015) ‘Nonmedical Uses of Antibiotics: Time to Restrict Their Use?’, *PLoS Biology*, 13(10), pp. 1–11. doi: 10.1371/JOURNAL.PBIO.1002266.

Mehrad, B., Clark, N. M., Zhanel, G. G., and Lynch, J. P. 3rd (2015) ‘Antimicrobial Resistance in Hospital-Acquired Gram-Negative Bacterial Infections’, *Chest*, 147(5), p. 1413. doi: 10.1378/CHEST.14-2171.

Von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H., and Dutilh, B. E. (2019) ‘Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT’, *Genome Biology*, 20(1), pp. 1–14. doi: 10.1186/s13059-019-1817-x.

Microsoft Corporation (2022) *Microsoft Excel* (Version 2209 Build 16.0.15629.20200) [Computer Program]. Available at: https://office.microsoft.com/excel.

Mikheenko, A., Saveliev, V., and Gurevich, A. (2016) ‘MetaQUAST: evaluation of metagenome assemblies’, *Bioinformatics (Oxford, England)*, 32(7), pp. 1088–1090. doi: 10.1093/BIOINFORMATICS/BTV697.

Miller, W. R., Munita, J. M., and Arias, C. A. (2014) ‘Mechanisms of antibiotic resistance in enterococci’, *Expert review of anti-infective therapy*, 12(10), p. 1221. doi: 10.1586/14787210.2014.956092.

Mizrahi-Man, O., Davenport, E. R., and Gilad, Y. (2013) ‘Taxonomic Classification of Bacterial 16S rRNA Genes Using Short Sequencing Reads: Evaluation of Effective Study Designs’, *PLoS ONE*, 8(1), p. e53608. doi: 10.1371/JOURNAL.PONE.0053608.

Mobatek (2022) *MobaXterm Personal Edition* (Version 21.2) [Computer program]. Available at: https://mobaxterm.mobatek.net/.

Mooiman, C., Bouwknegt, J., Dekker, W. J. C., Wiersma, S. J., Ortiz-Merino, R. A., De Hulster, E., and Pronk, J. T. (2021) ‘Critical parameters and procedures for anaerobic cultivation of yeasts in bioreactors and anaerobic chambers’, *FEMS Yeast Research*, 21(5), p. 35. doi: 10.1093/FEMSYR/FOAB035.

Moore, W. E. C. and Holdeman, L. V. (1974) ‘Human Fecal Flora: The Normal Flora of 20 Japanese-Hawaiians’, *Applied Microbiology*, 27(5), pp. 961–979. doi: 10.1128/AM.27.5.961-979.1974.

Moreau, P., Cournac, A., Palumbo, G. A., Marbouty, M., Mortaza, S., Thierry, A., Cairo, S., Lavigne, M., Koszul, R., and Neuveut, C. (2018) ‘Tridimensional infiltration of DNA viruses into the host genome shows preferential contact with active chromatin’, *Nature Communications*, 9(1), p. 4268. doi: 10.1038/s41467-018-06739-4.

Mullish, B. H. and Williams, H. R. T. (2018) ‘*Clostridium difficile* infection and antibiotic-associated diarrhoea’, *Clinical Medicine*, 18(3), p. 237. doi: 10.7861/CLINMEDICINE.18-3-237.

Munck, C., Sheth, R. U., Freedberg, D. E., and Wang, H. H. (2020) ‘Recording mobile DNA in the gut microbiota using an *Escherichia coli* CRISPR-Cas spacer acquisition platform’, *Nature Communications 2020 11:1*, 11(1), pp. 1–11. doi: 10.1038/s41467-019-14012-5.

Murray, C. J., Ikuta, K. S., Sharara, F., Swetschinski, L., Robles Aguilar, G., Gray, A., Han, C., Bisignano, C., Rao, P., Wool, E., Johnson, S. C., Browne, A. J., Chipeta, M. G., Fell, F., Hackett, S., Haines-Woodhouse, G., Kashef Hamadani, B. H., Kumaran, E. A. P., McManigal, B., Agarwal, R., Akech, S., Albertson, S., Amuasi, J., Andrews, J., Aravkin, A., Ashley, E., Bailey, F., Baker, S., Basnyat, B., Bekker, A., Bender, R., Bethou, A., Bielicki, J., Boonkasidecha, S., Bukosia, J., Carvalheiro, C., Castañeda-Orjuela, C., Chansamouth, V., Chaurasia, S., Chiurchiù, S., Chowdhury, F., Cook, A. J., Cooper, B., Cressey, T. R., Criollo-Mora, E., Cunningham, M., Darboe, S., Day, N. P. J., De Luca, M., Dokova, K., Dramowski, A., Dunachie, S. J., Eckmanns, T., Eibach, D., Emami, A., Feasey, N., Fisher-Pearson, N., Forrest, K., Garrett, D., Gastmeier, P., Giref, A. Z., Greer, R. C., Gupta, V., Haller, S., Haselbeck, A., Hay, S. I., Holm, M., Hopkins, S., Iregbu, K. C., Jacobs, J., Jarovsky, D., Javanmardi, F., Khorana, M., Kissoon, N., Kobeissi, E., Kostyanev, T., Krapp, F., Krumkamp, R., Kumar, A., Kyu, H. H., Lim, C., Limmathurotsakul, D., Loftus, M. J., Lunn, M., Ma, J., Mturi, N., Munera-Huertas, T., Musicha, P., Mussi-Pinhata, M. M., Nakamura, T., Nanavati, R., Nangia, S., Newton, P., Ngoun, C., Novotney, A., Nwakanma, D., Obiero, C. W., Olivas-Martinez, A., Olliaro, P., Ooko, E., Ortiz-Brizuela, E., Peleg, A. Y., Perrone, C., Plakkal, N., Ponce-de-Leon, A., Raad, M., Ramdin, T., Riddell, A., Roberts, T., Robotham, J. V., Roca, A., Rudd, K. E., Russell, N., Schnall, J., Scott, J. A. G., Shivamallappa, M., Sifuentes-Osornio, J., Steenkeste, N., Stewardson, A. J., Stoeva, T., Tasak, N., Thaiprakong, A., Thwaites, G., Turner, C., Turner, P., van Doorn, H. R., Velaphi, S., Vongpradith, A., Vu, H., Walsh, T., Waner, S., Wangrangsimakul, T., Wozniak, T., Zheng, P., Sartorius, B., Lopez, A. D., Stergachis, A., Moore, C., Dolecek, C., and Naghavi, M. (2022) 'Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis', *The Lancet*, 399(10325), pp. 629–655. doi: 10.1016/S0140-6736(21)02724-0.

Nagano, T., Várnai, C., Schoenfelder, S., Javierre, B.-M., Wingett, S. W., and Fraser, P. (2015) 'Comparison of Hi-C results using in-solution versus in-nucleus ligation', *Genome Biology*, 16(1), p. 175. doi: 10.1186/s13059-015-0753-7.

National Center for Biotechnology Information (2022) *Datasets - NCBI*. Available at: https://www.ncbi.nlm.nih.gov/datasets/ (Accessed: 24 May 2022).

National Library of Medicine (2022) *BLAST: Basic Local Alignment Search Tool*. Available at: https://blast.ncbi.nlm.nih.gov/Blast.cgi (Accessed: 24 April 2022).

Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S., and Kyrpides, N. C. (2019) 'New insights from uncultivated genomes of the global human gut microbiome', *Nature*, 568(7753), pp. 505–510. doi: 10.1038/s41586-019-1058-x.

Negi, S., Das, D. K., Pahari, S., Nadeem, S., and Agrewala, J. N. (2019) 'Potential Role of Gut Microbiota in Induction and Regulation of Innate Immune Memory', *Frontiers in Immunology*, 10(OCT), p. 2441. doi: 10.3389/FIMMU.2019.02441.

New England BioLabs (2022) *Luna® Universal qPCR Master Mix Protocol (#M3003)*. Available at: https://international.neb.com/protocols/2016/11/08/luna-universal-qpcr-master-mix-protocol-m3003 (Accessed: 25 October 2021).

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015) 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood

Phylogenies', *Molecular Biology and Evolution*, 32(1), pp. 268–274. doi: 10.1093/MOLBEV/MSU300.

Nikolaou, E., Hubbard, A. T. M., Botelho, J., Marschall, T. A. M., Ferreira, D. M., and Roberts, A. P. (2020) 'Antibiotic Resistance Is Associated with Integrative and Conjugative Elements and Genomic Islands in Naturally Circulating *Streptococcus pneumoniae* Isolates from Adults in Liverpool, UK', *Genes*, 11(6), pp. 1–9. doi: 10.3390/GENES11060625.

Nordmann, P., Dortet, L., and Poirel, L. (2012) 'Carbapenem resistance in Enterobacteriaceae: here is the storm!', *Trends in Molecular Medicine*, 18(5), pp. 263–272. doi: 10.1016/J.MOLMED.2012.03.003.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. (2016) 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic acids research*, 44(D1), pp. D733–D745. doi: 10.1093/NAR/GKV1189.

O'Neill, J. (2016) *Tackling drug-resistant infections globally: final report and recommendations*. Available at: https://amr-review.org/Publications.html.

Olekhnovich, E. I., Vasilyev, A. T., Ulyantsev, V. I., Kostryukova, E. S., and Tyakht, A. V. (2018) 'MetaCherchant: analyzing genomic context of antibiotic resistance genes in gut microbiota', *Bioinformatics*, 34(3), pp. 434–444. doi: 10.1093/BIOINFORMATICS/BTX681.

Oliphant, K. and Allen-Vercoe, E. (2019) 'Macronutrient metabolism by the human gut microbiome: major fermentation by-products and their impact on host health', *Microbiome 2019 7:1*, 7(1), pp. 1–15. doi: 10.1186/S40168-019-0704-8.

Ooijevaar, R. E., van Beurden, Y. H., Terveer, E. M., Goorhuis, A., Bauer, M. P., Keller, J. J., Mulder, C. J. J., and Kuijper, E. J. (2018) 'Update of treatment algorithms for Clostridium difficile infection', *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 24(5), pp. 452–462. doi: 10.1016/J.CMI.2017.12.022.

Palmer, K. L., Kos, V. N., and Gilmore, M. S. (2010) 'Horizontal Gene Transfer and the Genomics of Enterococcal Antibiotic Resistance', *Current opinion in microbiology*, 13(5), p. 632. doi: 10.1016/J.MIB.2010.08.004.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015) 'CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes', *Genome Research*, 25(7), pp. 1043–1055. doi: 10.1101/GR.186072.114.

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P. A., and Hugenholtz, P. (2018) ‘A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life’, *Nature Biotechnology 2018 36:10*, 36(10), pp. 996–1004. doi: 10.1038/nbt.4229.

Parks, D. H., Chuvochina, M., Chaumeil, P. A., Rinke, C., Mussig, A. J., and Hugenholtz, P. (2020) ‘A complete domain-to-species taxonomy for Bacteria and Archaea’, *Nature Biotechnology 2020 38:9*, 38(9), pp. 1079–1086. doi: 10.1038/s41587-020-0501-8.

Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P. A., and Hugenholtz, P. (2022) ‘GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy’, *Nucleic Acids Research*, 50(D1), pp. D785–D794. doi: 10.1093/NAR/GKAB776.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., Collado, M. C., Rice, B. L., DuLong, C., Morgan, X. C., Golden, C. D., Quince, C., Huttenhower, C., and Segata, N. (2019) ‘Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle’, *Cell*, 176(3), pp. 649-662.e20. doi: 10.1016/J.CELL.2019.01.001.

