Application of ‘-omics’ technologies for diagnosis and surveillance of gastrointestinal infections

Saskia Neuert

Quadram Institute Bioscience

A thesis submitted for the degree of Doctor of Philosophy to the University of East Anglia

March 2020

Abstract

Application of '–omics' technologies for diagnosis and surveillance of gastrointestinal infections

The burden of gastrointestinal infections remains high, even in industrialised countries. Accurate, timely identification of enteropathogens is crucial for targeted treatment and global disease management. Routine diagnostic methods can be time-consuming, often rely on pathogen isolation prior to phenotypic characterisation, and still commonly fail to detect the aetiological agent. This thesis explores the use of '-omics' technologies as an alternative to conventional diagnostic and surveillance approaches.

A supervised learning algorithm using taxonomic and metabolic gut microbiota profiles derived from 16S rRNA gene sequencing and ^{1}H nuclear magnetic resonance spectroscopy of faeces from 41 controls and 246 gastroenteritis patients allowed distinction of bacterial and protozoal infections with a prediction accuracy of 81.61%.

Metagenomic sequencing of twenty *Salmonella enterica*-positive stool samples confirmed the presence of *S. enterica* in 70% of samples. Data was compared with whole genome sequences from the corresponding pathogen isolates. Metagenome coverage was insufficient for reconstruction of phylogenies and antimicrobial resistance profiling.

Comparison of antimicrobial resistance profiles inferred from whole genome sequences of 3,941 non-typhoidal *S. enterica* isolates with the results of phenotypic testing showed that resistance profiles encompassing fifteen antimicrobials were in complete agreement for 97.82% of isolates, with discordant results for only 0.17% of all possible isolate/antimicrobial combinations.

A longitudinal metagenomic study monitoring changes in the gut antimicrobial resistome in 48 travellers visiting regions with high prevalence of resistance revealed that travel led to a 1.27-fold increase in the number of antimicrobial resistance determinants and increased carriage of potentially pathogenic species in the cohort, with some changes persisting up to six months after return.

The results presented in this thesis provide evidence that some '–omics'-based approaches are suitable alternatives to traditional methods used for the diagnosis and surveillance of gastrointestinal infections, and that other tools could be of clinical use after further development work.

Table of contents

List of figures

List of tables

Abbreviations

ARD	Antimicrobial resistance determinant
AMR	antimicrobial resistance
AN	accession number
AST	antimicrobial susceptibility testing
BH	Benjamini-Hochberg
BSC	Bristol Stool Chart
BWA	Burrows-Wheeler Aligner
CAP	cationic antimicrobial peptide
CCA	constrained correspondence analysis
CDI	*Clostridioides difficile* infection
CFU	colony-forming unit
clr	centered log-ratio
CSF	cerebrospinal fluid
DNA	deoxyribonucleic acid
EAEC	enteroaggregative *Escherichia coli*
ECDC	European Centre for Disease Preventon and Control
ESBL	extended spectrum β-lactamase
ESBL-E	extended spectrum β-lactamase-producing Enterobacteriaceae
ETEC	enterotoxigenic *Escherichia coli*
EU	European Union
EUCAST	European Committee on Antimicrobial Susceptibility Testing
FC	fold change
GBRU	Gastrointestinal Bacteria Reference Unit
GI	gastrointestinal
LCA	lowest common ancestor
MALDI-TOF	matrix-assisted laser desorption/ionisation time-of-flight
MDR	multidrug resistance/resistant
ME	major error
MIC	minimum inhibitory concentration
MLS	macrolide lincosamide streptogramin
MLST	multilocus sequence type
MS	mass spectroscopy
NAAT	nucleic acid amplification test
NCBI	National Center for Biotechnology Information
NGS	next-generation sequencing
NHS	National Health Service
NMR	nuclear magnetic resonance
NNUH	Norfolk and Norwich University Hospitals
NTS	non-typhoidal *Salmonella enterica*
OTU	operational taxonomic unit
PBS	phosphate-buffered saline
PC	principal component

PCA	principal component analysis
PCoA	principal coordinate analysis
PCR	polymerase chain reaction
PD	phylogenetic diversity
PHE	Public Health England
PMQR	plasmid-mediated quinolone resistance
QIB	Quadram Institute Bioscience
QRDR	quinolone resistance-determining region
rRNA	ribosomal ribonucleic acid
SCFA	short chain fatty acid
SF	scaling factor
SNP	single nucleotide polymorphism
sPLS	sparse partial least squares
SRA	Short Read Archive
STEC	Shiga toxin-producing *Escherichia coli*
subsp.	subspecies
TD	traveller's diarrhoea
TSP	sodium 3-(trimethylsilyl)propionate
UCL	University College London
UK	United Kingdom
USA	United States of America
VME	very major error
WGS	whole genome sequences/sequencing

Symbols

±	plus/minus
>	greater than
<	less than
≤	less than or equal to
%	percent
°C	degrees Celsius
^{1}H	proton
bp	base pairs
g	gram
g/L	gram per litre
h	hour
Hz	Hertz
kg	kilogram
log	logarithm
mg	milligram
mg/L	milligram per litre
MHz	megahertz
min	minute
ml	millilitre
µl	microliter
mm	millimetre
mM	millimoles per litre
µm	micrometre
ms	milliseconds
m/s	metre per second
p	probability
pH	power of hydrogen
ppm	parts per million
R^2	coefficient of correlation
s	second
w/v	weight/volume
x	times
x g	times gravity

Acknowledgements

First of all, I would like to thank my supervisory team: Prof. Arjan Narbad for giving me the opportunity to work on this project. I have learned a lot – a little bit about the subject and a great deal about myself and life. Dr. Tim Dallman for believing that someone who had never even seen a command line could perform bioinformatics analyses. I still cannot code for the life of me but I got so much better at googling things. Prof. John Wain for stimulating questions and stepping in whenever necessary. Dr. Lee Kellingray for being my go-to advisor on everyday matters, at work and even more so outside of it. This experience would have been impossible without your friendship, inappropriate humour, sarcasm, comma policing, wine and provision of an emergency cigarette. I am also grateful to the National Institute for Health Research Health Protection Research Unit in Gastrointestinal Infections for providing the funding for this project.

Past and present members of the Narbad lab: Special thanks to Dr. Melinda Mayer for being the only reason this group continues to exist. It is safe to say that we all want to be a little bit like you when we grow up. Henry Whiley for personal poop collection services and extended coffee and lunch breaks. Dr. Maria Diaz, the sweetest, kindest metalhead known to men. My fellow PhD students, Dr. Fatma Cebeci for brutal honesty, Dr. Nichola Wong who I would happily put into a suitcase so I can take her wherever I go, Dr. Enriqueta Garcia for discussions about the state of the world and teaching me all the Spanish swear words, Jacob Scadden – aka pineapple – for constantly winding me up and being exactly the kind of friend I need (there, I said it, in writing, friend). My absolute lifeline over the past years, Dr. Bhavika Parmanand, thank you for brilliant detective work, being my literal and figurative shoulder to cry on and for sharing all the little highs and lows with me.

At IFR/QIB, thank you to Dr. John Walshaw and Dr. Andrea Telatin for (a lot of) bioinformatics support, Dr. Gwenaelle Le Gall and Dr. Ian Colquhoun for their assistance with my metabolomics experiments, Mark Philo for preparing litres of buffer and for pleasant company in the NMR outhouse and Dr. Marianne Defernez for introducing me to the wondrous world of correlation analysis. The old Biorefinery gang: Adam, Ian, Sam, Rares, Ricardo, Fatima, Jose, Jay et al. for the most random

conversations and for introducing me to life in Norwich outside of work. And what a life it turned out to be. At the NNUH enteric laboratory, Dr. Ngozi Elumogo and Claire Stuart for their help with sample collection. At PHE in Colindale, thank you to the staff in the GBRU, especially David Greig for answering all my stupid lab-related questions and Dr. Claire Jenkins for aliquotting out pots of poo and the tremendous support whilst I was writing my first paper. The GutBack study team for doing an amazing job in hardly losing any volunteers to follow-up.

Norwich Community Choir, in particular Kyanna, Dawn and Megan, for reminding me every week that there is a world outside of science. Ryan for FOOTBALL, beer, drams, ridiculous names, overall silliness and generally being the best antidepressant I could wish for. Martin/Urmel for keeping up our friendship over the distance. I will always make time for your stories.

Jonathan for noticing my smile at a time I thought I had lost it, for accepting me with all my flaws and for dealing with my moods, especially over the last few months. We can have a life again soon! And remember, I am only doing this so that we can one day afford a dog.

Mama und Papa, ich habe vielleicht nicht unbedingt den Weg eingeschlagen, den ihr euch für mich vorgestellt hattet. Vielleicht bin ich etwas zu selbständig geworden und mit Sicherheit oftmals viel zu stur. Aber ich kann mir das nur erlauben, weil ich weiß, dass ich zwei Menschen hinter mir stehen habe, die im Notfall Himmel und Hölle in Bewegung setzen würden, um mir beizustehen. Nicht ein einziges meiner Abenteuer wäre ohne euch möglich gewesen. Danke für alles, ich liebe euch!

CHAPTER ONE

1. General introduction

1.1 The human gut microbiota

Human beings, like other complex organisms, exist in a symbiotic relationship with a multitude of microorganisms, collectively referred to as the microbiota. While these microbes include archaea, fungi, viruses and bacteria, the latter are thought to be the most abundant organisms[1] and many studies describing microbial communities are therefore focussed on bacteria. Initial estimates from the 1970s suggested that bacterial cells vastly outnumbered those making up the human body, by a ratio of 10:1[2,3]. More recent estimates suggest that their actual number is more likely to be equal to the number of human cells, making up approximately 0.2 kg of the body mass of an average 70 kg male[4,5]. Nonetheless, even at these revised lower numbers, microbial activity can be expected to have profound effects on the host. The collection of genes encoding this activity is termed the microbiome[6].

1.1.1 Composition of the human gut microbiota

Microbial communities of distinct composition have been described for various body sites, including the skin[7], oral cavity[8] and vaginal tract[9]. The largest number of microorganisms by far, however, resides in the gastrointestinal (GI) tract[3,5]. In general, descriptions of the gut microbiota refer to faecal, in other words colonic, content as the number of microbes in the stomach and small intestine is negligible in comparison, owing to low pH and short transit times[5].

An individual's colonic microbiota is established early in life. For a long time the scientific consensus was that, during development in the womb, foetuses remained sterile and that the first contact with microorganisms occurred at birth through exposure to the mother's vaginal flora and the environment. However, discovery of a microbiota in meconium[10,11], the first stool of a neonate, challenged this conception and suggests that humans might be exposed to microbes earlier than assumed. Further colonisation occurs rapidly after birth and generally follows a common pattern of succession: Facultative anaerobes like *Enterobacteria* and *Enterococci*, the first colonisers, pave the way for strict anaerobes such as *Bacteroides*, *Clostridia* and *Bifidobacteria* by depleting the lumen of oxygen[12]. Over time, microbiota diversity

increases, and by three to five years of age its composition resembles that of an adult's. At this stage, gut microbiota profiles are individually distinct, with the exact composition varying based on environmental exposures[13]. The early microbiota of babies delivered vaginally is dominated by *Lactobacilli*, derived from the maternal vaginal tract[14,15]. In children born via caesarean section, on the other hand, skin microbes are more prominent, bacterial diversity is decreased and colonisation with Bacteroidetes is delayed[16,17]. Preterm birth also leads to reduced diversity and additionally seems to favour colonisation with potentially pathogenic bacteria over *Bifidobacteria* and *Bacteroides*[18,19]. Infant diet is another factor influencing the taxonomic composition of the developing microbiota: Breast feeding leads to exposure to bacteria in the mother's milk, mainly *Streptococci* and *Staphylococci*[20,21], and complex oligosaccharides contained within the milk stimulate the growth of, for example, *Bifidobacteria*[22]. Introduction of solid foods during weaning plays an important role in further increasing gut microbiota diversity and leads to colonisation with butyrate producers such as *Clostridia* [23].

Once established within the first years of life, an individual's microbiota becomes less susceptible to modulation by environmental influences and, in the absence of serious insults, is relatively stable over time[24,25]. The typical adult gut microbiota is dominated by two phyla, Firmicutes and Bacteroidetes, and Actinobacteria, Proteobacteria, Fusobacteria and Verrucomicrobia make up most of the remaining microbes[9,26]. Different ratios of these common phyla, and their members at higher taxonomic resolution, give rise to individually distinct microbiota profiles. Apart from early-life exposures, the exact nature of the profiles is affected by host genetics[27-29], lifestyle and geographical location[30-33]. Although studied less extensively, the gut microbiota seems to become less stable again in later life, with decreased diversity and increased levels of facultative anaerobes reported in elderly cohorts[34-36].

In an attempt to categorise individuals based on the dominant taxa making up their gut microbiota, the concept of enterotypes was introduced[37]: Enterotype 1 is characterised by the abundance of *Bacteroides*, type 2 by *Prevotella* and type 3 by *Ruminococcus*. On top of compositional information, these enterotypes also define major functions exerted by members of the gut microbiota. Nowadays, enterotypes

are viewed as continuous rather than discrete entities and they are thought to fluctuate depending on an individual's diet[38].

1.1.2 Function of the human gut microbiota

The strong effect of diet on the composition of the gut microbiota is linked to one of the major benefits the host derives from the symbiotic relationship with microbes. Bacteria have the ability to ferment complex dietary residues which humans cannot process as they lack the necessary digestive enzymes. In doing so, the gut microbiota produces metabolites essential to host health, including short chain fatty acids (SCFAs), vitamins and secondary bile acids. The SCFA butyrate, primarily produced by Firmicutes, has anti-inflammatory[39,40] and anti-tumorigenic effects[41,42], is the main energy source for colonic epithelial cells[43,44], regulates their proliferation[45] and enhances gut barrier function by facilitating assembly of tight junctions[46]. Furthermore, SCFAs are involved in lipid and glucose homeostasis in the liver[47,48] and might play a role in appetite regulation[49]. Vitamin B_{12} and folate, important for haematopoiesis, are essential vitamins synthesised by lactic acid and *Bifidobacteria*[50,51]. The gut microbiota also contributes to bile acid metabolism by converting unabsorbed primary to secondary bile acids like deoxycholic and lithocholic acid[52], which act as signalling molecules regulating lipid[53], energy[54] and glucose homeostasis[55].

An individual's microbiota adapts to most efficiently process the dietary residues it is exposed to regularly. This explains why lifestyle choices or geographical location influence its composition. The GI tract of vegetarians and vegans, for example, was found to harbour high levels of *Prevotella* and *Lachnospira*, bacteria which degrade plant-based dietary fibre[31,56]. Similarly, Schnorr et al.[32] compared gut microbiota composition in an Italian cohort following a Mediterranean diet with that of a traditional hunter-gatherer community in Tanzania. An enrichment of specialised *Prevotella* and *Treponema* in the latter allows them to extract the maximum amount of nutrients from the foods they have access to.

The second major function of the gut microbiota is to promote host immune defence. Commensal microorganisms, the normal, largely beneficial inhabitants of the GI tract, have developed strategies to directly stop establishment of foreign,

potentially pathogenic microbes within the gut and to prevent the overgrowth of fellow commensals, a phenomenon referred to as colonisation resistance[57]. This can be achieved through a variety of mechanisms, one of which is the competition for nutrients. Many of the major commensal taxa, such as *Bacteroides*, have the ability to process a wide range of different dietary residues while pathogens may be more limited in the number of compounds they can use for energy generation[58,59]. In a balanced microbiota, most nutrients are scavenged by the commensals leaving little behind for utilisation by other organisms. Prebiotics are dietary nutrients that particularly favour the growth of beneficial taxa and their consumption by the host can enhance colonisation resistance, although most of the evidence supporting this theory has come from murine models[60,61]. The beneficial taxa themselves are referred to as probiotics and nutritional supplements enriched with such microorganisms are available. On top of having a competitive advantage, many commensals also secrete bacteriostatic or bactericidal compounds, such as bacteriocins, that directly inhibit or kill other microbes. Their ability to control the growth of pathogens has made them the target of many studies searching for alternatives to conventional antibiotics[62]. Additionally, contact-dependent inhibition mechanisms mediated by secretion systems have been described[63,64]. By-products of bacterial metabolism also contribute to colonisation resistance: SCFAs and secondary bile acids were shown to inhibit the growth of enteropathogens, such as *Salmonella* Typhimurium[65] and *Clostridioides difficile*[66], in mice.

Moreover, the gut microbiota plays an important part in development of the host's immune system by priming immune cells in early life. The T-cells of newborns preferentially develop tolerance in response to antigen exposure[67]. The mucosal immune system therefore learns to tolerate the early colonisers of the infant gut, preventing an autoimmune response to commensals[68]. During the weaning period in mice, high levels of pro-inflammatory cytokines have been observed in the gut in response to exposure to new microorganisms through solid food[69].

1.1.3 The gut microbiota in dysbiosis

In the healthy state, the mechanism of colonisation resistance makes sure that the constituents of the gut microbiota occupy their individual niche and in conjunction

perform their beneficial functions. Disturbances that disproportionately affect certain taxa can therefore imbalance the entire finely tuned machinery, inducing a dysbiotic state. The nature of these disturbances can be varied with transient or long-term effects on the host's health. A common cause of gut microbiota dysbiosis is antibiotic treatment. The majority of antibiotics are not specific and can therefore kill other species apart from the intended ones. Accordingly, antibiotic treatment is usually accompanied by a decrease in bacterial diversity and changes in taxon abundance[70]. Most groups are able to recover within a month after treatment but some changes appear to persist up to two years after treatment[71]. During the antibiotic-induced dysbiotic state, colonisation resistance is reduced increasing the likelihood of successful pathogen invasion. This forms the basis for development of a recurrent *C. difficile* infection (CDI)[72]. Faecal microbiota transplants are more successful at treating CDI than antibiotics as the former aim at restoring the protective function executed by gut microbiota commensals instead of utilising substances that could cause further damage[73]. The effects of antibiotic use are particularly detrimental in early life as even short-term antibiotic treatment can disturb the evolution of the infant gut microbiota, delaying the colonisation with *Bifidobacteria* and increasing the abundance of Proteobacteria[74]. Since this can also negatively affect the development of the immune system, antibiotic treatment of neonates can increase the risk of allergies in later life[75].

Obesity was found to change the healthy composition of the microbiota. The microbial community of obese individuals is characterised by a reduction in Bacteroidetes and higher levels of Actinobacteria, which results in an increased energy harvest capacity[76]. Obesity also causes chronic low-level inflammation that might be modulated by members of the microbiota[77] and plays a part in the pathogenesis of non-alcoholic fatty liver disease via microbial metabolites[78]. Abnormal microbiota compositions were observed in patients suffering from disorders of the GI tract like Crohn's disease[79], ulcerative colitis and irritable bowel syndrome[80]. While it is easy to see why changes in the microbial ecosystem of the gut might contribute to the progression of GI diseases, alterations in microbiota composition have also been linked to other conditions less obviously related to the GI tract, such as depression[81], anxiety[82] and autism spectrum disorders[83]. These links are thought

to be based on the gut-brain axis, signalling between the GI tract and the central nervous system[84].

1.2 Acute gastroenteritis

Failure of the gut microbiota's protective mechanism against changes to the ecosystem can allow colonisation with or overgrowth of enteropathogens and lead to acute gastroenteritis. Enteropathogens cause symptoms by either secreting toxins, which tends to result in milder disease characterised by vomiting and/or diarrhoea, or by invasion of the intestinal mucosa[85]. The latter results in an inflammatory condition referred to as dysentery that is associated with bloody diarrhoea, abdominal cramps and fever[86].

Worldwide, there are an estimated 1.7 billion cases of gastroenteritis annually, resulting in 1.5 million deaths[87]. Due to inadequate sanitation and hygiene, the highest burden of disease is carried by developing countries, and here especially by children. Children under five years in these regions experience three to four diarrhoeal episodes per year[88] and often suffer from long-term health consequences such as developmental defects[89]. Individuals in developed regions, in comparison, only experience an estimated 0.3 episodes of gastroenteritis annually, usually do not require medical attention and infections are rarely fatal[90]. The economic and societal impact of the disease can therefore easily be overlooked. Nonetheless, every year an estimated quarter of the UK population will suffer from gastroenteritis, leading to approximately 1 million general practice consultations[90] and an average absence from work or school for six days per case[91]. Prolonged or recurrent disease increases the risk of long-term health complications and hospitalisation, putting further financial strain on the affected individuals and the healthcare sector.

1.2.1 Aetiological agents of acute gastroenteritis

A variety of pathogens of bacterial, viral, parasitic and fungal origin have been implicated as aetiological agents of GI infections. Rotavirus and norovirus are the most common viral causes of gastroenteritis, the former primarily affecting children[92] and the latter often associated with outbreaks in healthcare settings[93]. In the UK,

bacteria are involved in 20-40% of diagnoses and therefore of equal importance to viruses[90]. With an estimated 500,000 cases annually and 80,000 GP consultations, *Campylobacter* is the most common cause of bacterial gastroenteritis[90]. A potential complication resulting from campylobacteriosis is Guillain-Barré syndrome[94], an autoimmune condition affecting the peripheral nervous system. Enteroaggregative and Shiga toxin-producing *Escherichia coli* (STEC) and non-typhoidal *S. enterica* (NTS) follow *Campylobacter* as the most frequently encountered aetiological agents of bacterial gastroenteritis[90]. Less common pathogens, such as *Vibrio* species[95], are often associated with travel to countries where they are endemic. Transmission of bacteria usually occurs through interaction with contaminated surfaces[96], close contact to animals, which may be asymptomatic carriers of pathogens[97,98], or ingestion of contaminated water and food[99,100]. *Salmonellae* traditionally were a major cause of foodborne gastroenteritis but outbreaks linked to this pathogen have declined[101], likely thanks to successful public health interventions targeting the poultry industry[102]. Foodborne outbreaks due to *Campylobacter* and STEC, on the other hand, seem to be on the rise[101].