Petersson, J., Schreiber, O., Hansson, G. C., Gendler, S. J., Velcich, A., Lundberg, J. O., Roos, S., Holm, L., and Phillipson, M. (2011) ‘Importance and regulation of the colonic mucus barrier in a mouse model of colitis’, *American Journal of Physiology - Gastrointestinal and Liver Physiology*, 300(2), p. G327. doi: 10.1152/AJPGI.00422.2010.

Pevzner, P. A., Tang, H., and Waterman, M. S. (2001) ‘An Eulerian path approach to DNA fragment assembly’, *Proceedings of the National Academy of Sciences of the United States of America*, 98(17), pp. 9748–9753. doi: 10.1073/PNAS.171285098.

Pfleiderer, A., Lagier, J. C., Armougom, F., Robert, C., Vialettes, B., and Raoult, D. (2013) ‘Culturomics identified 11 new bacterial species from a single anorexia nervosa stool sample’, *European Journal of Clinical Microbiology and Infectious Diseases*, 32(11), pp. 1471–1481. doi: 10.1007/S10096-013-1900-2/TABLES/4.

Phan, H. T. T., Stoesser, N., Maciuca, I. E., Toma, F., Szekely, E., Flonta, M., Hubbard, A. T. M., Pankhurst, L., Do, T., Peto, T. E. A., Walker, A. S., Crook, D. W., and Timofte, D. (2018) ‘Illumina short-read and MinION long-read WGS to characterize the molecular epidemiology of an NDM-1 *Serratia marcescens* outbreak in Romania’, *Journal of Antimicrobial Chemotherapy*, 73(3), pp. 672–679. doi: 10.1093/jac/dkx456.

Phase Genomics Inc. (2022) *ProxiMeta Hi-C Metagenomic Deconvolution Platform*. Available at: https://proximeta.phasegenomics.com/ (Accessed: 26 August 2021).

Pokharel, S., Raut, S., and Adhikari, B. (2019) ‘Tackling antimicrobial resistance in low-income and middle-income countries’, *BMJ Global Health*, 4(6), p. e002104. doi: 10.1136/BMJGH-2019-002104.

Poolman, J. T. and Wacker, M. (2016) ‘Extraintestinal Pathogenic *Escherichia coli*, a Common Human Pathogen: Challenges for Vaccine Development and Progress in the Field’, *The Journal of Infectious Diseases*, 213(1), p. 6. doi: 10.1093/INFDIS/JIV429.

Porse, A., Schou, T. S., Munck, C., Ellabaan, M. M. H., and Sommer, M. O. A. (2018) ‘Biochemical mechanisms determine the functional compatibility of heterologous genes’, *Nature Communications 2018 9:1*, 9(1), pp. 1–11. doi: 10.1038/s41467-018-02944-3.

Portillo, A., Ruiz-Larrea, F., Zarazaga, M., Alonso, A., Martinez, J. L., and Torres, C. (2000) ‘Macrolide resistance genes in *Enterococcus* spp.’, *Antimicrobial Agents and Chemotherapy*, 44(4), pp. 967–971. doi: 10.1128/AAC.44.4.967-971.2000/ASSET/1D8AFE39-9F47-4B47-9F0B-1EB3143380B3/ASSETS/GRAPHIC/AC0400579002.JPEG.

van Prehn, J., Reigadas, E., Vogelzang, E. H., Bouza, E., Hristea, A., Guery, B., Krutova, M., Norén, T., Allerberger, F., Coia, J. E., Goorhuis, A., van Rossen, T. M., Ooijevaar, R. E., Burns, K., Scharvik Olesen, B. R., Tschudin-Sutter, S., Wilcox, M. H., Vehreschild, M. J. G. T., Fitzpatrick, F., and Kuijper, E. J. (2021) ‘European Society of Clinical Microbiology and Infectious Diseases: 2021 update on the treatment guidance document for *Clostridioides difficile* infection in adults’, *Clinical Microbiology and Infection*, 27, pp. S1–S21. doi: 10.1016/J.CMI.2021.09.038/ATTACHMENT/F1291CE1-176A-4FF5-8721-566BD0F0433E/MMC2.DOCX.

Press, M. O., Wiser, A. H., Kronenberg, Z. N., Langford, K. W., Shakya, M., Lo, C.-C., Mueller, K. A., Sullivan, S. T., Chain, P. S. G., and Liachko, I. (2017) ‘Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions’, *bioRxiv*, p. 198713. doi: 10.1101/198713.

Prestinaci, F., Pezzotti, P., and Pantosti, A. (2015) ‘Antimicrobial resistance: a global multifaceted phenomenon’, *Pathogens and Global Health*, 109(7), p. 309. doi: 10.1179/2047773215Y.0000000030.

Price, L. B., Hungate, B. A., Koch, B. J., Davis, G. S., and Liu, C. M. (2017) ‘Colonizing opportunistic pathogens (COPs): The beasts in all of us’, *PLOS Pathogens*, 13(8), p. e1006369. doi: 10.1371/JOURNAL.PPAT.1006369.

Prikryl, M. (2022) *WinSCP* (Version 5.21.3) [Computer program]. Available at: https://winscp.net.

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., and Korobeynikov, A. (2020) ‘Using SPAdes De Novo Assembler’, *Current Protocols in Bioinformatics*, 70(1), p. e102. doi: 10.1002/CPBI.102.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, Shaochuan, Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J. M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, Shengting, Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, Songgang, Qin, N., Yang, H., Wang, Jian, Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S. D., Wang, Jun, Antolin, M., Artiguenave, F., Blottiere, H., Borruel, N., Bruls, T., Casellas, F., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Forte, M., Friss, C., Van De Guchte, M., Guedon, E., Haimet, F., Jamet, A., Juste, C., Kaci, G., Kleerebezem, M., Knol, J., Kristensen, M., Layec, S., Le Roux, K., Leclerc, M., Maguin, E., Melo Minardi, R., Oozeer, R., Rescigno, M., Sanchez, N., Tims, S., Torrejon, T., Varela, E., De Vos, W., Winogradsky, Y., and Zoetendal, E. (2010) 'A human gut microbial gene catalogue established by metagenomic sequencing', *Nature 2010 464:7285*, 464(7285), pp. 59–65. doi: 10.1038/nature08821.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017) 'Shotgun metagenomics, from sampling to analysis', *Nature biotechnology*, 35(9), pp. 833–844. doi: 10.1038/NBT.3935.

Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26(6), pp. 841–842. doi: 10.1093/bioinformatics/btq033.

R Core Team (2022) *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.r-project.org/.

Rajilić-Stojanović, M. and de Vos, W. M. (2014) 'The first 1000 cultured species of the human gastrointestinal microbiota', *Fems Microbiology Reviews*, 38(5), p. 996. doi: 10.1111/1574-6976.12075.

Ramsay, J. P. and Firth, N. (2017) 'Diverse mobilization strategies facilitate transfer of non-conjugative mobile genetic elements', *Current Opinion in Microbiology*, 38, pp. 1–9. doi: 10.1016/J.MIB.2017.03.003.

Raplee, I., Walker, L., Xu, L., Surathu, A., Chockalingam, A., Stewart, S., Han, X., Rouse, R., and Li, Z. (2021) 'Emergence of nosocomial associated opportunistic pathogens in the gut microbiome after antibiotic treatment', *Antimicrobial Resistance and Infection Control*, 10(1), pp. 1–11. doi: 10.1186/S13756-021-00903-0/FIGURES/6.

Rettedal, E. A., Gumpert, H., and Sommer, M. O. A. (2014) 'Cultivation-based multiplex phenotyping of human gut microbiota allows targeted recovery of previously uncultured bacteria', *Nature Communications 2014 5:1*, 5(1), pp. 1–9. doi: 10.1038/ncomms5714.

Ricaboni, D., Mailhe, M., Cadoret, F., Vitton, V., Fournier, P. E., Raoult, D., and Lagier, J. C. (2017) '"*Merdibacter massiliensis*" gen. nov., sp. nov., isolated from human ileum', *New Microbes and New Infections*, 15, p. 89. doi: 10.1016/J.NMNI.2016.11.017.

Ricaboni, D., Mailhe, M., Vitton, V., Andrieu, C., Fournier, P. E., and Raoult, D. (2017) '"*Negativibacillus massiliensis*" gen. nov., sp. nov., isolated from human left colon', *New Microbes and New Infections*, 17, p. 36. doi: 10.1016/J.NMNI.2016.11.002.

Rice, L. B. (1998) 'Tn*916* Family Conjugative Transposons and Dissemination of Antimicrobial Resistance Determinants', *Antimicrobial Agents and Chemotherapy*, 42(8), p. 1871. doi: 10.1128/AAC.42.8.1871.

Richter, M. and Rosselló-Móra, R. (2009) 'Shifting the genomic gold standard for the prokaryotic species definition', *Proceedings of the National Academy of Sciences of the United States of America*, 106(45), p. 19126. doi: 10.1073/PNAS.0906412106.

Riley, L. W. (2014) ‘Pandemic lineages of extraintestinal pathogenic *Escherichia coli*’, *Clinical Microbiology and Infection*, 20(5), pp. 380–390. doi: 10.1111/1469-0691.12646.

Rinninella, E., Raoul, P., Cintoni, M., Franceschi, F., Miggiano, G. A. D., Gasbarrini, A., and Mele, M. C. (2019) ‘What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases’, *Microorganisms*, 7(1), p. 14. doi: 10.3390/MICROORGANISMS7010014.

Roberts, A. P., Cheah, G., Ready, D., Pratten, J., Wilson, M., and Mullany, P. (2001) ‘Transfer of Tn*916*-Like Elements in Microcosm Dental Plaques’, *Antimicrobial Agents and Chemotherapy*, 45(10), p. 2943. doi: 10.1128/AAC.45.10.2943-2946.2001.

Roberts, A. P. and Mullany, P. (2011) ‘Tn*916*-like genetic elements: a diverse group of modular mobile elements conferring antibiotic resistance’, *FEMS Microbiology Reviews*, 35(5), pp. 856–871. doi: 10.1111/J.1574-6976.2011.00283.X.

Rodley, C. D. M., Bertels, F., Jones, B., and O’Sullivan, J. M. (2009) ‘Global identification of yeast chromosome interactions using Genome conformation capture’, *Fungal Genetics and Biology*, 46(11), pp. 879–886. doi: 10.1016/j.fgb.2009.07.006.

Rosero, J. A., Killer, J., Sechovcová, H., Mrázek, J., Benada, O., Fliegerová, K., Havlík, J., and Kopečný, J. (2016) ‘Reclassification of *Eubacterium rectale* (Hauduroy *et al.* 1937) Prévot 1938 in a new genus *Agathobacter* gen. nov. as *Agathobacter rectalis* comb. nov., and description of *Agathobacter ruminis* sp. nov., isolated from the ru’, *International journal of systematic and evolutionary microbiology*, 66(2), pp. 768–773. doi: 10.1099/IJSEM.0.000788.

Le Roy, T., Van der Smissen, P., Paquot, A., Delzenne, N., Muccioli, G. G., Collet, J. F., and Cani, P. D. (2020) ‘*Dysosmobacter welbionis* gen. nov., sp. nov., isolated from human faeces and emended description of the genus *Oscillibacter*’, *International Journal of Systematic and Evolutionary Microbiology*, 70(9), pp. 4851–4858. doi: 10.1099/IJSEM.0.003547/CITE/REFWORKS.