Similar to commensal bacteria, many protozoal eukaryotes are constituents of the healthy gut[103,104], even though their exact role in the ecosystem is poorly understood[105]. Nonetheless, some protozoa are enteropathogens, with *Giardia lamblia*, *Cryptosporidium*, *Entamoeba* and *Cyclospora cayetanensis* being the most common ones in developed countries[106]. Other protozoa, such as *Blastocystis*[103,107] or *Dientamoeba fragilis*[104,108], are generally associated with asymptomatic carriage but can become pathogenic under certain conditions. Although foodborne outbreaks have also been reported[109,110], protozoal infection is mainly transmitted through contaminated water[111,112] and therefore incidence in developing countries with poor sanitation is higher. *Giardia* prevalence in human populations, for example, is estimated to be 4-43% in developing and 1-7% in developed countries[113,114]. Further risk factors for protozoal gastroenteritis seem to be contact with livestock[115,116] as well as gender since several studies have observed an increased incidence of infection in men[117,118].

1.2.2 Antimicrobial resistance in enteropathogens

While the availability of antimicrobials undoubtedly plays a part in keeping fatality rates from GI infections low, the overuse of broad-spectrum antibiotics is one of the underlying factors of the global threat posed by antimicrobial-resistant organisms. When Alexander Fleming received his Nobel Prize for the discovery of penicillin in 1945, he warned that misuse of the drug could lead to selection for resistant bacteria. Within ten years, penicillin-resistant strains were reported[119]. Antimicrobial resistance (AMR) occurs when microbes acquire certain traits by spontaneous mutation or incorporation of foreign DNA into their own genome, allowing them to survive exposure to compounds that would normally kill them. Since these resistant organisms have a selective advantage over their susceptible counterparts they quickly outgrow the normal population[120]. Resistance strategies may include decreased cell wall penetration, export of drugs out of the cell or enzymatic inactivation. How quickly bacteria respond to environmental pressures imposed by antibiotics through accumulation of mutations, was demonstrated by the Baym laboratory's MEGA-plate experiment, which showed that *E. coli* evolve to tolerate very high concentrations of antimicrobials within eleven days[121].

Annually, more than 700,000 deaths are attributed to resistant microorganisms already and this number is expected to rise to 10 million deaths per year by 2050, overtaking cancer as the main cause of mortality[120] . While not all of these deaths will be linked to resistant enteropathogens, the World Health Organisation has included extended spectrum β-lactamase (ESBL)-producing, carbapenem-resistant Enterobacteriaceae and fluoroquinolone-resistant *Salmonellae, Campylobacter* and *Shigella* species on their list of priority pathogens for the development of new antibiotics due to a worrying increase in treatment failures[122]. Enrichment of AMR genes in the gut microbiota can occur via ingestion of contaminated food[123] or water[124] or by close contact with environments harbouring resistant bacteria[125]. Alternatively, chronic exposure to low levels of antibiotic residues in food and water can select for resistant bacteria[125,126]. Even animals exposed to environments enriched in antibiotic residues or resistant bacteria can act as reservoirs for the transfer of resistance genes to humans. Rats living in hospital sewage systems, for

example, were found to carry pathogens resistant to vancomycin and an increased number of ESBL-producing *E. coli* [127,128].

1.3 '-omics' technologies

1.3.1 Next-generation sequencing

Many of the insights into the composition and function of the gut microbiota would not have been possible without the development of advanced sequencing technologies. Initial studies relied on culturing and Sanger sequencing of the 16S rRNA gene of the isolated bacteria or PCR-Denaturing Gradient Gel Electrophoresis. These approaches were biased through the selection of media and limited in the number of organisms that could be detected, thus only capturing a fraction of the diversity of the gut microbiota[26]. Next-generation sequencing (NGS), on the other hand, allows massively parallel sequencing of millions of DNA fragments derived from isolated microorganisms or complex communities. The available NGS platforms exploit different sequencing technologies and the most widely used ones are manufactured by Illumina® and Oxford Nanopore Technologies. Illumina® platforms employ a sequencing-by-synthesis approach after attachment of DNA fragments to a flow cell and bridge amplification[129]. Oxford Nanopore sequencers guide single-stranded DNA molecules through a grid of protein nanopores, with readouts based on disruption of an electrical current[130]. The technologies can be used for metataxonomic, metagenomic and whole genome sequencing (WGS) studies of individual isolates.

Metataxonomics refers to amplicon-based profiling of microbial communities with the help of conserved marker genes to determine taxonomic composition. Investigations of bacterial communities are carried out using the 16S rRNA marker gene, which encodes part of the prokaryotic 30S ribosomal subunit. The gene is approximately 1.5 kb in length and contains nine hypervariable regions, which species identification is based on, flanked by highly conserved regions ideal for primer binding[131]. Metagenomics, in comparison, is the untargeted study of all the genetic material of a complex community after random fragmentation of larger DNA sequences. On top of taxonomic information, the resulting data can be used to infer functional characteristics of the community through annotation of gene fragments.

1.3.2 Metabolomics

Metabolomics is the study of all the metabolites present in a biological matrix[132] that can be used for functional characterisation of microbial communities. Technologies available for metabolomic investigations include high-performance liquid chromatography, mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy. Although less sensitive than MS, NMR spectroscopy allows higher sample throughput and is therefore more suitable for large-scale studies[133]. NMR is based on absorption of electromagnetic radiation by stable isotopes, such as ^{1}H or ^{13}C, at a frequency characteristic of the isotope. The exact position of the isotope relative to other atoms further modifies its absorption characteristics so that unique spectra are obtained for different compounds.

1.3.3 Multivariate statistics

Classical statistical analyses focussing on easily identifiable changes in a single variable are rarely suitable for the large datasets produced by the high-throughput technologies described in the previous sections. To investigate trends in the data, multivariate statistics are employed instead, which incorporate measurements for multiple variables, for example the relative abundance of all the individual bacterial species present in a single stool sample. Many multivariate statistics tools make use of ordination, which aims to determine gradients of variation between samples based on the values measured for the investigated variables. Ordination generally involves dimensionality reduction so that samples can be represented in a system of coordinates of few dimensions with a larger distance between sample coordinates indicating higher dissimilarity. Ordination techniques can be unconstrained, reflecting overall data variance, or constrained. The latter approach uses a set of explanatory variables within the dataset and aims to explain the variation in the variables of interest by the variation in the explanatory variables.

One of the most widely used ordination methods is principal component analysis (PCA), where each principal component (PC) is a linear combination of the original variables constituting an axis in the multidimensional data space with the first PC representing the largest gradient of variation in the data[437]. Correspondence analysis

(CA) similarly displays data in a low-dimensional space but while PCA is employed for continuous data, CA is used for categorical data[438]. In PCA and CA, samples are organised based on (weighted) Euclidean distances to one another. Principal coordinates analysis (PCoA), on the other hand, can be applied to any distance matrix, including those incorporating phylogenetic distance measures.

1.4 Thesis aims and objectives

This thesis seeks to prove the advantages of '–omics' technologies over conventional techniques in the context of diagnosis and surveillance of GI infections and to provide guidance on future implementation of these technologies as routine services in diagnostic and reference laboratories. The following approaches will be taken to achieve this aim:

1. Use of 16S rRNA gene sequencing of the gut microbiota of patients with symptoms of acute gastroenteritis for culture-independent detection of aetiological agents of disease.
2. Identification of pathogen-specific taxonomic and metabolic biomarkers of GI disease with the help of 16S rRNA gene sequencing and NMR spectroscopy.
3. Detection and characterisation of *S. enterica* strains in stool samples derived from patients diagnosed with salmonellosis using metagenomic sequencing.
4. Evaluation of the sensitivity and specificity of WGS-derived AMR profile predictions compared to phenotypic antimicrobial susceptibility testing for NTS isolates.
5. Examination of the prevalence and persistence of changes in gut microbiota composition and resistome induced by exposure to a foreign environment by conducting a longitudinal metagenomic study involving international travellers.

CHAPTER TWO

---

2. Taxonomic and metabolic markers of bacterial and protozoal gastroenteritis

2.1 Summary

Due to the limitations of current diagnostic methods for acute gastroenteritis, identification of an aetiological agent is only possible in approximately half of the cases referred to laboratories. Increasing the proportion of successful diagnoses would allow targeted treatment and provide more surveillance data for enteric pathogens. To identify pathogen-specific changes in gut microbiota taxonomic and metabolic profiles during infection that could aid successful diagnosis when other methods fail, 16S rRNA gene sequencing and ^{1}H nuclear magnetic resonance (NMR) spectroscopy were performed on stool samples from 41 healthy controls and 246 patients diagnosed with bacterial or protozoal gastroenteritis. Infection decreased bacterial diversity and the abundance of Firmicutes. A decrease in *Blautia* abundance and in the levels of caprate and caprylate were characteristic of bacterial, an increase in butyrate, hypoxanthine, isovalerate and phenylacetate of protozoal infection. Random Forest classification of samples based on both taxonomic and metabolic features resulted in prediction of the type of infection with an accuracy of 81.61%, and prediction of the aetiological agent with an accuracy of 69%.

2.2 Introduction

2.2.1 Conventional approaches to diagnosis of gastrointestinal infections

Since gastroenteritis is generally self-limiting[134] and unlikely to cause long-term health consequences for otherwise healthy individuals, or even prompt patients to access healthcare, a large number of cases remain undiagnosed. Establishment of an aetiological agent of disease is, however, necessary for patients with severe or prolonged symptoms, an indication of invasive disease and a history of complications associated with gastrointestinal (GI) disease[134]. Mere assessment of symptoms is of limited use as clinical presentation is highly similar for many enteric pathogens[135] so that laboratory testing is required for precise diagnoses. To this end, faeces (and to a lesser extent rectal swabs) are collected during the acute phase of diarrhoea and ideally processed within two hours[136].

Initial examination includes visual inspection of the sample for blood and mucous as well as description of stool colour and consistency according to the Bristol Stool Chart (BSC)[137]. The UK Standards for Microbiology Investigations[86] recommend routine screening of stool samples for *Campylobacter*, *Salmonella*, *Shigella* and Shiga toxin-producing *Escherichia coli* (STEC). Inclusion of further pathogens in the testing panel depends on the season, the patient's age and travel history, and whether the case occurred in a community or healthcare setting. Cases linked to recent travel to Asia, Africa and Latin America, for example, should additionally be tested for *Vibrio* and *Plesiomonas* species and several cases from the same hospital or nursing home might be indicative of a nosocomial outbreak, often caused by Norovirus or *Clostridioides difficile*.