RStudio Team (2022) *RStudio: Integrated Development for R* (Version 2022.7.1.554) [Computer program]. Available at: http://www.rstudio.com/.

Rumbo, C., Fernández-Moreira, E., Merino, M., Poza, M., Mendez, J. A., Soares, N. C., Mosquera, A., Chaves, F., and Bou, G. (2011) ‘Horizontal Transfer of the OXA-24 Carbapenemase Gene via Outer Membrane Vesicles: a New Mechanism of Dissemination of Carbapenem Resistance Genes in *Acinetobacter baumannii*’, *Antimicrobial Agents and Chemotherapy*, 55(7), p. 3084. doi: 10.1128/AAC.00929-10.

Saha, S., Kapoor, S., Tariq, R., Schuetz, A. N., Tosh, P. K., Pardi, D. S., and Khanna, S. (2019) ‘Increasing antibiotic resistance in *Clostridioides difficile*: A systematic review and meta-analysis’, *Anaerobe*, 58, pp. 35–46. doi: 10.1016/J.ANAEROBE.2019.102072.

Sakamoto, M. and Benno, Y. (2006) ‘Reclassification of *Bacteroides distasonis*, *Bacteroides goldsteinii* and *Bacteroides merdae* as *Parabacteroides distasonis* gen. nov., comb. nov., *Parabacteroides goldsteinii* comb. nov. and *Parabacteroides merdae* com’, *International journal of systematic and evolutionary microbiology*, 56(Pt 7), pp. 1599–1605. doi: 10.1099/IJS.0.64192-0.

Salonen, A., Nikkilä, J., Jalanka-Tuovinen, J., Immonen, O., Rajilić-Stojanović, M., Kekkonen, R. A., Palva, A., and de Vos, W. M. (2010) ‘Comparative analysis of fecal DNA extraction methods with phylogenetic microarray: Effective recovery of bacterial and archaeal DNA using mechanical cell lysis’, *Journal of Microbiological Methods*, 81(2), pp. 127–134. doi: 10.1016/j.mimet.2010.02.007.

Salyers, A. A., Gupta, A., and Wang, Y. (2004) ‘Human intestinal bacteria as reservoirs for antibiotic resistance genes’, *Trends in microbiology*, 12(9), pp. 412–416. doi: 10.1016/J.TIM.2004.07.004.

Sangwan, N., Xia, F., and Gilbert, J. A. (2016) ‘Recovering complete and draft population genomes from metagenome datasets’, *Microbiome*, 4. doi: 10.1186/S40168-016-0154-5.

Santiago, A., Panda, S., Mengels, G., Martinez, X., Azpiroz, F., Dore, J., Guarner, F., and Manichanh, C. (2014) ‘Processing faecal samples: A step forward for standards in microbial community analysis’, *BMC Microbiology*, 14(1), pp. 1–9. doi: 10.1186/1471-2180-14-112.

Sarangi, A. N., Goel, A., and Aggarwal, R. (2019) ‘Methods for Studying Gut Microbiota: A Primer for Physicians’, *Journal of Clinical and Experimental Hepatology*, 9(1), pp. 62–73. doi: 10.1016/J.JCEH.2018.04.016.

Sati, S. and Cavalli, G. (2017) ‘Chromosome conformation capture technologies and their impact in understanding genome function’, *Chromosoma*, 126(1), pp. 33–44. doi: 10.1007/s00412-016-0593-6.

Savage, D. C. and Dubos, R. J. (1968) ‘Alterations in the mouse cecum and its flora produced by antibacterial drugs’, *Journal of Experimental Medicine*, 128(1), pp. 97–110. doi: 10.1084/JEM.128.1.97.

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Tse, T., Wang, J., Williams, R., Trawick, B. W., Pruitt, K. D., and Sherry, S. T. (2022) ‘Database resources of the National Center for Biotechnology Information’, *Nucleic acids research*, 50(D1), pp. D20–D26. doi: 10.1093/NAR/GKAB1112.

van Schaik, W. (2015) ‘The human gut resistome’, *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370(1670), p. 20140087. doi: 10.1098/rstb.2014.0087.

Schloss, P. D. and Handelsman, J. (2005) ‘Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness’, *Applied and Environmental Microbiology*, 71(3), pp. 1501–1506. doi: 10.1128/AEM.71.3.1501-1506.2005/ASSET/2DDD2F7C-C368-4869-A3C9-FE0F6D56BCA1/ASSETS/GRAPHIC/ZAM0030552650004.JPEG.

Schluter, J. and Foster, K. R. (2012) ‘The Evolution of Mutualism in Gut Microbiota Via Host Epithelial Selection’, *PLoS Biology*, 10(11), p. 1001424. doi: 10.1371/JOURNAL.PBIO.1001424.

Schmieder, R. and Edwards, R. (2011) ‘Quality control and preprocessing of metagenomic datasets’, *BIOINFORMATICS APPLICATIONS NOTE*, 27(6), pp. 863–864. doi: 10.1093/bioinformatics/btr026.

Scott, K. P., Melville, C. M., Barbosa, T. M., and Flint, H. J. (2000) ‘Occurrence of the New Tetracycline Resistance Gene *tet*(W) in Bacteria from the Human Gut’, *ANTIMICROBIAL AGENTS AND CHEMOTHERAPY*, 44(3), pp. 775–777.

Seemann, T. (2014) ‘Prokka: rapid prokaryotic genome annotation’, *Bioinformatics*, 30(14), pp. 2068–2069. doi: 10.1093/bioinformatics/btu153.

Seemann, T. (2021) *ABRicate* (Version 1.0.1) [Computer program]. Available at: https://github.com/tseemann/abricate.

Seemann, T. (2022) *mlst* (Version 2.19.0) [Computer program]. Available at: https://github.com/tseemann/mlst.

Sekirov, I., Russell, S. L., Antunes, L. C. M., and Finlay, B. B. (2010) ‘Gut Microbiota in Health and Disease’, *Physiological Reviews*, 90(3), pp. 859–904. doi: 10.1152/physrev.00045.2009.

Sender, R., Fuchs, S., and Milo, R. (2016) ‘Revised Estimates for the Number of Human and Bacteria Cells in the Body’. doi: 10.1371/journal.pbio.1002533.

Seville, L. A., Patterson, A. J., Scott, K. P., Mullany, P., Quail, M. A., Parkhill, J., Ready, D., Wilson, M., Spratt, D., and Roberts, A. P. (2009) ‘Distribution of tetracycline and erythromycin resistance genes among human oral and fecal metagenomic DNA’, *Microbial drug resistance (Larchmont, N.Y.)*, 15(3), pp. 159–166. doi: 10.1089/MDR.2009.0916.

Shambaugh, G. E. (1966) ‘History of Sulfonamides’, *Archives of Otolaryngology*, 83(1), pp. 1–2. doi: 10.1001/ARCHOTOL.1966.00760020003001.

Shapiro, J. A. and Von Sternberg, R. (2005) ‘Why repetitive DNA is essential to genome function’, *Biological Reviews*, 80(2), pp. 227–250. doi: 10.1017/S1464793104006657.

Shawa, M., Furuta, Y., Paudel, A., Kabunda, O., Mulenga, E., Mubanga, M., Kamboyi, H., Zorigt, T., Chambaro, H., Simbotwe, M., Hang’ombe, B., and Higashi, H. (2022) ‘Clonal relationship between multidrug-resistant *Escherichia coli* ST69 from poultry and humans in Lusaka, Zambia’, *FEMS Microbiology Letters*, 368(21–24), pp. 1–11. doi: 10.1093/FEMSLE/FNAC004.

Sheridan, P. O., Duncan, S. H., Walker, A. W., Scott, K. P., Louis, P., and Flint, H. J. (2016) 'Objections to the proposed reclassification of *Eubacterium rectale* as *Agathobacter rectalis*', *International Journal of Systematic and Evolutionary Microbiology*, 66(5), p. 2016. doi: 10.1099/IJSEM.0.000969/CITE/REFWORKS.

Sherrard, L. J., Schaible, B., Graham, K. A., McGrath, S. J., McIlreavey, L., Hatch, J., Wolfgang, M. C., Muhlebach, M. S., Gilpin, D. F., Schneiders, T., Stuart Elborn, J., and Tunney, M. M. (2014) 'Mechanisms of reduced susceptibility and genotypic prediction of antibiotic resistance in *Prevotella* isolated from cystic fibrosis (CF) and non-CF patients', *Journal of Antimicrobial Chemotherapy*, 69(10), p. 2690. doi: 10.1093/JAC/DKU192.

Shoemaker, N. B., Vlamakis, H., Hayes, K., and Salyers, A. A. (2001) 'Evidence for Extensive Resistance Gene Transfer among *Bacteroides* spp. and among *Bacteroides* and Other Genera in the Human Colon', *Applied and Environmental Microbiology*, 67(2), pp. 561–568. doi: 10.1128/AEM.67.2.561-568.2001.

Sholeh, M., Krutova, M., Forouzesh, M., Mironov, S., Sadeghifard, N., Molaeipour, L., Maleki, A., and Kouhsari, E. (2020) 'Antimicrobial resistance in *Clostridioides* (*Clostridium*) *difficile* derived from humans: A systematic review and meta-analysis', *Antimicrobial Resistance and Infection Control*, 9(1), pp. 1–11. doi: 10.1186/S13756-020-00815-5/FIGURES/2.

Shreiner, A. B., Kao, J. Y., and Young, V. B. (2015) 'The gut microbiome in health and in disease', *Current opinion in gastroenterology*, 31(1), p. 69. doi: 10.1097/MOG.0000000000000139.

Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., and Banfield, J. F. (2018) 'Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy', *Nature Microbiology 2018 3:7*, 3(7), pp. 836–843. doi: 10.1038/s41564-018-0171-1.

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006) 'ISfinder: the reference centre for bacterial insertion sequences', *Nucleic Acids Research*, 34, pp. D32–D36. doi: 10.1093/nar/gkj014.

Siguier, P., Gourbeyre, E., and Chandler, M. (2014) 'Bacterial insertion sequences: Their genomic impact and diversity', *FEMS Microbiology Reviews*, 38(5), pp. 865–891. doi: 10.1111/1574-6976.12067.

Singh, S., Verma, N., and Taneja, N. (2019) 'The human gut resistome: Current concepts & future prospects', *The Indian Journal of Medical Research*, 150(4), p. 345. doi: 10.4103/IJMR.IJMR_1979_17.

Smith, J. L., Fratamico, P. M., and Gunther, N. W. (2007) 'Extraintestinal pathogenic *Escherichia coli*', *Foodborne pathogens and disease*, 4(2), pp. 134–163. doi: 10.1089/FPD.2007.0087.

Smith, L. (2020) *Converting Metaphlan profile to Phyloseq objects*. Available at: https://github.com/flannsmith/metaphlan-plot-by-taxa/blob/72c0db5302e8b1e05732ef46a1252cad72ff8075/Converting Metaphlan profile to Phyloseq objects.ipynb (Accessed: 13 January 2021).

Sóki, J., Wybo, I., Hajdú, E., Toprak, N. U., Jeverica, S., Stingu, C. S., Tierney, D., Perry, J. D., Matuz, M., Urbán, E., and Nagy, E. (2020) ‘A Europe-wide assessment of antibiotic resistance rates in *Bacteroides* and *Parabacteroides* isolates from intestinal microbiota of healthy subjects’, *Anaerobe*, 62, p. 102182. doi: 10.1016/J.ANAEROBE.2020.102182.

Soucy, S. M., Huang, J., and Gogarten, J. P. (2015) ‘Horizontal gene transfer: building the web of life’, *Nature Reviews Genetics*, 16(8), pp. 472–482. doi: 10.1038/nrg3962.

Spencer, S. J., Tamminen, M. V, Preheim, S. P., Guo, M. T., Briggs, A. W., Brito, I. L., A Weitz, D., Pitkänen, L. K., Vigneault, F., Virta, M. Pj., and Alm, E. J. (2016) ‘Massively parallel sequencing of single cells by epicPCR links functional genes with phylogenetic markers’, *The ISME Journal*, 10(2), pp. 427–436. doi: 10.1038/ismej.2015.124.

SRA Toolkit Development Team (2022) *Sequence Read Archive Toolkit* (Version 2.10.7) [Computer program]. Available at: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software.

Stalder, T., Press, M. O., Sullivan, S., Liachko, I., and Top, E. M. (2019) ‘Linking the resistome and plasmidome to the microbiome’, *ISME Journal*, 13(10), pp. 2437–2446. doi: 10.1038/s41396-019-0446-4.

Stämmler, F., Gläsner, J., Hiergeist, A., Holler, E., Weber, D., Oefner, P. J., Gessner, A., and Spang, R. (2016) ‘Adjusting microbiome profiles for differences in microbial load by spike-in bacteria’, *Microbiome*, 4(1), pp. 1–13. doi: 10.1186/s40168-016-0175-0.

Stanton, T. B., Humphrey, S. B., Sharma, V. K., and Zuerner, R. L. (2008) ‘Collateral effects of antibiotics: Carbadox and metronidazole induce VSH-I and facilitate gene transfer among *Brachyspira hyodysenteriae* strains’, *Applied and Environmental Microbiology*, 74(10), pp. 2950–2956. doi: 10.1128/AEM.00189-08/ASSET/FB93B185-52EE-41C6-8369-D871867D064D/ASSETS/GRAPHIC/ZAM0100888320004.JPEG.

Stentz, R., Horn, N., Cross, K., Salt, L., Brearley, C., Livermore, D. M., and Carding, S. R. (2015) ‘Cephalosporinases associated with outer membrane vesicles released by *Bacteroides* spp. protect gut pathogens and commensals against β-lactam antibiotics’, *Journal of Antimicrobial Chemotherapy*, 70(3), pp. 701–709. doi: 10.1093/JAC/DKU466.

Stewart, E. J. (2012) ‘Growing Unculturable Bacteria’, *Journal of Bacteriology*, 194(16), p. 4151. doi: 10.1128/JB.00345-12.

Stewart, R. D., Auffret, M. D., Warr, A., Wiser, A. H., Press, M. O., Langford, K. W., Liachko, I., Snelling, T. J., Dewhurst, R. J., Walker, A. W., Roehe, R., and Watson, M. (2018) ‘Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen.’, *Nature communications*, 9(1), p. 870. doi: 10.1038/s41467-018-03317-6.

Stinear, T. P., Olden, D. C., Johnson, P. D. R., Davies, J. K., and Grayson, M. L. (2001) ‘Enterococcal *vanB* resistance locus in anaerobic bacteria in human faeces’, *Lancet*, 357(9259), pp. 855–856. doi: 10.1016/S0140-6736(00)04206-9.

Stojković, V., Ulate, M. F., Hidalgo-Villeda, F., Aguilar, E., Monge-Cascante, C., Pizarro-Guajardo, M., Tsai, K., Tzoc, E., Camorlinga, M., Paredes-Sabja, D., Quesada-Gómez, C., Fujimori, D. G., and Rodríguez, C. (2019) '*cfr*(B), *cfr*(C), and a New *cfr*-Like Gene, *cfr*(E), in *Clostridium difficile* Strains Recovered across Latin America', *Antimicrobial agents and chemotherapy*, 64(1). doi: 10.1128/AAC.01074-19.

Stracy, M., Snitser, O., Yelin, I., Amer, Y., Parizade, M., Katz, R., Rimler, G., Wolf, T., Herzel, E., Koren, G., Kuint, J., Foxman, B., Chodick, G., Shalev, V., and Kishony, R. (2022) 'Minimizing treatment-induced emergence of antibiotic resistance in bacterial infections', *Science*, 375(6583), pp. 889–894. doi: 10.1126/SCIENCE.ABG9868/SUPPL_FILE/SCIENCE.ABG9868_DATA_S1.ZIP.

Sun, D. L., Jiang, X., Wu, Q. L., and Zhou, N. Y. (2013) 'Intragenomic heterogeneity of 16S rRNA genes causes overestimation of prokaryotic diversity', *Applied and Environmental Microbiology*, 79(19), pp. 5962–5969. doi: 10.1128/AEM.01282-13.

Sun, J., Huang, T., Chen, C., Cao, T.-T., Cheng, K., Liao, X.-P., and Liu, Y.-H. (2017) 'Comparison of Fecal Microbial Composition and Antibiotic Resistance Genes from Swine, Farm Workers and the Surrounding Villagers', *Scientific Reports*, 7(1), p. 4965. doi: 10.1038/s41598-017-04672-y.

Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. A., Kultima, J. R., Coelho, L. P., Arumugam, M., Tap, J., Nielsen, H. B., Rasmussen, S., Brunak, S., Pedersen, O., Guarner, F., De Vos, W. M., Wang, J., Li, J., Doré, J., Dusko Ehrlich, S., Stamatakis, A., and Bork, P. (2013) 'Metagenomic species profiling using universal phylogenetic marker genes', *Nature Methods*, 10(12), pp. 1196–1199. doi: 10.1038/nmeth.2693.

Takahashi, K., Nishida, A., Fujimoto, T., Fujii, M., Shioya, M., Imaeda, H., Inatomi, O., Bamba, S., Andoh, A., and Sugimoto, M. (2016) 'Reduced Abundance of Butyrate-Producing Bacteria Species in the Fecal Microbial Community in Crohn's Disease', *Digestion*, 93(1), pp. 59–65. doi: 10.1159/000441768.

Tang-Feldman, Y. J., Henderson, J. P., Ackermann, G., Feldman, S. S., Bedley, M., Silva, J., and Cohen, S. H. (2005) 'Prevalence of the *ermB* gene in *Clostridium difficile* strains isolated at a university teaching hospital from 1987 through 1998', *Clinical Infectious Diseases*, 40(10), pp. 1537–1540. doi: 10.1086/428835/2/40-10-1537-FIG002.GIF.

Tange, O. (2020) *GNU Parallel* (Version 20201122 ('Biden')) [Computer program]. doi: https://doi.org/10.5281/zenodo.4284075.

Tanimoto, K. and Ike, Y. (2008) 'Complete nucleotide sequencing and analysis of the 65-kb highly conjugative *Enterococcus faecium* plasmid pMG1: identification of the transfer-related region and the minimum region required for replication', *FEMS Microbiology Letters*, 288(2), pp. 186–195. doi: 10.1111/J.1574-6968.2008.01342.X.

Tanizawa, H., Iwasaki, O., Tanaka, A., Capizzi, J. R., Wickramasinghe, P., Lee, M., Fu, Z., and Noma, K.-I. (2010) 'Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation', *Nucleic Acids Research*, 38(22), pp. 8164–8177. doi: 10.1093/nar/gkq955.

Tanizawa, H. and Noma, K. ichi (2012) 'Unravelling global genome organization by 3C-seq', *Seminars in Cell and Developmental Biology*, 23(2), pp. 213–221. doi: 10.1016/j.semcdb.2011.11.003.

Tchesnokova, V. L., Rechkina, E., Chan, D., Haile, H. G., Larson, L., Ferrier, K., Schroeder, D. W., Solyanik, T., Shibuya, S., Hansen, K., Ralston, J. D., Riddell, K., Scholes, D., and Sokurenko, E. V. (2020) 'Pandemic Uropathogenic Fluoroquinolone-resistant *Escherichia coli* Have Enhanced Ability to Persist in the Gut and Cause Bacteriuria in Healthy Women', *Clinical Infectious Diseases*, 70(5), pp. 937–939. doi: 10.1093/CID/CIZ547.

Tenaillon, O., Skurnik, D., Picard, B., and Denamur, E. (2010) 'The population genetics of commensal *Escherichia coli*', *Nature Reviews Microbiology*, 8(3), pp. 207–217. doi: 10.1038/nrmicro2298.

Textco BioSoftware, I. (2022) *Gene Construction Kit* (Version 4.5) [Computer program]. Available at: http://www.textco.com/gene-construction-kit.php.

Thermo Fisher Scientific (2017) *qPCR Efficiency Calculator*. Available at: https://www.thermofisher.com/uk/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/qpcr-efficiency-calculator.html (Accessed: 23 May 2019).

Theuretzbacher, U., Bush, K., Harbarth, S., Paul, M., Rex, J. H., Tacconelli, E., and Thwaites, G. E. (2020) 'Critical analysis of antibacterial agents in clinical development', *Nature Reviews Microbiology 2020 18:5*, 18(5), pp. 286–298. doi: 10.1038/s41579-020-0340-0.

Thursby, E. and Juge, N. (2017) 'Introduction to the human gut microbiota', *Biochemical Journal*, 474(11), p. 1823. doi: 10.1042/BCJ20160510.

Di Tommaso, N., Gasbarrini, A., and Ponziani, F. R. (2021) 'Intestinal Barrier in Human Health and Disease', *International Journal of Environmental Research and Public Health*, 18(23), p. 12836. doi: 10.3390/IJERPH182312836.

Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., Gladstone, R. A., Lo, S., Beaudoin, C., Floto, R. A., Frost, S. D. W., Corander, J., Bentley, S. D., and Parkhill, J. (2020) 'Producing polished prokaryotic pangenomes with the Panaroo pipeline', *Genome Biology*, 21(1), pp. 1–21. doi: 10.1186/S13059-020-02090-4/FIGURES/7.

Tóth, A. G., Tóth, I., Rózsa, B., Dubecz, A., Patai, Á. V., Németh, T., Kaplan, S., Kovács, E. G., Makrai, L., and Solymosi, N. (2022) 'Canine Saliva as a Possible Source of Antimicrobial Resistance Genes', *Antibiotics 2022, Vol. 11, Page 1490*, 11(11), p. 1490. doi: 10.3390/ANTIBIOTICS11111490.

Tramontano, M., Andrejev, S., Pruteanu, M., Klünemann, M., Kuhn, M., Galardini, M., Jouhten, P., Zelezniak, A., Zeller, G., Bork, P., Typas, A., and Patil, K. R. (2018) 'Nutritional preferences of human gut bacteria reveal their metabolic idiosyncrasies', *Nature Microbiology 2018 3:4*, 3(4), pp. 514–522. doi: 10.1038/s41564-018-0123-9.

Tran, C. M., Tanaka, K., and Watanabe, K. (2013) 'PCR-based detection of resistance genes in anaerobic bacteria isolated from intra-abdominal infections', *Journal of*

Infection and Chemotherapy, 19(2), pp. 279–290. doi: 10.1007/S10156-012-0532-2/TABLES/4.