In the majority of frontline laboratories, stool culture on selective media remains the routine diagnostic method for identification of bacterial pathogens. An example is the isolation of *Salmonella* species by initial inoculation on mannitol selenite broth and subsequent subculture on xylose lysine desoxycholate agar[138]. For higher taxonomic resolution, culturing can be followed up with biochemical tests, such as the mannitol fermentation test for differentiation of *Shigella* species[139]. Over the last

decade, some laboratories have started using matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) mass spectroscopy (MS) for follow-up testing instead, which identifies species based on the size of characteristic peptides[140]. Diagnosis of protozoal infections is traditionally based on the detection of trophozoites or oocysts using a microscope, sometimes with the help of stains such as the fluorescent Auramine O employed for *Cryptosporidium* oocysts[141].

As an alternative to these classical methods, antigenic tests have been developed for many pathogens, including viruses, and are the recommended approach for detection of *C. difficile* toxins[142]. Increasingly, laboratories use nucleic acid amplification tests (NAATs), either for a single pathogen and its virulence factors or combining several reactions into a multiplex panel. These panels range in the number and type of their targets - some being specific for bacteria[143,144], viruses[145] or parasites[146,147], and others covering a wider selection of pathogens[148] – and also in the type of analysis platform required. The BioFire® Film Array GI panel[149], for example, detects 22 pathogens but requires specialised equipment, whereas the EntericBio Gastro Panel 2[150] only targets six organisms but can be used with a standard real-time polymerase chain reaction (PCR) instrument.

2.2.2 Challenges of gastrointestinal infection diagnostics

Improving the outcome of certain enteric infections depends heavily on timely, adequate interventions. Renal problems in children suffering from haemolytic uraemic syndrome after STEC gastroenteritis, for example, were found to be less severe when administration of intravenous fluids occurred early during the infection[151]. The high turnaround time of stool culture – a minimum of 48 hours – is prohibitive to fast, targeted interventions. Successful recovery of viable organisms depends heavily on bacterial load, time of sample collection after the onset of symptoms and the time that has passed between sample collection and processing[152], and the approach is not cost-effective[153]. Furthermore, some pathogens, such as the anaerobe *C. difficile*, do not grow under standard conditions. Thus, culturing is thought to underestimate even the burden of common enteropathogens like *Campylobacter* and *Shigella*[154,155]. Some of these challenges can be overcome by the use of MALDI-TOF MS: The method is more cost-effective

and decreases turnaround times as it allows identification of many organisms after a 24-hour inoculation period. However, it does not eliminate the initial culturing step and many platforms cannot distinguish *Shigella* from *E. coli* and are unable to type *Salmonella* at the serovar level[140,156].

Microscopic identification of protozoa requires experienced staff but since some organisms are encountered very infrequently, training of additional staff can only be undertaken sporadically. Oocysts are often small and thus easily missed, especially for organisms that do not stain well such as *Cyclospora cayetanensis*[157]. Shedding of protozoa in faeces is irregular so to increase the likelihood of detection collection of three specimens over several days is generally required[157]. Due to these issues, the sensitivity of microscopy-based diagnosis of protozoal infections can be as low as 20%[157]. On top of this, some pathogenic species are morphologically indistinguishable from non-pathogenic relatives, as is the case for *Entamoeba histolytica* and *dispar*[146].

While antigenic tests can be more cost-effective and decrease turnaround times, their sensitivity and specificity varies[158,159] so that initial results frequently need to be followed up with a confirmatory test. Implementation of NAATs, although at first associated with higher expenses due to procurement of additional equipment, has been shown to lead to cost savings for healthcare facilities in the long term[160]. When designed to amplify the appropriate pathogen-specific target sequences, sensitivity and specificity of NAATs is generally high for all types of pathogens. However, due to their ability to detect pathogen nucleic acids at very low levels, PCR-based methods do not necessarily prove the presence of viable organisms. The clinical importance of positive results can therefore sometimes be questionable.

2.2.3 Potential use of '–omics' technologies for diagnosis of gastrointestinal infections

While multiplex PCR panels simultaneously target the most commonly encountered enteropathogens, even the most elaborate ones will only detect a restricted number of organisms. However, for bacteria alone, more than 40 species have been associated with GI infections[161]. Adding this fact to the aforementioned limitations of enteropathogen diagnostics, it is unsurprising that an aetiological agent is

successfully identified in only 40-50% of gastroenteritis cases referred for laboratory diagnostics[162]. A diagnostic approach simultaneously detecting all microorganisms in the gut would circumvent the problem of missing aetiological agents by simply not targeting them.

Acute GI infections have been shown to introduce changes in the commensal gut microbiota based on the pathogens outcompeting established bacteria and disrupting the network of interactions between organisms. This effect has been described for bacterial[163-165], protozoal[166] and viral pathogens[167,168]. Since more than 10% of metabolites found in faeces are estimated to be of microbial origin or at least microbially modified[136], changes in gut microbiota community composition are expected to also have an effect on the metabolome. Given that a pathogen's specific virulence factors will influence the exact manifestation of disease[169], different pathogens might induce characteristic changes in the established gut microbiota. In cases where direct identification of an aetiological agent is not possible, pathogen-specific taxonomic and metabolic biomarkers of infection could therefore provide an indication of the organism responsible for the patient's symptoms, increasing the number of successful diagnoses and allowing more targeted treatment.

2.3 Objectives

The purpose of the work presented in this chapter was to assess the potential of 16S rRNA gene sequencing for identification of bacterial enteropathogens and to identify pathogen-specific taxonomic and metabolic biomarkers of bacterial and protozoal disease using metataxonomics and metabolomics. It was hypothesised that characteristic changes in gut microbiota composition and function could be used to predict the aetiological agent of acute gastroenteritis.

2.4 Materials and Methods

Routine diagnostics were carried out by staff at the Norfolk and Norwich University Hospital's (NNUH) enteric laboratory. Stool samples were collected by Marieke Pape, Dr. Lee Kellingray, Henry Whiley and the author. Faecal waters and DNA extracts were prepared by Marieke Pape and the author. NMR spectra were acquired and transformed by Dr. Gwenaelle Le Gall and Dr. Ian Colquhoun. Unless otherwise stated, default software parameters were used.

2.4.1 Processing of faecal samples

2.4.1.1 Sample collection

Faecal samples from patients with symptoms of GI infections were referred to the NNUH enteric laboratory from local hospitals and general practitioners. For samples received up until November 2017, stool culture on selective media was performed for pathogen detection. From November 2017 onwards, testing was carried out by PCR using the EntericBio Gastro Panel 2 (Serosep Limited, Limerick, Ireland), which detects *C. jejuni/coli/lari*, *C. parvum/hominis*, *Giardia lamblia*, *S. enterica*, *Shigella* species and STEC[150]. For the latter three pathogens, a positive PCR result was confirmed by stool culture. The faecal material remaining after testing was stored at 4°C in the enteric laboratory. Aliquots of pathogen-positive samples were transferred to Quadram Institute Bioscience (QIB) and stored at -20°C until further processing. Sample collection was undertaken between December 2015 and December 2017. The study was approved by the University of East Anglia's Faculty of Medicine and Health Sciences Research Ethics Committee (Reference: 20152016 31 HT). Ethics application form and approval letter can be found in **Appendix I**.

Pre-intervention faecal samples from participants in the BERI (The Effects of Bilberry Fruit and Black Rice Derived Anthocyanins on Lipid Status in Adults) (NCT number: NCT03213288) and EBL (Effects of Brassica on Human Gut Lactobacilli) (NCT number: NCT02291328; described in Kellingray et al.[170]) studies as well as samples obtained from volunteers on the QIB Colon Model study (NCT number: NCT02653001) were used as healthy controls.

2.4.1.2 Faecal water preparation

Samples were kept on ice whenever possible. Approximately 200 mg of faecal material was homogenised in 12x (w/v) NMR buffer using a Kimble® Kontes cordless pellet pestle motor with reusable pestles (Merck, Darmstadt, Germany). The buffer composition is shown below:

- 0.26 g $NaH_2PO_4 \cdot H_2O$ (BDH AnalaR NORMAPUR®, Safat, Kuwait)
- 1.44 g K_2HPO_4 (Sigma-Aldrich, St. Louis, MO, USA)
- 56.1 mg NaN_3 (Sigma-Aldrich)
- 17 mg sodium 3-(trimethylsilyl)propionate (TSP) (CK Isotopes Ltd., Desford, UK)
- 100 ml D_2O (CK Isotopes Ltd.)

For negative controls, 1 ml of buffer was processed without faecal matter. Samples were centrifuged at 3,220 x g and 4°C for 15 min and the supernatants sterile-filtered through 0.2 µm Minisart® syringe filters (Sartorius, Göttingen, Germany). Filtered faecal waters were stored at -20°C until acquisition of NMR spectra. Pellets were stored at -20°C until further processing.

2.4.1.3 Faecal DNA extraction

Total genomic DNA was extracted from the pellets obtained after faecal water preparation by mechanical lysis using the FastDNA™ SPIN Kit for Soil (MP Biomedicals, Santa Ana, CA, USA). Samples were kept on ice until the first centrifugation step. After addition of 978 µl sodium phosphate buffer and 122 µl MT buffer, samples were left to stand at 4°C for 1 h with a vortexing step every 15 min. Samples were transferred into Lysing Matrix E tubes and bead beating was carried out with a FastPrep-24™ tissue homogeniser (MP Biomedicals) using three 60 s pulses at 6.5 m/s with 2-min intervals between pulses. Lysates were centrifuged at 16,800 x g for 1 min and 250 µl of protein precipitation solution were added to the supernatant. Samples were mixed by inversion and centrifuged at 16,800 x g for 5 min. Supernatants were mixed with 1 ml binding matrix solution by inversion for 2 min

and left to stand for 3 min. After removal of approximately 1 ml of supernatant, the matrix was resuspended in the remaining supernatant, transferred to SPIN filter tubes and centrifuged at 14,500 x g for 1 min. The binding matrix was washed with 500 µl of salt/ethanol solution and centrifuged at 14,500 x g for 1 min. The washing and centrifugation steps were repeated twice more followed by an additional 2-min centrifugation step at 16,800 x g. The flow-through was discarded and the binding matrix left to air-dry for 5 min. The matrix was resuspended in 50 µl of DNA elution solution-ultrapure water, left to stand for 3 min and centrifuged at 16,800 x g for 1 min. Eluates were stored at -20°C until sequencing. DNA from the EBL study samples had been pre-extracted by Lee Kellingray and stored at -40°C.

2.4.2 Metataxonomics

2.4.2.1 Preparation of 16S rRNA gene libraries and amplicon sequencing

Extracted DNA was quantified using a Qubit fluorometer with broad-range reagents (Thermo Fisher Scientific, Waltham, MA, USA) prior to being sent to Earlham Institute (formerly The Genome Analysis Centre, Norwich, UK), the Centre for Genomic Research (University of Liverpool, Liverpool, UK) or Novogene (Beijing, China) for further processing in nine separate batches. The V4 hypervariable region of the 16S rRNA gene was amplified with 515F (GTGCCAGCMGCCGCGGTAA) and 806R (GGACTACHVGGGTWTCTAAT) primers[171] using the Phusion® High-Fidelity PCR Master Mix (New England Biolabs, Ipswich, MA, USA). Quality control of amplicons was carried out on a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) and only amplicons 400-450 bp in size were included in the sequencing reaction. Sequencing libraries were generated with the NEBNext® Ultra™ DNA Library Prep Kit for Illumina® (New England Biolabs). Libraries were purified using the AMPure XP system (Beckman Coulter, Brea, CA, USA), analysed for size distribution using the Bioanalyzer instrument and quantified by real-time PCR. Cluster generation was performed on a cBot System (Illumina®, San Diego, CA, USA). Paired-end 150 bp sequencing was performed on a HiSeq 2500 instrument (Illumina®).

2.4.2.2 Trimming and quality filtering of sequencing reads

Quality trimming and filtering of sequencing reads was carried out by the sequencing providers. Reads were demultiplexed and adapters as well as index sequences were removed. Reads containing >10% of ambiguous base calls as well as those for which >50% of total base calls had a quality score ≤5 were filtered out. Only the remaining high-quality reads were included in subsequent analyses.

2.4.2.3 Sequence analysis

The following analyses were performed using the Medical Research Council's Cloud Infrastructure for Microbial Bioinformatics[34]. Taxonomic analysis of demultiplexed sequencing reads was carried out with QIIME 2 (Quantitative Insights Into Microbial Ecology) v. 2018.8 and software versions implemented within the environment[172,173] as described in the "Moving Pictures" tutorial (https://docs.qiime2.org/2018.8/tutorials/moving-pictures/). Denoising, quality filtering, construction of a feature table containing the frequency of each unique sequence and mapping of identifiers to the sequences were performed with DADA2[174] for each sequencing batch separately. If median quality scores dropped below 30, reads were truncated at the base position this drop occurred for the first time.

Representative sequences obtained for the different batches were merged. Taxonomic labels were assigned to the sequences using a Naïve Bayes classifier pre-trained on the SILVA 132 99% operational taxonomic unit (OTU) database[175]. The feature table with taxonomic assignments was either converted to a tab-separated file directly using the 'biom convert' command or collapsed at the taxonomic level desired, converted to a relative abundance table and then to a tab-separated file.

Multiple sequence alignment of the representative sequences was performed using MAFFT (Multiple Alignment using Fast Fourier Transform)[176] and the alignment was filtered to remove highly variable positions. A phylogenetic tree was built with FastTree[177] and midpoint rooting was applied to the resulting tree.

Feature tables obtained for the different sequencing batches were merged. From the previously generated rooted phylogenetic tree and the merged feature tables with the sampling depth set to 12,150, alpha-diversity- a measure of microbial diversity

within a sample – was estimated in the form of Faith's phylogenetic diversity (PD)[178], Shannon's index and observed OTUs and β-diversity – a measure of microbial diversity between samples – was computed in the form of weighted UniFrac distances[179].

Predictions of functional metagenome content from the merged feature tables and the representative sequences was carried out using the QIIME 2 PICRUSt2 (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States) plugin v. 2019.10 (https://github.com/gavinmdouglas/q2-picrust2)[180,181]. Maximum parsimony was selected as hidden-state prediction method and the cut-off for exclusion of sequences based on distance from the reference phylogeny was set to 2. KEGG (Kyoto Encyclopaedia of Genes and Genomes) pathways were reconstructed from the resulting KEGG orthology metagenome predictions with the HUMAnN2 (HMP Unified Metabolic Analysis Network) pipeline[182] v. 2.8.0 and the KEGG pathway database from the HUMAnN v.0.99 software[183].

2.4.3 Metabolomics

2.4.3.1 ^{1}H nuclear magnetic resonance spectroscopy

Faecal waters were transferred to 5 mm borosilicate glass NMR tubes (Wilmad, Vineland, NJ, USA). ^{1}H NMR spectra were recorded on a 600 MHz AVANCE™ spectrometer (Bruker, Billerica, MA, USA) fitted with a cryoprobe and a 60-slot autosampler using the TopSpin® software, v. 3.2 (Bruker). Each spectrum was acquired with 1,024 scans, a spectral width of 12,295 Hz, an acquisition time of 2.67 s and a relaxation delay of 3 s. The "noesygppr1d" pre-saturation sequence was used to suppress the residual water signal with a low-power selective irradiation at the water frequency during the recycle delay and a mixing time of 10 ms.

2.4.3.2 Metabolite quantification

The recorded spectra were transformed with a 0.3 Hz line broadening, and manually phased, baseline-corrected and referenced by setting the TSP methyl signal to 0 ppm using the TopSpin® software. Metabolite identification and quantification based on reference compounds was carried out with the Chenomx NMR Suite software v.

8.2 (Chenomx, Edmonton, Canada). Reference libraries were provided with the software or had previously been custom-built at QIB.

2.4.4 Statistical analysis

2.4.4.1 Univariate analysis

Differences in α-diversity or metabolite concentrations were assessed using the dunnTest function with Benjamini-Hochberg (BH) correction in R's FSA (Fisheries Stock Assessment) package v. 0.8.25 (https://github.com/droglenc/FSA)[184]. An adjusted p-value ≤0.05 was considered statistically significant.

2.4.4.2 Multivariate analysis

Principal coordinate analysis (PCoA) of taxonomic sample composition was performed in the QIIME 2 environment on the weighted UniFrac distance matrix generated in section 2.4.2.2. Biplots were generated by projecting the information contained in the feature table onto the resulting PCoA matrix. Visualisation of data was carried out in R using the qiime2R package v. 0.99.12 (https://github.com/jbisanz/qiime2R). Differences in sample group dissimilarity were assessed using PERMANOVA (Permutational Multivariate Analysis of Variance)[185] in the QIIME2 environment.

Principal component analysis (PCA) of sample metabolite profiles was performed using the mixOmics package v. 6.8.4 in R[186]. The contribution of individual metabolites to variance was visualised using the package's plotLoadings function.

2.4.4.3 Normalisation of sequencing reads

Sequencing reads assigned to individual OTUs were aggregated at the desired level of the QIIME 2 taxonomy output using the metagenomeSeq package v. 1.26.3 in R[187]. Read counts were normalised with the help of the package's cumNorm function, which calculates scaling factors (SFs) equal to the sum of counts up to a specified quantile. The 0.5 quantile (i.e. median) was selected for this analysis. Read counts assigned to individual KEGG pathways were normalised in the same manner without prior aggregation.

2.4.4.4 Differential abundance testing

Differences in specific features at the desired level of taxonomy or in specific KEGG pathways between sample groups were estimated using the limma package v. 3.40.6 in R[188]. A design matrix was generated from the (aggregated) normalised read count data with the model.matrix function, adding storage time of the sample at 4°C, year of DNA extraction and sequencing centre as covariates. Furthermore, normalisation factors, derived from the SFs described in section 2.4.4.3, were added to the matrix. Normalisation factors were calculated as follows:

$$\log_2(SF/median(SF) + 1)$$

The lmFit function was used to produce weighted fits for all features with the previously generated design matrix as input. Pairwise comparisons between sample groups were carried out by constructing a contrast matrix and re-orientating the fitted model object from the coefficients of the original design matrix to any set of contrasts of the original coefficients. From the resulting linear model fits, fold change (FC) estimates and t-statistics with BH correction were computed using the eBayes function. An adjusted p-value ≤0.05 was considered statistically significant.

2.4.4.5 Correlation analysis

Centered log-ratio (clr) transformation of relative abundance data for genera making up >0.1% of the community in at least ten samples was carried out using the compositions package v. 1.40-2 in R[189]. Concentrations of metabolites present at >0.1 mM in at least ten samples were $\log_{10}$-transformed. A correlation matrix was generated using sparse partial least squares (sPLS) analysis in canonical mode, implemented in the mixOmics package. Modelling was performed with ten components and five metabolite and 20 taxonomic features on each component. A clustered image map was generated from the correlation matrix by determining the scalar products of sample coordinates on the sPLS axes.

2.4.5 Random Forest classification

Relative abundance data for genera making up >0.1% of the community and concentrations of metabolites present at >0.1 mM in at least ten samples as well as Faith's PD were used as features to classify samples based on the type of infection or the aetiological agent identified during routine diagnostics with the randomForest package v. 4.6-14 in R[190]. To ensure reproducibility, the seed of R's random number generator was set to 100. The training dataset was generated by randomly choosing 70% of samples and a Random Forest model was built with the number of trees set to 500 and the number of features sampled at each split set to 14 for the classification of the type of infection and to 16 for the classification of the aetiological agent identified. The importance of individual features in the model was assessed in the form of the mean decrease in prediction accuracy. Predictions of type of infection or aetiological agent were performed on the training set first, followed by predictions on the remaining 30% of samples comprising the validation dataset.

2.5 Results

2.5.1 Study samples

A total of 276 faecal samples from patients with symptoms of GI infections were collected during the study period. Of these, 159 tested positive for a bacterial and 117 for a protozoal pathogen. *Campylobacter* was identified in 111, *Cryptosporidium* in 24, *G. lamblia* in 90, *S. enterica* in 31, *Shigella* in 8 and STEC in 7 samples. A single patient suffered from a *Blastocystis hominis* infection and another from a *Campylobacter*/*Shigella* co-infection. For two samples, pathogen identification was ambiguous and stated as either *Cryptosporidium* or *G. lamblia*. Of the 43 faecal samples from healthy controls, nine were provided by QIB Colon Model study participants, ten from the EBL study and 24 from the BERI study.

Eight samples did not yield a sufficient amount of DNA for successful amplicon sequencing and a further five showed low sequencing read counts and thus fell below the sampling depth chosen for diversity analysis. To guarantee identical sample sets, these were also excluded from the metabolomics analysis. Furthermore, samples testing positive for a pathogen detected in less than ten samples overall, and those for which a single pathogen could not unambiguously be identified as the aetiological agent, were excluded from both datasets.