Tyagi, A., Singh, B., Billekallu Thammegowda, N. K., and Singh, N. K. (2019) 'Shotgun metagenomics offers novel insights into taxonomic compositions, metabolic pathways and antibiotic resistance genes in fish gut microbiome', *Archives of Microbiology*, 201(3), pp. 295–303. doi: 10.1007/s00203-018-1615-y.

Uelze, L., Grützke, J., Borowiak, M., Hammerl, J. A., Juraschek, K., Deneke, C., Tausch, S. H., and Malorny, B. (2020) 'Typing methods based on whole genome sequencing data', *One Health Outlook 2020 2:1*, 2(1), pp. 1–19. doi: 10.1186/S42522-020-0010-1.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., and Rozen, S. G. (2012) 'Primer3—new capabilities and interfaces', *Nucleic Acids Research*, 40(15), p. e115. doi: 10.1093/NAR/GKS596.

Uritskiy, G., Press, M., Sun, C., Huerta, G. D., Zayed, A. A., Wiser, A., Grove, J., Auch, B., Eacker, S. M., Sullivan, S., Bickhart, D. M., Smith, T. P. L., Sullivan, M. B., and Liachko, I. (2021) 'Accurate viral genome reconstruction and host assignment with proximity-ligation sequencing', *bioRxiv*, p. 2021.06.14.448389. doi: 10.1101/2021.06.14.448389.

Valerio, M., Pedromingo, M., Muñoz, P., Alcalá, L., Marin, M., Peláez, T., Giannella, M., and Bouza, E. (2012) 'Potential protective role of linezolid against *Clostridium difficile* infection', *International Journal of Antimicrobial Agents*, 39(5), pp. 414–419. doi: 10.1016/J.IJANTIMICAG.2012.01.005.

Veloo, A. C. M., Baas, W. H., Haan, F. J., Coco, J., and Rossen, J. W. (2019) 'Prevalence of antimicrobial resistance genes in *Bacteroides* spp. and *Prevotella* spp. Dutch clinical isolates', *Clin Microbiol Infect*, 25. doi: 10.1016/j.cmi.2019.02.017.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y. H., and Smith, H. O. (2004) 'Environmental Genome Shotgun Sequencing of the Sargasso Sea', *Science*, 304(5667), pp. 66–74. doi: 10.1126/SCIENCE.1093857/SUPPL_FILE/VENTER.SOM.PDF.

Vidor, C. J., Bulach, D., Awad, M., and Lyras, D. (2019) '*Paeniclostridium sordellii* and *Clostridioides difficile* encode similar and clinically relevant tetracycline resistance loci in diverse genomic locations', *BMC Microbiology*, 19(1), pp. 1–12. doi: 10.1186/S12866-019-1427-5/FIGURES/3.

Vincent, J. L. (2003) 'Nosocomial infections in adult intensive-care units', *The Lancet*, 361(9374), pp. 2068–2077. doi: 10.1016/S0140-6736(03)13644-6.

Viswanathan, V. K. (2014) 'Off-label abuse of antibiotics by bacteria', *Gut Microbes*, 5(1), p. 3. doi: 10.4161/GMIC.28027.

Wall, J. D., Weaver, P. F., and Gest, H. (1975) 'Gene transfer agents, bacteriophages, and bacteriocins of *Rhodopseudomonas capsulata*', *Archives of Microbiology 1975 105:1*, 105(1), pp. 217–224. doi: 10.1007/BF00447140.

Wang, Ziye, Wang, Zhengyang, Lu, Y. Y., Sun, F., and Zhu, S. (2019) ‘SolidBin: improving metagenome binning with semi-supervised normalized cut’, *Bioinformatics (Oxford, England)*, 35(21), pp. 4229–4238. doi: 10.1093/BIOINFORMATICS/BTZ253.

Warren, Y. A., Tyrrell, K. L., Citron, D. M., and Goldstein, E. J. C. (2006) ‘*Clostridium aldenense* sp. nov. and *Clostridium citroniae* sp. nov. Isolated from Human Clinical Infections’, *Journal of Clinical Microbiology*, 44(7), p. 2416. doi: 10.1128/JCM.00116-06.

Waters, J. L. and Salyers, A. A. (2013) ‘Regulation of CTnDOT conjugative transfer is a complex and highly coordinated series of events’, *mBio*, 4(6). doi: 10.1128/mBio.00569-13.

Webber, M. A., Whitehead, R. N., Mount, M., Loman, N. J., Pallen, M. J., and Piddock, L. J. V. (2015) ‘Parallel evolutionary pathways to antibiotic resistance selected by biocide exposure’, *Journal of Antimicrobial Chemotherapy*, 70(8), p. 2241. doi: 10.1093/JAC/DKV109.

Weingarten, R. A., Johnson, R. C., Conlan, S., Ramsburg, A. M., Dekker, J. P., Lau, A. F., Khil, P., Odom, R. T., Deming, C., Park, M., Thomas, P. J., Program, N. C. S., Henderson, D. K., Palmore, T. N., Segre, J. A., and Frank, K. M. (2018) ‘Genomic Analysis of Hospital Plumbing Reveals Diverse Reservoir of Bacterial Plasmids Conferring Carbapenem Resistance’, *mBio*, 9(1), pp. e02011-17. doi: 10.1128/MBIO.02011-17.

Wick, R. (2022) *Filtlong* (Version 0.2.1) [Computer program]. Available at: https://github.com/rrwick/Filtlong.

Wick, R. R., Schultz, M. B., Zobel, J., and Holt, K. E. (2015) ‘Bandage: interactive visualization of de novo genome assemblies’, *Bioinformatics*, 31(20), pp. 3350–3352. doi: 10.1093/BIOINFORMATICS/BTV383.

Wick, R. R., Judd, L. M., Gorrie, C. L., and Holt, K. E. (2017) ‘Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads’, *PLOS Computational Biology*, 13(6), p. e1005595. doi: 10.1371/JOURNAL.PCBI.1005595.

Williams, O. M., Brazier, J., Peraino, V., and Goldstein, E. J. C. (2010) ‘A review of three cases of *Clostridium aldenense* bacteremia’, *Anaerobe*, 16(5), pp. 475–477. doi: 10.1016/J.ANAEROBE.2010.08.004.

Von Wintersdorff, C. J. H., Penders, J., Van Niekerk, J. M., Mills, N. D., Majumder, S., Van Alphen, L. B., Savelkoul, P. H. M., and Wolffs, P. F. G. (2016) ‘Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer’, *Frontiers in Microbiology*, 7(FEB), p. 173. doi: 10.3389/FMICB.2016.00173/BIBTEX.

de Wit, E. and de Laat, W. (2012) ‘A decade of 3C technologies: insights into nuclear organization.’, *Genes & development*, 26(1), pp. 11–24. doi: 10.1101/gad.179804.111.

Wong, D., Aoki, F., and Rubinstein, E. (2014) ‘Bacteremia caused by *Eggerthella lenta* in an elderly man with a gastrointestinal malignancy: A case report’, *The Canadian Journal of Infectious Diseases & Medical Microbiology*, 25(5), p. e85. doi: 10.1155/2014/802481.

Wood, D. E., Lu, J., and Langmead, B. (2019) 'Improved metagenomic analysis with Kraken 2', *Genome Biology*, 20(1), p. 257. doi: 10.1186/s13059-019-1891-0.

Wood, D. E. and Salzberg, S. L. (2014) 'Kraken: Ultrafast metagenomic sequence classification using exact alignments', *Genome Biology*, 15(3), pp. 1–12. doi: 10.1186/GB-2014-15-3-R46/FIGURES/5.

Woodcroft, B. J. (2022) *CoverM* (Version 0.6.0) [Computer program]. Available at: https://github.com/wwood/CoverM.

World Health Organization (2020) *Antibiotic resistance*. Available at: https://www.who.int/news-room/fact-sheets/detail/antibiotic-resistance (Accessed: 4 September 2020).

Wu, W. K., Chen, C. C., Panyod, S., Chen, R. A., Wu, M. S., Sheen, L. Y., and Chang, S. C. (2019) 'Optimization of fecal sample processing for microbiome study — The journey from bathroom to bench', *Journal of the Formosan Medical Association*, 118(2), pp. 545–555. doi: 10.1016/J.JFMA.2018.02.005.

Xiao, L., Feng, Q., Liang, S., Sonne, S. B., Xia, Z., Qiu, X., Li, X., Long, H., Zhang, J., Zhang, D., Liu, C., Fang, Z., Chou, J., Glanville, J., Hao, Q., Kotowska, D., Colding, C., Licht, T. R., Wu, D., Yu, J., Sung, J. J. Y., Liang, Q., Li, J., Jia, H., Lan, Z., Tremaroli, V., Dworzynski, P., Nielsen, H. B., Bäckhed, F., Doré, J., Le Chatelier, E., Ehrlich, S. D., Lin, J. C., Arumugam, M., Wang, J., Madsen, L., and Kristiansen, K. (2015) 'A catalog of the mouse gut metagenome', *Nature Biotechnology*, 33(10), pp. 1103–1108. doi: 10.1038/nbt.3353.

Xie, H., Yang, C., Sun, Y., Igarashi, Y., Jin, T., and Luo, F. (2020) 'PacBio Long Reads Improve Metagenomic Assemblies, Gene Catalogs, and Genome Binning', *Frontiers in Genetics*, 11, p. 1077. doi: 10.3389/FGENE.2020.516269/BIBTEX.

Yaffe, E. and Relman, D. A. (2020) 'Tracking microbial evolution in the human gut using Hi-C reveals extensive horizontal gene transfer, persistence and adaptation', *Nature Microbiology*, 5(2), pp. 343–353. doi: 10.1038/s41564-019-0625-0.

Yang, C., Chowdhury, D., Zhang, Z., Cheung, W. K., Lu, A., Bian, Z., and Zhang, L. (2021) 'A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data', *Computational and Structural Biotechnology Journal*, 19, pp. 6301–6314. doi: 10.1016/J.CSBJ.2021.11.028.

Yang, F., Sun, J., Luo, H., Ren, H., Zhou, H., Lin, Y., Han, M., Chen, B., Liao, H., Brix, S., Li, J., Yang, H., Kristiansen, K., and Zhong, H. (2020) 'Assessment of fecal DNA extraction protocols for metagenomic studies', *GigaScience*, 9(7), pp. 1–12. doi: 10.1093/gigascience/giaa071.

Ye, S. H., Siddle, K. J., Park, D. J., and Sabeti, P. C. (2019) 'Benchmarking Metagenomics Tools for Taxonomic Classification', *Cell*, 178(4), p. 779. doi: 10.1016/J.CELL.2019.07.010.

You, D., Su, Y., Sun, X., Wang, J., Zheng, Y., and Liu, Y. (2022) 'Linezolid in the treatment of severe intraabdominal infection: A STROBE-compliant retrospective study', *Medicine*, 101(33), p. E30038. doi: 10.1097/MD.0000000000030038.

Zhang, X., De Maat, V., Guzmán Prieto, A. M., Prajsnar, T. K., Bayjanov, J. R., De Been, M., Rogers, M. R. C., Bonten, M. J. M., Mesnage, S., Willems, R. J. L., and Van Schaik, W. (2017) 'RNA-seq and Tn-seq reveal fitness determinants of vancomycin-resistant *Enterococcus faecium* during growth in human serum', *BMC Genomics*, 18(1), p. 893. doi: 10.1186/s12864-017-4299-9.

Zhang, X., Shi, L., Sun, T., Guo, K., and Geng, S. (2021) 'Dysbiosis of gut microbiota and its correlation with dysregulation of cytokines in psoriasis patients', *BMC Microbiology*, 21(1), pp. 1–10. doi: 10.1186/S12866-021-02125-1/FIGURES/4.

Zou, Y., Xue, W., Luo, G., Deng, Z., Qin, P., Guo, R., Sun, H., Xia, Y., Liang, S., Dai, Y., Wan, D., Jiang, R., Su, L., Feng, Q., Jie, Z., Guo, T., Xia, Z., Liu, C., Yu, J., Lin, Y., Tang, S., Huo, G., Xu, X., Hou, Y., Liu, X., Wang, J., Yang, H., Kristiansen, K., Li, J., Jia, H., and Xiao, L. (2019) '1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses', *Nature Biotechnology 2019 37:2*, 37(2), pp. 179–185. doi: 10.1038/s41587-018-0008-8.

APPENDIX

A.1 Extracting ARGs from contigs

```
grep -v "SEQUENCE" [Abricate_file].tsv | while read sample contig start
end strand arg_full rest ; do arg=$(echo "$arg_ful" | cut -f6 | sed
s/"'"/""/g | sed s/"-"/""/g | sed s/"\."/""/g | sed 's/[()]//g') &&
size=$(echo "${end} - ${start} + 1" | bc) && length=$(grep
"\<${contig}\>" [Assembly].fa | cut -d" " -f4 | sed 's!len=!!') &&
distance=$(echo "${length} - ${end} + 1" | bc) && sequence=$(grep -A1
"\<${contig}\>" [Assembly].fa | awk 'BEGIN{ cmd = "cut -c '${start}'- |
rev | cut -c '${distance}'- | rev" } NR % 2 { print } NR % 2 == 0 {
print | cmd; close(cmd) }' | tail -1) && title=$(grep "\<${contig}\>"
[Assembly].fa | cut -d" " -f1) && (echo ${title}_${arg}_len=${size} &&
echo $sequence) >> all_args.fa ; done ; done
```

A.2 Bash script to remove intercontig reads mapping within first or last 500 nt of a contig

```
# Output sam file containing reads mapping within 500 nt of the start
of end of a contig, and a separate sam file for reads mapping outside
of 500 nt of the start or end of a contig
for i in *intercontig.sam ; do cat $i | while read -r name qual
mappedto position rest ; do grep "\<$mappedto\>"
contig_lengths/${i/_intercontig.sam/}_contig_lengths.tsv | while read -
r contig length ; do if (( $position < 501 )) ; then echo -e
${name}'\t'${qual}'\t'${mappedto}'\t'${length}'\t'${position}'\t'"${res
t}" >> ${i/.sam/_within.sam} ; else if distance=$(echo "${length} -
${position}" | bc) && (( $distance < 501 )) ; then echo -e
${name}'\t'${qual}'\t'${mappedto}'\t'${length}'\t'${position}'\t'"${res
t}" >> ${i/.sam/_within.sam} ; else echo -e
${name}'\t'${qual}'\t'${mappedto}'\t'${length}'\t'${position}'\t'"${res
t}" >> ${i/.sam/_not.sam} ; fi ; fi ; done ; done ; done
```

A.3 Bash script to link ARGs to hosts (H-LARGe v1)

```
#create data files:
# [SAMPLE]_arg_contigs.tsv - two columns showing contig_name,ARG_name
# [SAMPLE]_is_list - file containing IS element contig names only
# [SAMPLE]_contigs.fa - assembly file

#create variables to easily access all the data files:
export READ_SUFFIX=[SAMPLE]
export DATASET_DIR=[DATA/DIRECTORY]
export HEATMAP_DIR=[DIRECTORY/WHERE/YOU/WANT/YOUR/HEATMAP/TABLE/FILES]
export ARG_CONTIGS=${DATASET_DIR}/${READ_SUFFIX}_arg_contigs.tsv
export IS_LIST=${DATASET_DIR}/${READ_SUFFIX}_is_list
export ASSEMBLY=${DATASET_DIR}/${READ_SUFFIX}_contigs.fa

#ONLY APPLICABLE FOR SAMPLES WITH MULTIPLE EZNYMES USED FOR 3C/Hi-C
(with separate reads):
export ENZYME_1=[enzyme 1 suffix e.g. _H]
export ENZYME_2=[enzyme 2 suffix e.g. _M]

#in mapped directory:
#find contigs linked to ARGs (individual file for each ARG, named as
the ARG contig name e.g. k141_1234)
mkdir ${READ_SUFFIX}_linked_contigs
mkdir ${READ_SUFFIX}_linked_contigs/cut
tab=$'\t'
for i in ${READ_SUFFIX}*_intercontig_not.sam ; do cut -f1 $ARG_CONTIGS
| while read contig ; do grep "$contig$tab" $i >
${READ_SUFFIX}_linked_contigs/${contig}_${i/_intercontig_not.sam/} &&
cat ${READ_SUFFIX}_linked_contigs/${contig}_${i/_intercontig_not.sam/}
| while read line ; do column=$(echo "$line" | awk -v b="${contig}"
'{for (i=1;i<=NF;i++) { if ($i == b) { print i } }}') && if [[
"${column}" == "3" ]] ; then echo "$line" | cut -f1,9 >>
${READ_SUFFIX}_linked_contigs/cut/${contig}_${i/_intercontig_not.sam/}
; else if [[ "${column}" == "9" ]] ; then echo "$line" | cut -f1,3 >>
${READ_SUFFIX}_linked_contigs/cut/${contig}_${i/_intercontig_not.sam/}
; fi ; fi ; done ; done ; done
cd ${READ_SUFFIX}_linked_contigs/cut

#remove duplicates (currently have duplicate lines for both read 1 and
2 from the same pair)
mkdir unique
for i in k141_* ; do cat $i | sort | uniq > unique/$i ; done
cd unique

#get list of unique contigs linked to ARG
mkdir uniq
for i in k141_* ; do cut -f2 $i | sort | uniq > uniq/$i ; done
cd uniq

#get count for how many times each unique contig is linked to ARG
mkdir counts
for i in k141_* ; do cat $i | while read line ; do grep "\<$line\>"
../$i | wc -l | cat | while read word ; do echo -e ${word}'\t'${line}
>> counts/$i ; done ; done ; done
```

```
cd counts

#filter links so that a contig is only considered linked if linked by
>=5 intercontig read pairs
#also removes linked IS element contigs
mkdir unique_filtered
for i in k141* ; do sort $i | uniq | sort -nr | grep -v -P
'^1\tk141'\|'^2\tk141'\|'^3\tk141'\|'^4\tk141' | grep -vf $IS_LIST >
unique_filtered/${i} ; done
cd unique_filtered

#get sequences for linked contigs
mkdir contigs
for i in k141_* ; do cat $i | while read count contig ; do grep -A1
"\<${contig}\>" $ASSEMBLY >> contigs/${i}_contigs.fa ; done ; done

#classify linked contigs
cd contigs
#first blast them:
mkdir blasted
for i in k141* ; do blastn -query $i -db /DB/BLAST_DB/nt2/nt -
num_threads 16 -max_target_seqs 5 -max_hsps 1 -outfmt "7 qseqid sacc
bitscore evalue qcovs pident stitle" >
blasted/${i/_contigs.fa/_blasted} ; done
#then classify
mkdir kraken
for i in k141* ; do kraken2 --threads 16 --use-names --db
/DB/KRAKEN_DB/kraken2-microbial $i --output
kraken/${i/_contigs.fa/_kraken} ; done
cd ..
#get top BLAST results for each contig:
mkdir contigs/blast_first_results
for i in k141* ; do cut -f2 $i | while read line ; do grep "\<$line\>"
contigs/blasted/${i}_blasted | head -2 | tail -1 >>
contigs/blast_first_results/${i} ; done ; done
mkdir pasted
for i in k141* ; do cat $i | while read count contig ; do blast=$(grep
"\<$contig\>" contigs/blast_first_results/$i | cut -f7) &&
kraken=$(grep "\<$contig\>" contigs/kraken/${i}_kraken | cut -f3) &&
echo -e ${count}'\t'${contig}'\t'${blast}'\t'${kraken} >> pasted/$i ;
done ; done
cd pasted

#start making heatmap file
#get list of linked contig classifications and counts. At this point,
contigs that BLAST identified as plasmid DNA are labelled as "Plasmid
DNA", and the rest are labelled with their Kraken2 classification (down
to genus-level)
#also this is written for zsh - if using BASH, change ":u" to "^^" e.g.
[[ "${classification^^}" == *"PLASMID"* ]]
mkdir names
for i in k141* ;
do cat $i | while IFS=$'\t' read count contig blast kraken ;
do if [[ "${blast:u}" == *"PLASMID"* ]] ; then
echo -e ${count}'\t'"Plasmid DNA" ;
```

```
else kraken_classification=$(echo "${kraken}" | cut -d" " -f1 | sed
's/(taxid//' | sed 's/[][]//g' | sed s/"'"/""/g | sed s/"-"/""/g | sed
s/"\."/""/g)
&& echo -e ${count}'\t'${kraken_classification} ;
fi >> names/$i ; done ; done
cd names
#if multiple enzymes are used, now combine the files. If not then skip:
mkdir combined
for i in k141*$ENZYME_1* ; do cat $i >> combined/${i/$ENZYME_1/} ; done
for i in k141*$ENZYME_2* ; do cat $i >> combined/${i/$ENZYME_2/} ; done
cd combined

#get proportions for links to each unique classification (contigs
labelled "Plasmid DNA" removed at this point):
mkdir added
for i in k141_* ; do total=$(grep -v "Plasmid DNA" $i | cut -f1 | paste
-sd+ | bc) && grep -v "Plasmid DNA" $i | while read count name ; do
combined=$(grep "\<${name}\>$" $i | cut -f1 | paste -sd+ | bc) &&
proportion=$(echo "scale=6 ; ${combined} / ${total}" | bc | awk
'{printf "%.6f\n", $0}') && echo -e
${combined}'\t'${name}'\t'${proportion} ; done | sort | uniq > added/$i
; done
cd added
#get list of all classifications linked to ARGs
for i in k141_* ; do cat $i | cut -f2 ; done | sort | uniq >
classification_list

#make heatmap table
#first convert each file into a list of all classifications and
proportion of links to each (no links = 0, all links = 1)
mkdir columns
for i in k141_* ; do cat classification_list | while read name ; do cat
$i | while IFS=$'\t' read count title proportion ; do if [[ "${title}"
== "${name}" ]] ; then echo -e ${name}'\t'${proportion} ; else echo -e
${name}'\t'0 ; fi ; done | sort -r | uniq | grep -m1 "\<${name}\>" ;
done > columns/columns_${i} ; done
cd columns
#add ARG names to top of list (the arg_rem part is so the filename
contains the ARG without any special characters)
for i in columns_k141_* ; do b=${i/columns_/} &&
c=${b/_${READ_SUFFIX}/} && grep "\<${c}\>" $ARG_CONTIGS | while read
contig arg ; do arg_rem=$(echo ${arg//[\(\)]/} | sed s/"'"/""/g | sed
s/"-"/""/g | sed s/"\."/""/g) && (echo ${arg} && cut -f2 $i) >
${i/columns_/}_${arg_rem} ; done ; done

#get list of classifications linked to ARGs with blank first line (for
column 1 of heatmap)
(echo -en '\n' && cat ../classification_list) > heatmap_list

#make and run command for creating heatmap table, and copy the table to
your heatmap output directory
(echo paste heatmap_list && for i in k141_* ; do echo $i ; done && echo
"| sed 's/\t/,/g' > ${READ_SUFFIX}_heatmap_table_genus_no_plasmid.csv")
| sed ':a;N;$!ba;s/\n/ /g' > command.txt
parallel -j1 < command.txt
cp ${READ_SUFFIX}_heatmap_table_genus_no_plasmid.csv ${HEATMAP_DIR}
```