Table 2.1 shows the characteristics of the 287 samples included in the following analyses. Factors other than infection which might influence taxonomic and metabolite composition were also recorded. These comprised the storage time of the faeces at 4°C in the NNUH enteric laboratory before their transfer to -20°C at QIB, the time period of sample processing, the provider of the amplicon sequencing service and the stool's consistency. Stools defined as Type 1-4 according to the BSC[137] were categorised as solid, Types 5 and 6 as semi-liquid and Type 7 as loose.

Table 2.1: **Characteristics of study stool samples from controls and patients with gastrointestinal infections.**

	Control (n=41)	**Bacteria (total)** (n=134)	***Campylobacter*** (n=107)	***S. enterica*** (n=27)	**Protozoa (total)** (n=112)	***Cryptosporidium*** (n=24)	***G. lamblia*** (n=88)
Storage time at 4°C							
0 days	41 (100%)	2 (1.49%)	0	2 (7.41%)	8 (7.14%)	3 (12.50%)	5 (5.68%)
1 day	0	11 (8.21%)	10 (9.35%)	1 (3.70%)	34 (30.57%)	6 (25.00%)	28 (31.82%)
2 days	0	33 (24.63%)	28 (26.17%)	5 (18.52%)	29 (25.89%)	5 (20.83%)	24 (27.27%)
3 days	0	54 (40.3%)	44 (41.12%)	10 (37.04%)	31 (27.68%)	8 (33.33%)	23 (26.14%)
4 days	0	32 (23.88%)	24 (22.43%)	8 (29.63%)	8 (7.14%)	2 (8.33%)	6 (6.82%)
5 days	0	2 (1.49%)	1 (0.93%)	1 (3.70%)	2 (1.79%)	0	2 (2.27%)
Stool consistency							
solid	40 (97.56%)	14 (10.45%)	10 (9.35%)	4 (14.81%)	34 (30.57%)	8 (33.33%)	26 (29.55%)
semi-liquid	1 (2.44%)	64 (47.76%)	52 (48.60%)	12 (44.44%)	57 (50.89%)	9 (37.50%)	48 (54.55%)
loose	0	51 (38.06%)	40 (37.38%)	11 (40.74%)	19 (16.96%)	7 (29.17%)	12 (13.64%)
No information	0	5 (3.73%)	5 (4.67%)	0	2 (1.79%)	0	2 (2.27%)
Extraction year							
2015	8 (19.51%)	0	0	0	0	0	0
2016	0	104 (77.61%)	98 (91.59%)	6 (22.22%)	14 (12.50%)	2 (8.33%)	12 (13.64%)
2017	0	17 (12.69%)	9 (8.41%)	8 (29.63%)	35 (31.25%)	10 (41.67%)	25 (28.41%)
2018	33 (80.49%)	13 (9.70%)	0	13 (48.15%)	63 (56.25%)	12 (50.00%)	51 (57.95%)
Sequencing Centre							
CGR	0	16 (11.94%)	8 (7.48%)	8 (29.63%)	34 (30.57%)	9 (37.50%)	25 (28.41%)
EI	0	104 (77.61%)	98 (91.59%)	6 (22.22%)	14 (12.5%)	2 (8.33%)	12 (13.64%)
Novogene	41 (100%)	14 (10.45%)	1 (0.93%)	13 (48.15%)	64 (57.14%)	13 (54.17%)	51 (57.95%)

Patient samples are stratified by type of infection and aetiological agent identified during routine diagnostics in the enteric laboratory. Values denote the number/percentage of samples in each category showing the characteristic. CGR - Centre for Genomic Research, EI - Earlham Institute.

2.5.2 Taxonomic signatures of gastrointestinal infections

2.5.2.1 Effect of gastrointestinal infections on gut microbiota diversity

Read counts after amplicon sequencing for the final study samples ranged from 12,150 to 332,157 with a median of 56,303. A total of 6,790 different OTUs were detected. However, 5,749 of these were present in less than ten samples.

To assess the effect of bacterial and protozoal infection on the number of gut microbiota organisms within an individual, α-diversity was estimated. Compared to controls, Faith's PD was significantly decreased in patients suffering from a protozoal infection (22.95±3.34 vs. 15.86±6.33) and further decreased during bacterial infection (10.28±3.98) (**Fig. 2.1A**). This trend was also apparent when samples were stratified by aetiological agent with patients testing positive for *G. lamblia* exhibiting the smallest, yet still significant, difference in Faith's PD from controls (16.32±6.28) and

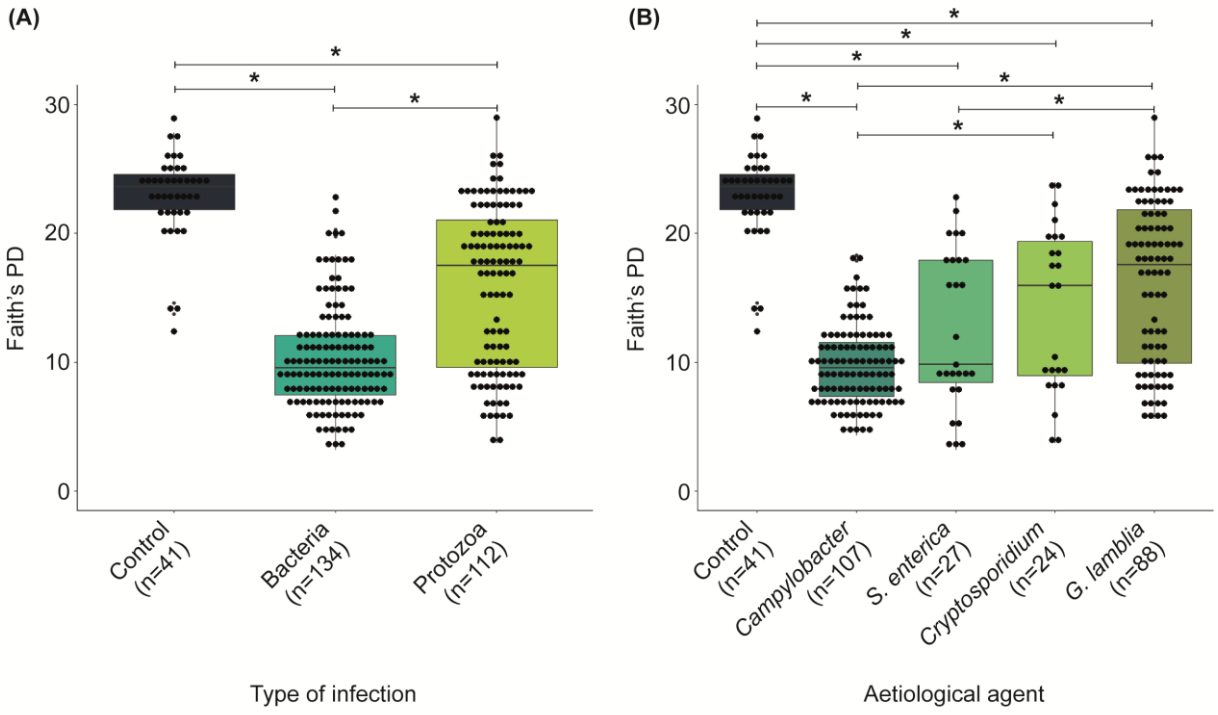

Figure 2.1: Effect of gastrointestinal (GI) infection on gut microbiota α-diversity. Faith's phylogenetic diversity (PD) after taxonomic classification of 16S rRNA gene sequences detected in 287 stool samples from healthy controls and patients with GI infections. (A) Samples stratified by type of infection. (B) Samples stratified by pathogen identified during routine diagnostics. Boxes show medians and upper and lower quartiles, dots represent individual measurements. Asterisks denote a significant difference in Faith's PD between groups, as determined by Dunn's test with Benjamini-Hochberg correction (p≤0.05).

Campylobacter-positive patients showing the lowest diversity (9.72±3.03) (**Fig. 2.1B**). Patients suffering from salmonellosis or cryptosporidiosis could not be distinguished based on their gut microbiota diversity and neither could patients who carried different pathogens causing the same type of infection. The pattern described could also be reproduced when using the number of observed OTUs or Shannon's index as α-diversity metric.

Other sample characteristics were also found to have an effect on Faith's PD: There was a decrease in diversity with increasing storage time at 4°C as well as significant differences between the three sequencing centres and the years the samples were processed. Furthermore, solid stools showed a higher diversity than loose ones.

After estimation of diversity within samples, differences in overall taxonomic composition between samples were investigated using PCoA of weighted UniFrac distances. Principal component (PC) 1 and PC2 explained 41.44% of the total variance. While most control samples formed a cluster, indicating similarities in taxonomic composition, samples testing positive for bacterial pathogens were more dispersed (**Fig. 2.2A**). This suggests that bacterial infection causes a shift in microbiota composition away from the healthy state with the exact nature of this shift varying between individuals. The difference to healthy samples was less pronounced during protozoal infection as the majority of patients testing positive for protozoal pathogens clustered close to the controls. To understand which taxa drove the observed separation, a biplot was generated. Lower PC1 scores, which were obtained for the control cluster and many samples from patients suffering from protozoal gastroenteritis, were associated with higher abundances of *Blautia* and *Collinsella* while lower PC2 scores, obtained for a subgroup of samples positive for protozoal pathogens, were characterised by a higher abundance of *Prevotella* (**Fig. 2.2B**). Higher PC2 scores as seen for most controls, on the other hand, were linked to higher levels of *Lactobacilli*. A greater distance from the control cluster, observed for the majority of samples testing positive for *Campylobacter* and *Salmonella*, was accompanied by higher abundances of Enterobacteriaceae, *Escherichia-Shigella* and *Bacteroides*. Pairwise comparisons of weighted UniFrac distances further highlighted the patterns observed in the PCoA plots: Controls were more dissimilar to infection