A.4 R script to find annotated region of genome being mapped to by each read

```
# First, on command line:
# First step is to strip relevant info from the .gff file
# [Genome] = corresponding spike-in genome
grep "CDS" [genome].gff| sed -e 's/;/\t/g' |cut -f 4,5,9,13 >
[Genome]_genes.tsv
# grep "CDS" e745.gff # Strips lines containing "CDS", ignoring headers
and sequence.
# | sed -e 's/;/\t/g' # Replaces ; with a tab, to allow separation of
important info
# |cut -f 4,5,9,13 # Uses tabs to differentiate between fields, prints
start, stop, Locus tag and product.
# Also strip the sam header to load into R:
tail -n +9 [mapping file of either WGS or 3C reads].sam >
[Genome]_sam_lookup.tsv

# Now in R
# Install and load packages:
install.packages("dplyr")
install.packages("tidyr")
install.packages("purrr")
library(dplyr)
library(tidyr)
library(purrr)
# Set working directory
setwd('/[working_directory]')
# Read file generated from .gff
[Genome]_genes <-
read.csv(file="[Genome]_genes.tsv",header=FALSE,sep='\t')
# Set column names
colnames([Genome]_genes) <- c('Start','Stop','Locus', 'Product')
# Create new df with seq 1:max
# Can check max coord position with max([Genome]_genes$Stop)
lookup_[Genome] <- data.frame(Pos = seq(1,5270022))
# Match columns based on start/stop co-ords
matched_[Genome] <- [Genome]_genes %>% mutate(Pos = map2(Start, Stop,
`:`)) %>%
      unnest(Pos) %>% select(3:5) %>% right_join(lookup_[Genome]) %>%
      arrange(Pos) %>% select(3,1,2)
# Read processed sam file
[Genome]_sam <- read.csv(file='[Genome]_sam_lookup.tsv', header=FALSE,
sep='\t')
# Extract interesting columns
sam_lookup_[Genome] <- [Genome]_sam[,c(1,4)]
# Set column names
colnames(sam_lookup_[Genome]) <- c('Info', 'Pos')
# Match mapped reads to genes
output_[Genome] <- merge(x = matched_[Genome], y = sam_lookup_[Genome],
by = 'Pos')
write.table(output_[Genome], file = "[Genome]_[WGS/3C]_links.tsv",
quote=FALSE, sep='\t', row.names = FALSE)
```

A.5 Bash script to assign labels to genome regions

```
for i in *links.tsv ; do cat $i | while read line ; do product=$(echo
"$line" | cut -f3) && if [[ $product == "NA" ]] ; then echo -e
"${line}"'\t'intergenic ; else if [[ $product == *"IS"* ]] ; then echo
-e "${line}"'\t'is_element ; else if [[ $product == *"transposase"* ]]
; then echo -e "${line}"'\t'transposon ; else if [[ $product ==
*"hypothetical protein"* ]] ; then echo -e
"${line}"'\t'hypotehtical_protein ; else if [[ $product ==
*"prediction"* ]] ; then echo -e "${line}"'\t'hypotehtical_protein ;
else if [[ $product == *"product=putative protein"* ]] ; then echo -e
"${line}"'\t'hypotehtical_protein ; else if [[ $product == *"gene="* ]]
; then echo -e "${line}"'\t'annotated_gene ; else if [[ $product ==
*"locus_tag"* ]] ; then echo -e "${line}"'\t'annotated_gene ; else if
[[ $product == *"db_xref"* ]] ; then echo -e
"${line}"'\t'annotated_gene ; else if [[ $product == *"protein"* ]] ;
then echo -e "${line}"'\t'annotated_gene ; else if [[ $product ==
*"note"* ]] ; then echo -e "${line}"'\t'annotated_gene ; else if [[
$product == *"product"* ]] ; then echo -e "${line}"'\t'annotated_gene ;
else echo -e "${line}"'\t'other ; fi ; fi ; fi ; fi ; fi ; fi ; fi ; fi
; fi ; fi ; fi ; fi ; done > ${i/.tsv/_labelled.tsv} ; done
```

A.6 Bash script to calculate proportion of reads mapping within the first or last 500 nt of a contig

```
# First find lengths of contigs in assembly
grep ">" [Assembly].fa | cut -d" " -f1,4 | sed 's!len=!!' | sed
's/\s/\t/'g > [Assembly]_contig_lengths.tsv
# Prepare cut first, third, and fourth column from sam file
for i in *intercontig.sam ; do cut -f1,3,4 $i > ${i/.sam/.tsv} ; done
# Calculate position of alignment for all reads
for i in *intercontig.tsv ; do cat $i | while read -r name mappedto
position ; do grep "\<$mappedto\>" [Assembly]_contig_lengths.tsv |
while read -r contig length ; do if (( $position < 501 )) ; then echo -
e
${name}'\t'${mappedto}'\t'${length}'\t'${position}'\t'${position}'\t'wi
thin'\t'start ; else if distance=$(echo "${length} - ${position}" | bc)
&& (( $distance < 501 )) ; then echo -e
${name}'\t'${mappedto}'\t'${length}'\t'${position}'\t'${distance}'\t'wi
thin'\t'end ; else echo -e
${name}'\t'${mappedto}'\t'${length}'\t'${position}'\t'${distance}'\t'no
t'\t' ; fi ; fi ; done ; done > ${i/.tsv/_positions.tsv} ; done
# Find total number of reads mapping within 500 nt of the ends of a
contig or not
# File types.txt = file containing lines "within" and "not"
for i in *positions.tsv ; do cat types.txt| while read line ; do grep
"\<$line\>" $i | wc -l | cat | while read word ; do echo -e
${word}'\t'${line} ; done ; done > ${i/.tsv/_totals.tsv} ; done
```

A.7 Bash script to link ARGs to hosts (H-LARGe v2)

```
#create data files:
# [SAMPLE]_arg_contigs.tsv - two columns showing contig_name,ARG_name
# [SAMPLE]_arg_list - file containing ARG-contig names only
# [SAMPLE]_is_list - file containing IS element contig names only
# [SAMPLE]_contigs.fa - assembly file
# [SAMPLE]_full_contigs_list.tsv - file showing classification for each
contig. Columns showing contig_name,bin_name,bin_classification

#create variables to easily access all the data files:
export READ_SUFFIX=[SAMPLE]
export DATASET_DIR=[DATA/DIRECTORY]
export HEATMAP_DIR=[DIRECTORY/WHERE/YOU/WANT/YOUR/HEATMAP/TABLE/FILES]
export ARG_CONTIGS=${DATASET_DIR}/${READ_SUFFIX}_arg_contigs.tsv
export ARG_LIST=${DATASET_DIR}/${READ_SUFFIX}_arg_list
export IS_LIST=${DATASET_DIR}/${READ_SUFFIX}_is_list
export ASSEMBLY=${DATASET_DIR}/${READ_SUFFIX}_contigs.fa
export CONTIGS_LIST=${DATASET_DIR}/${READ_SUFFIX}_full_contigs_list.tsv

#in mapped directory:
#find contigs linked to ARGs (individual file for each ARG, named as
the ARG contig name e.g. k141_1234)
mkdir ${READ_SUFFIX}_linked_contigs
mkdir ${READ_SUFFIX}_linked_contigs/cut
tab=$'\t'
for i in ${READ_SUFFIX}*_intercontig_not.sam ; do cut -f1 $ARG_CONTIGS
| while read contig ; do grep "$contig$tab" $i >
${READ_SUFFIX}_linked_contigs/${contig}_${i/_intercontig_not.sam/} &&
cat ${READ_SUFFIX}_linked_contigs/${contig}_${i/_intercontig_not.sam/}
| while read line ; do column=$(echo "$line" | awk -v b="${contig}"
'{for (i=1;i<=NF;i++) { if ($i == b) { print i } }}') && if [[
"${column}" == "3" ]] ; then echo "$line" | cut -f1,9 >>
${READ_SUFFIX}_linked_contigs/cut/${contig}_${i/_intercontig_not.sam/}
; else if [[ "${column}" == "9" ]] ; then echo "$line" | cut -f1,3 >>
${READ_SUFFIX}_linked_contigs/cut/${contig}_${i/_intercontig_not.sam/}
; fi ; fi ; done ; done ; done
cd ${READ_SUFFIX}_linked_contigs/cut

#remove duplicates (currently have duplicate lines for both read 1 and
2 from the same pair)
mkdir unique
for i in k141_* ; do cat $i | sort | uniq > unique/$i ; done
cd unique

#get list of unique contigs linked to ARG
mkdir uniq
for i in k141_* ; do cut -f2 $i | sort | uniq > uniq/$i ; done
cd uniq

#get count for how many times each unique contig is linked to ARG
mkdir counts
for i in k141_* ; do cat $i | while read line ; do grep "\<$line\>"
../$i | wc -l | cat | while read word ; do echo -e ${word}'\t'${line}
>> counts/$i ; done ; done ; done
```

```
cd counts

#filter links so that a contig is only considered linked if linked by
>=5 intercontig read pairs
#also removes linked IS element contigs, and links to other ARG-contigs
mkdir unique_filtered
for i in k141* ; do sort $i | uniq | sort -nr | grep -v -P
'^1\tk141'\|'^2\tk141'\|'^3\tk141'\|'^4\tk141' | grep -vf $IS_LIST |
grep -vf $ARG_LIST > unique_filtered/${i} ; done
cd unique_filtered

#get classifications of linked contigs
mkdir binned
for i in k141_* ; do cat $i | while read count contig ; do echo -e
${count}'\t'${contig}'\t'"$(grep "\<${contig}\>" $CONTIGS_LIST | cut -
f2,3)" >> binned/${i} ; done ; done
cd binned

#start making heatmap file
#get list of linked contig classifications and counts. Includes links
to plasmid and viral bins - if not applicable then remove:
#also this is written for zsh - if using BASH, change ":u" to "^^" e.g.
[[ "${classification^^}" == *"PLASMID"* ]]
mkdir names
for i in k141_* ; do cut -f1,4 $i | while read count classification ;
do if [[ "${classification:u}" == *"PLASMID"* ]] ; then echo -e
"${count}"'\t'"Plasmid bin" ; else if [[ "${classification:u}" ==
*"VMAG"* ]] ; then echo -e "${count}"'\t'"Viral bin" ; else if [[
"${classification}" == *"discarded"* ]] ; then echo -e
"${count}"'\t'"Discarded bin" ; else echo -e $count'\t'$classification
; fi ; fi ; fi ; done > names/$i ; done
cd names

#get proportions for links to each unique classification:
mkdir added
for i in k141_* ; do total=$(cut -f1 $i | paste -sd+ | bc) && cat $i |
while read count name ; do combined=$(grep "\<${name}\>$" $i | cut -f1
| paste -sd+ | bc) && proportion=$(echo "scale=6 ; ${combined} /
${total}" | bc | awk '{printf "%.6f\n", $0}') && echo -e
${combined}'\t'${name}'\t'${proportion} ; done | sort | uniq > added/$i
; done
cd added

#get list of all classifications linked to ARGs
for i in k141_* ; do cat $i | cut -f2 ; done | sort | uniq >
classification_list

#make heatmap table
#first convert each file into a list of all classifications and
proportion of links to each (no links = 0, all links = 1)
mkdir columns
for i in k141_* ; do cat classification_list | while read name ; do cat
$i | while IFS=$'\t' read count title proportion ; do if [[ "${title}"
== "${name}" ]] ; then echo -e ${name}'\t'${proportion} ; else echo -e
${name}'\t'0 ; fi ; done | sort -r | uniq | grep -m1 "\<${name}\>" ;
done > columns/columns_${i} ; done
```

```
cd columns

#add ARG names to top of list (the arg_rem part is so the filename
contains the ARG without any special characters)
for i in columns_k141_* ; do b=${i/columns_/} &&
c=${b/_${READ_SUFFIX}/} && grep "\<${c}\>" $ARG_CONTIGS | while read
contig arg ; do arg_rem=$(echo ${arg//[\(\)]/} | sed s/"'"/""/g | sed
s/"-"/""/g | sed s/"\."/""/g) && (echo ${arg} && cut -f2 $i) >
${i/columns_/}_${arg_rem} ; done ; done

#get list of classifications linked to ARGs with blank first line (for
column 1 of heatmap)
(echo -en '\n' && cat ../classification_list) > heatmap_list

#make and run command for creating heatmap table, and copy the table to
your heatmap output directory
(echo paste heatmap_list && for i in k141_* ; do echo $i ; done && echo
"| sed 's/\t/,/g' > ${READ_SUFFIX}_heatmap_table.csv") | sed
':a;N;$!ba;s/\n/ /g' > command.txt
parallel -j1 < command.txt
cp ${READ_SUFFIX}_heatmap_table.csv ${HEATMAP_DIR}
```

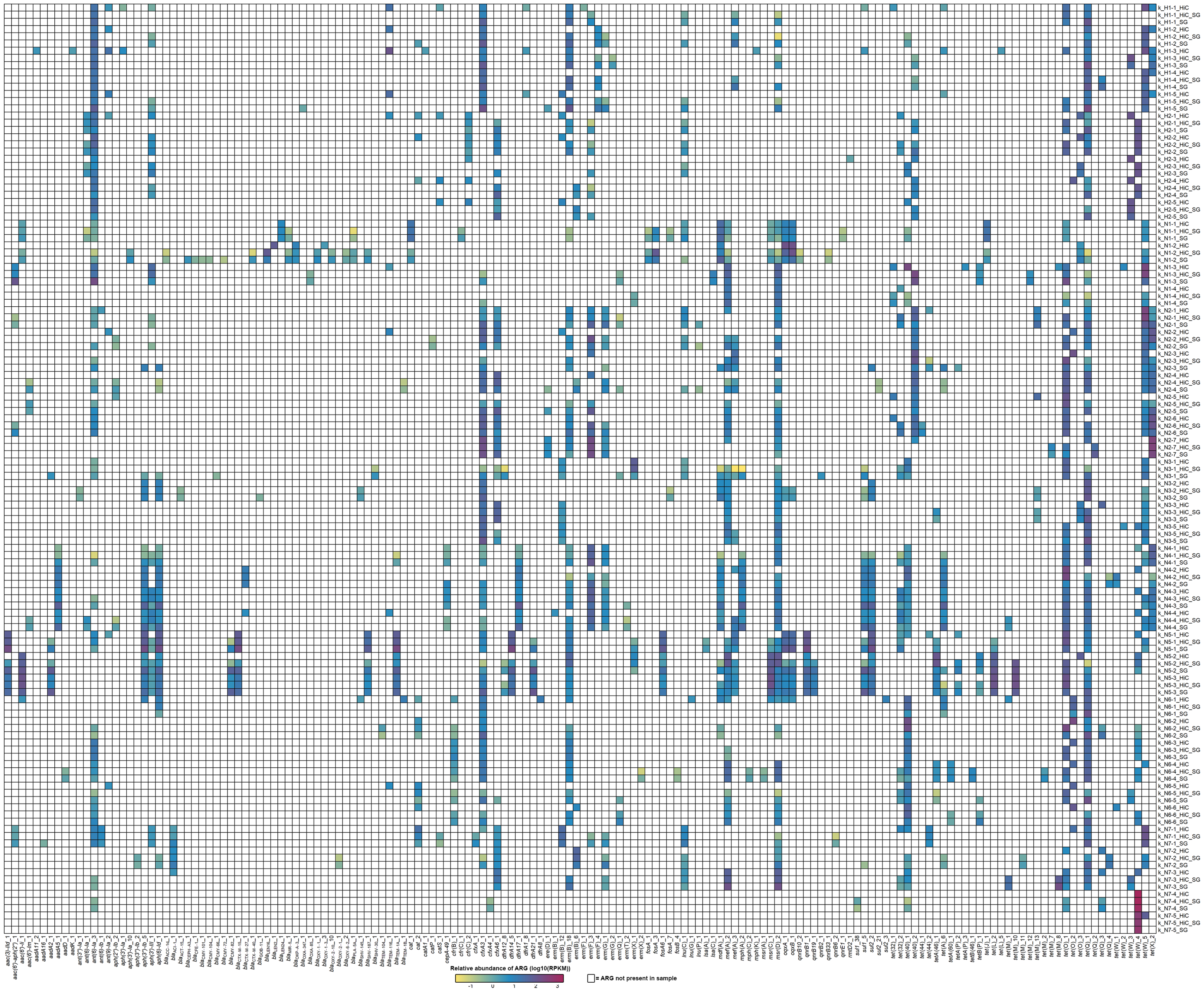


Figure A3.1. Relative abundance of antimicrobial resistance genes (ARGs) in the K_HiC dataset. Reads from each sample were mapped to the ARGs (columns) from the respective assembly. The relative abundance was calculated as reads per kilobase per million mapped reads (RPKM). White cells mean the ARG was not present, and coloured cells show that the ARG was present, with the colour relating to the relative abundance of the ARG within that set of reads (log(10) transformed RPKM values). Each row represents a different set of Hi-C or shotgun metagenomic reads (*_HiC/*_SG). Rows labelled *_HiC_SG show the RPKM of the Hi-C reads mapping to the ARGs identified in the shotgun assembly.

Table A4.1. Read statistics for the shallow sequenced Hi-C libraries

Dataset	H1 - shallow	H2 - shallow	H3 - shallow	H4 - shallow
Raw Reads	13,524,650	14,641,206	12,725,916	10,612,766
Deduplicated	9,173,078	12,804,868	10,547,650	9,304,936
% remaining	67.8%	87.5%	82.9%	87.7%
Adapter removal & quality filter	9,092,510	12,667,082	10,450,396	9,230,446
% remaining	67.2%	86.5%	82.1%	87.0%
Human reads removed	9,082,150	12,664,862	10,442,208	9,229,890
% remaining	67.2%	86.5%	82.1%	87.0%

Table A4.2. Intercontig read estimation for P_HiC and D_HiC datasets

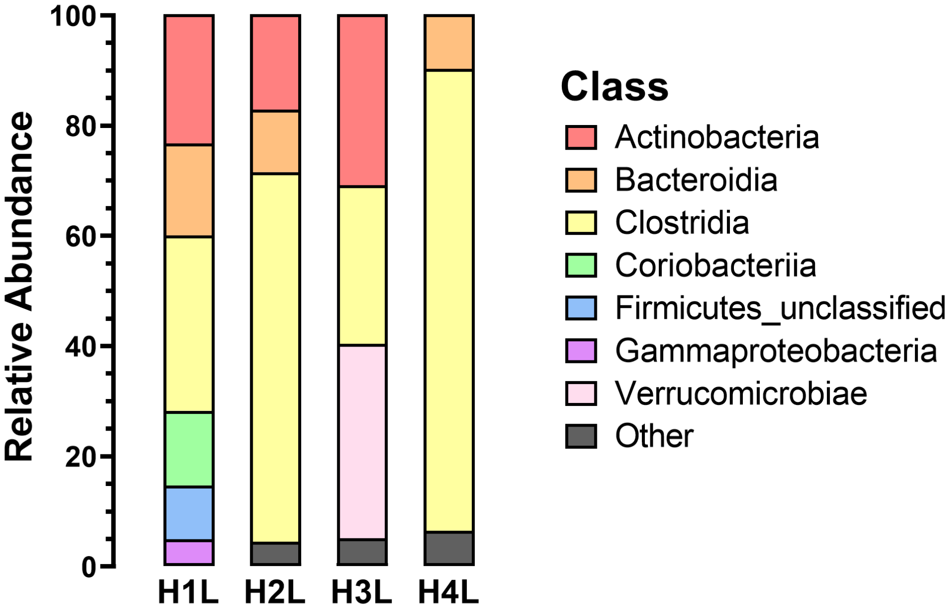
Dataset	P_HiC	D_HiC
Processed reads	157,755,162	133,509,800
Mapped to G_3C assembly	71,305,065	87,583,777
Percentage mapped	45.2%	65.6%
Mapped to G_3C assembly (MAPQ>20)	66,515,345	81,227,506
Percentage mapped (MAPQ>20)	42.2%	60.8%
Intercontig reads	8,874,710	6,804,105
% Intercontig (intercontig/processed)	5.6%	5.1%
% Intercontig (intercontig/mapped)	12.4%	7.8%
% Intercontig (intercontig/mapped MAPQ>20)	13.3%	8.4%
Actual intercontig %	13.7%	7.1%

MAPQ = mapping quality

Table A4.3. Intercontig read estimation for shallow sequenced Hi-C libraries

Dataset	H1 - shallow	H2 - shallow	H3 - shallow	H4 - shallow
Processed reads	9,082,150	2,664,862	10,442,208	9,229,890
Mapped to G_3C assembly	7,513,115	8,209,378	2,881,111	5,181,940
Percentage mapped	82.7%	64.8%	27.6%	56.1%
Mapped to G_3C assembly (MAPQ>20)	7,287,478	7,838,884	2,675,835	4,726,337
Percentage mapped (MAPQ>20)	80.2%	61.9%	25.6%	51.2%
Intercontig reads	1,938,029	4,046,260	279,954	1,568,924
% Intercontig (intercontig/processed)	21.3%	31.9%	2.7%	17.0%
% Intercontig (intercontig/mapped)	25.8%	49.3%	9.7%	30.3%
% Intercontig (intercontig/mapped>20)	26.6%	51.6%	10.5%	33.2%

MAPQ = mapping quality



% unclassified:	73.3	96.7	60.7	77.7
% OB7 spike-in:	1.1	0.01	0.2	0.08

Figure A4.1. Class-level composition of Hi-C reads from shallow sequencing.
The shallow sequenced Hi-C reads from each dataset were taxonomically profiled using MetaPhlAn3. The stacked bars show the relative abundance (%) of each class for the classified reads. Reads that could not be classified by MetaPhlAn3 (proportion unclassified showed below bars) are excluded here. The relative abundance of the *Acinetobacter pittii* strain OB7 spike-in are shown below the bars.