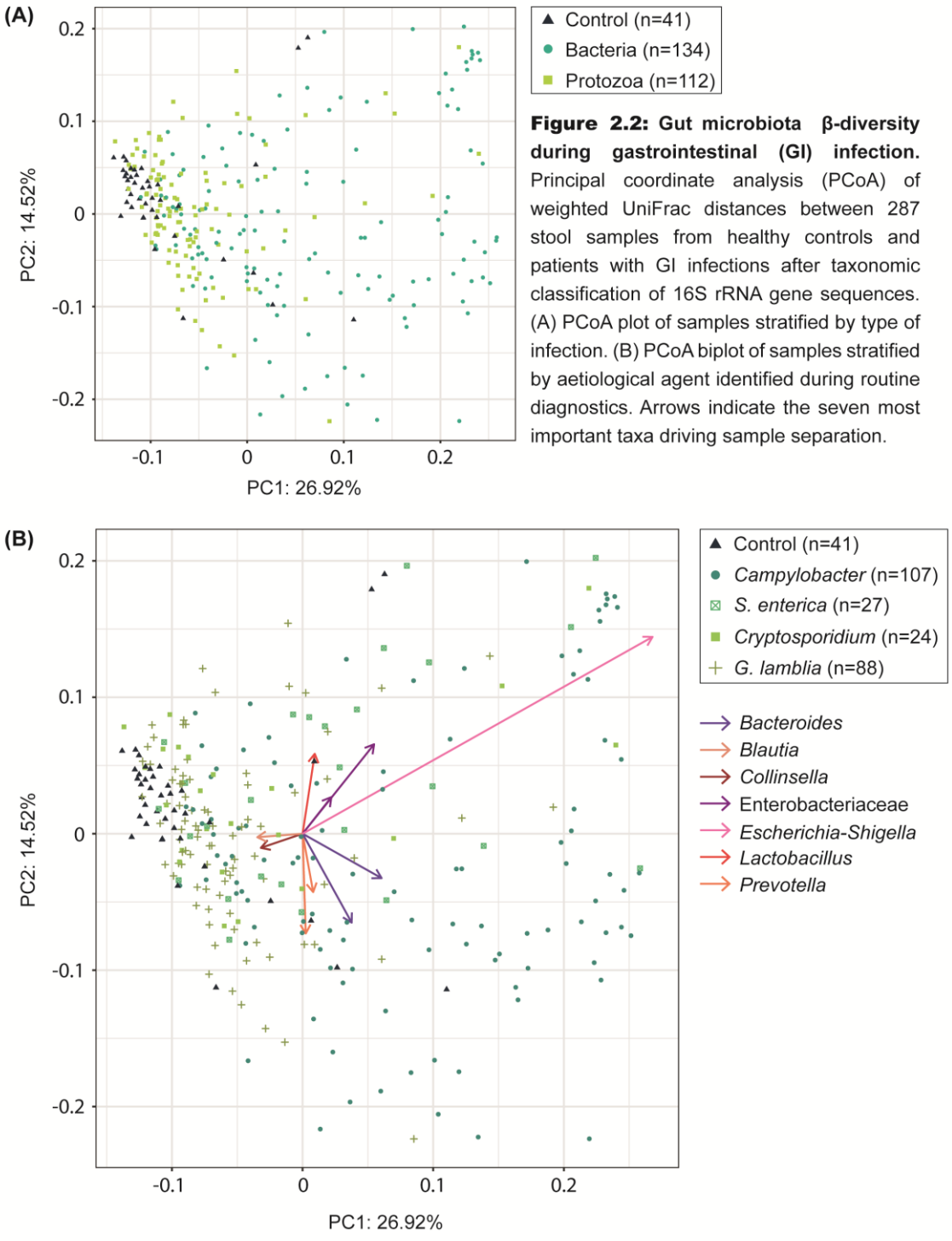

Figure 2.2: Gut microbiota β-diversity during gastrointestinal (GI) infection. Principal coordinate analysis (PCoA) of weighted UniFrac distances between 287 stool samples from healthy controls and patients with GI infections after taxonomic classification of 16S rRNA gene sequences. (A) PCoA plot of samples stratified by type of infection. (B) PCoA biplot of samples stratified by aetiological agent identified during routine diagnostics. Arrows indicate the seven most important taxa driving sample separation.

samples than to one another (**Fig. 2.3A**). On the other hand, as indicated by their dispersal on the PCoA plot, samples testing positive for a bacterial pathogen were equally dissimilar to one another than they were to controls and samples positive for a protozoal infection (**Fig. 2.3B**). PERMANOVA revealed significant dissimilarity

between controls and infection samples and between samples positive for bacterial and protozoal infection (q=0.001).

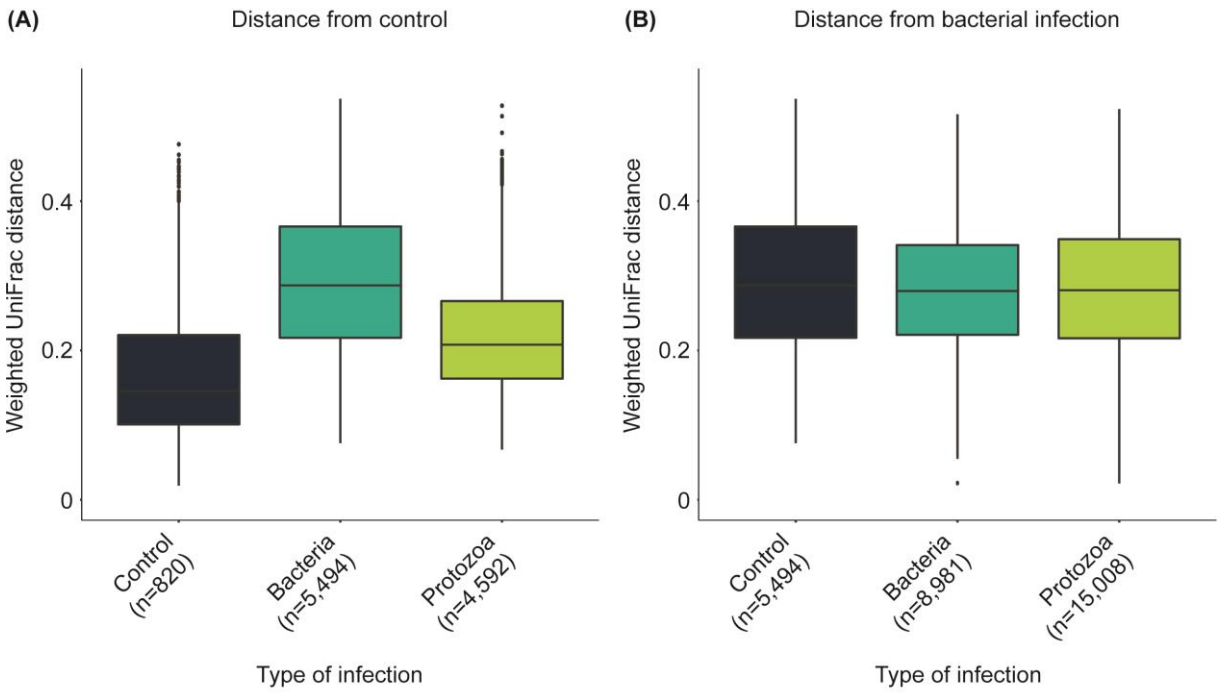

Figure 2.3: Gut microbiota dissimilarity during gastrointestinal (GI) infection. Weighted UniFrac distances for 287 stool samples from healthy controls and patients with GI infections after taxonomic classification of 16S rRNA gene sequences. (A) Pairwise distances from control samples. (B) Pairwise distances from samples testing positive for bacterial infections. Samples are stratified by type of infection. Boxes show medians and upper and lower quartiles.

2.5.2.2 Changes in abundance of specific taxa during gastrointestinal infections

Given that 16S rRNA gene sequencing targets a region not present in eukaryotic genomes, it is unsuitable for direct detection of *Cryptosporidium* and *G. lamblia*. However, identification of prokaryotic pathogens should be possible. Only few bacteria could be classified down to species level due to the inherent limitations of the method and these organisms did not include *Campylobacter* species and *S. enterica*. At genus level, while *Campylobacter* was detected in 79 out of 107 (73.83%) samples testing positive for the pathogen during routine diagnostics, with relative abundances ranging from <0.01% to 11.71%, no sequences were assigned to *Salmonella* in the *S. enterica*-positive samples. *Campylobacter* was additionally

detected in 13 of the 108 (12.04%) remaining samples at relative abundances <0.01%-1.23%.

Due to the limited suitability of 16S rRNA gene sequencing for direct pathogen detection, the possibility of using other taxa which undergo changes in relative abundance during infection as biomarkers was investigated using linear model fits. Fold changes determined based on these models illustrate the ratio of relative abundances between two sample groups. Storage time of the sample at 4°C, year of DNA extraction and sequencing centre were added as covariates. To ensure that the potential biomarkers are carried by a large proportion of the population, only taxa present at a relative abundance >0.1% in at least ten samples were included in the analysis.

At the phylum level, the abundance of Firmicutes was decreased during both bacterial and protozoal infection compared to controls (FC=16 and FC=15.89, respectively) (**Figure 2.4**). Furthermore, bacterial infection led to a reduction in the abundance of Actinobacteria (FC=13.35) No phyla were found to show significant

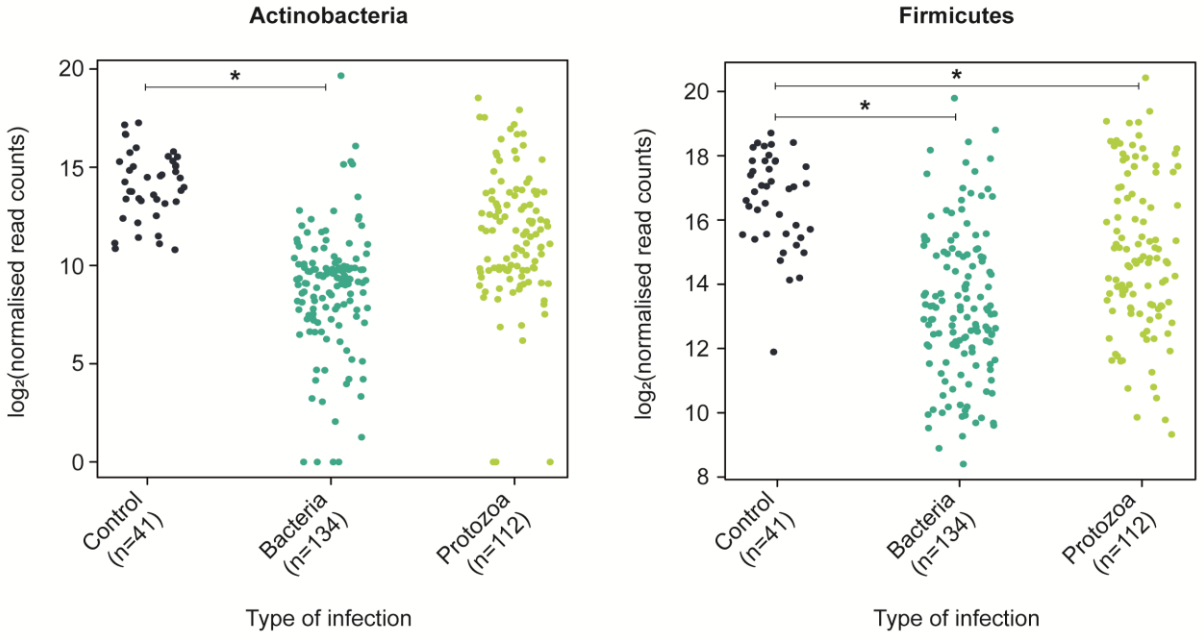

Figure 2.4: Phylum-level differences in gut microbiota composition during gastrointestinal (GI) infection. Log_2 of normalised read counts for Actinobacteria (left) and Firmicutes (right) in 287 samples from healthy controls and patients with GI infections after taxonomic classification of 16S rRNA gene sequences. Samples are stratified by type of infection. Asterisks denote a significant difference in read counts between groups, as determined by empirical Bayes statistics on linear model fits (adjusted p-value ≤0.05). Storage time of the stool sample at 4°C, year of DNA extraction and sequencing centre were added to the analysis as covariates.

differences in abundance between patients suffering from bacterial and protozoal infection. When stratified by aetiological agent, the abundance of Firmicutes was decreased in all groups compared to controls (FC range: 15.82-16.04) and *Campylobacter*- (FC=13.49) as well as *S. enterica*-positive patients (FC=13.19) showed a decreased abundance of Actinobacteria (FC=13.49 and FC=13.19, respectively).

A total of 41 genera present at a relative abundance >0.1% in at least ten samples were differentially abundant between groups. The decrease in Firmicutes observed during infection was due to changes in abundance of members of the Erysipelotrichaceae, Lachnospiraceae, Peptostreptococcaceae, Ruminococcaceae and Veillonellaceae families, members of the Clostridiales Family XIII AD3011 and Christensenellaceae R-7 groups as well as *Clostridium sensu stricto* 1 and *Lactococcus*. Actinobacteria genera undergoing changes included *Collinsella*, *Libanicoccus* and members of the Eggerthellaceae family. Further changes were observed in *Akkermansia*, *Barnesiella*, the gammaproteobacteria *Morganella* and *Parasutterella*, and genera of Prevotellaceae.

Inspection of normalised read counts revealed that most differentially abundant genera were enriched in controls (**Figure 2.5**). Only *Akkermansia*, *Barnesiella*, *Catenibacterium*, *Eggerthella*, *Flavonifractor*, *Libanicoccus*, *Parasutterella*, *Prevotella* 2, *Veillonella*, and members of the *Ruminococcus torques* and Lachnospiraceae UCG-010 groups were more abundant in infection samples. The majority of significant changes in abundance were universal to all aetiological agents and could therefore not be used for their distinction. Exceptions were *Blautia,* showing similar abundance in controls and during protozoal infection but differing from both during

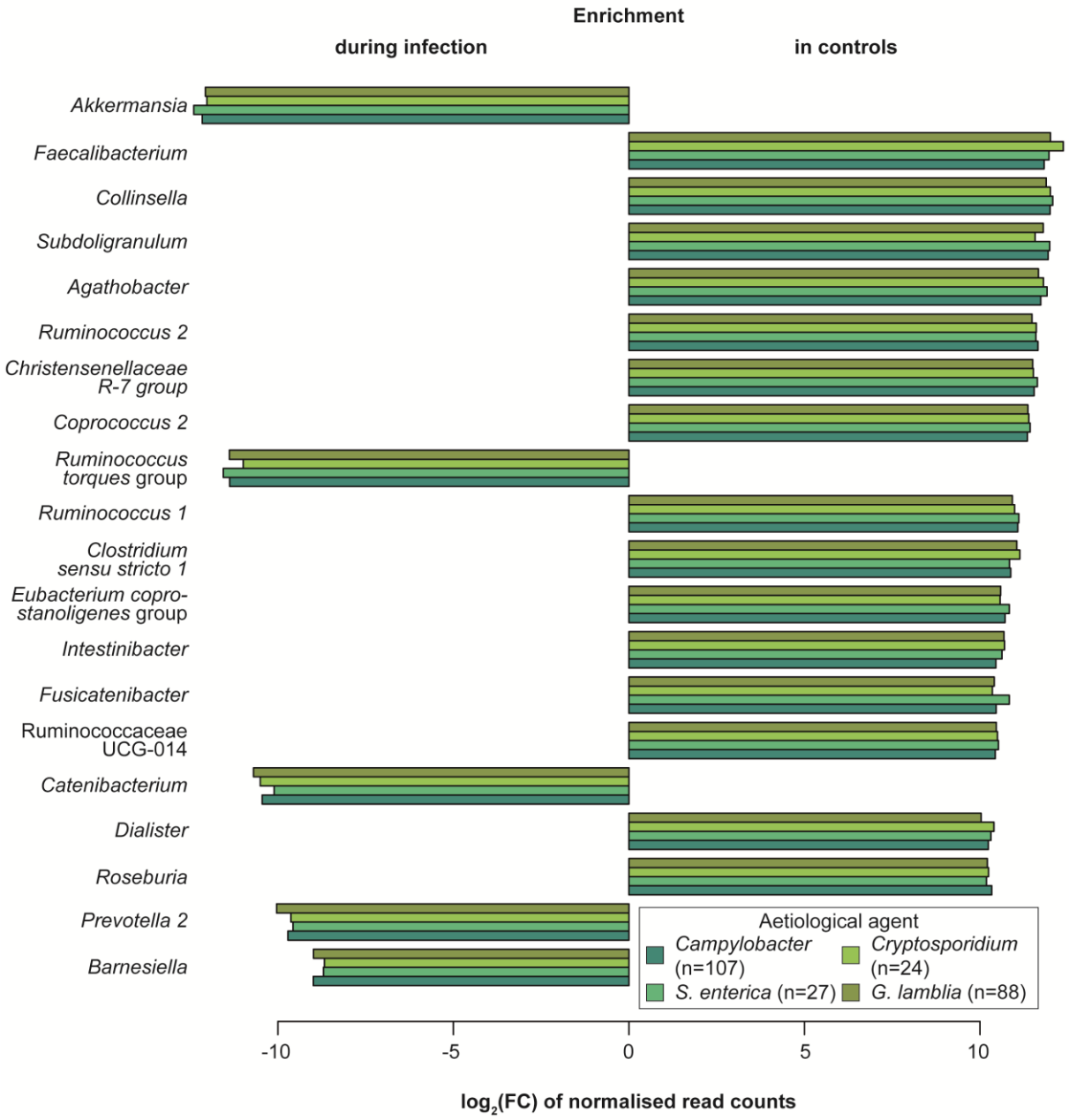

Figure 2.5: Specific changes in abundance of gut microbiota genera during infection. Log_2-transformed fold changes (FCs) of normalised read counts assigned to the genera listed, between 41 controls and samples testing positive for *Campylobacter*, *S. enterica*, *Cryptosporidium* and *G. lamblia*. Positive FCs represent enrichment in control samples, negative FCs enrichment during infection. Only the top 15 genera enriched in controls and the top five genera enriched in infection samples are depicted. All genera shown were differentially abundant between sample groups according to empirical Bayes statistics on linear model fits (adjusted p-value ≤0.05) and present at a relative abundance >0.1% in at least ten samples. Storage time of the stool sample at 4°C, year of DNA extraction and sequencing centre were added to the analysis as covariates.

bacterial infection (FC=12.49 and FC=11.77, respectively), and *Morganella*, which could be a potential biomarker for distinction of *Cryptosporidium* and *G. lamblia* infection (FC=5.87) (**Figure 2.6**).

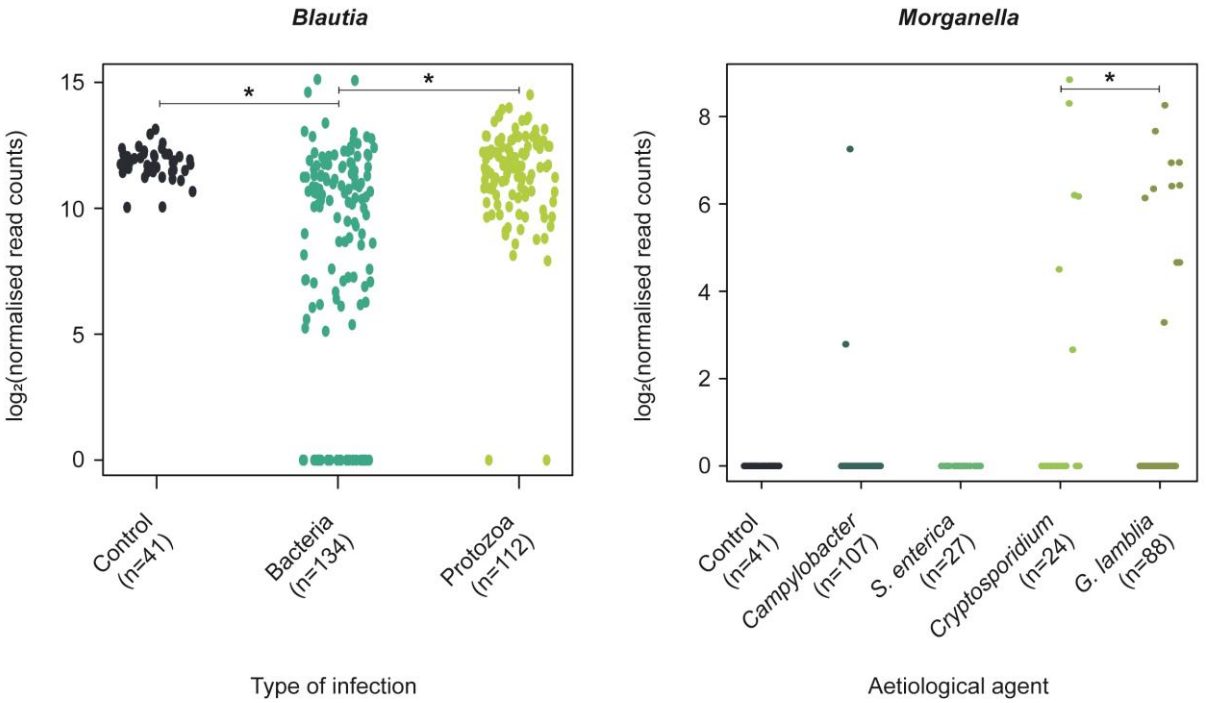

Figure 2.6: Potential genus-level taxonomic biomarkers for distinction of aetiological agents of gastrointestinal (GI) infection. Log_2 of normalised read counts for *Blautia* and *Morganella* in 287 samples from healthy controls and patients with GI infections after taxonomic classification of 16S rRNA gene sequences. Samples are stratified by type of infection (left) or aetiological agent (right) identified during routine diagnostics. Asterisks denote a significant difference in read counts between groups, as determined by empirical Bayes statistics on linear model fits (adjusted p-value ≤0.05). Storage time of the stool sample at 4°C, year of DNA extraction and sequencing centre were added to the analysis as covariates.

2.5.2.3 Prediction of microbiome function from metataxonomic data

Although the metataxonomic approach employed does not provide direct information about the presence of other functional genes, the metagenomic content of the microbiota can be inferred with the help of tools mapping the functions of known sequenced genomes to the taxa identified through 16S rRNA gene sequencing. The PICRUSt2 pipeline identified a total of 99 differentially abundant KEGG pathways between sample groups. In line with previously observed taxonomic differences, most changes were universal to all aetiological agents and most pathways were enriched in controls (**Figure 2.7**). Cell functions these pathways are associated with

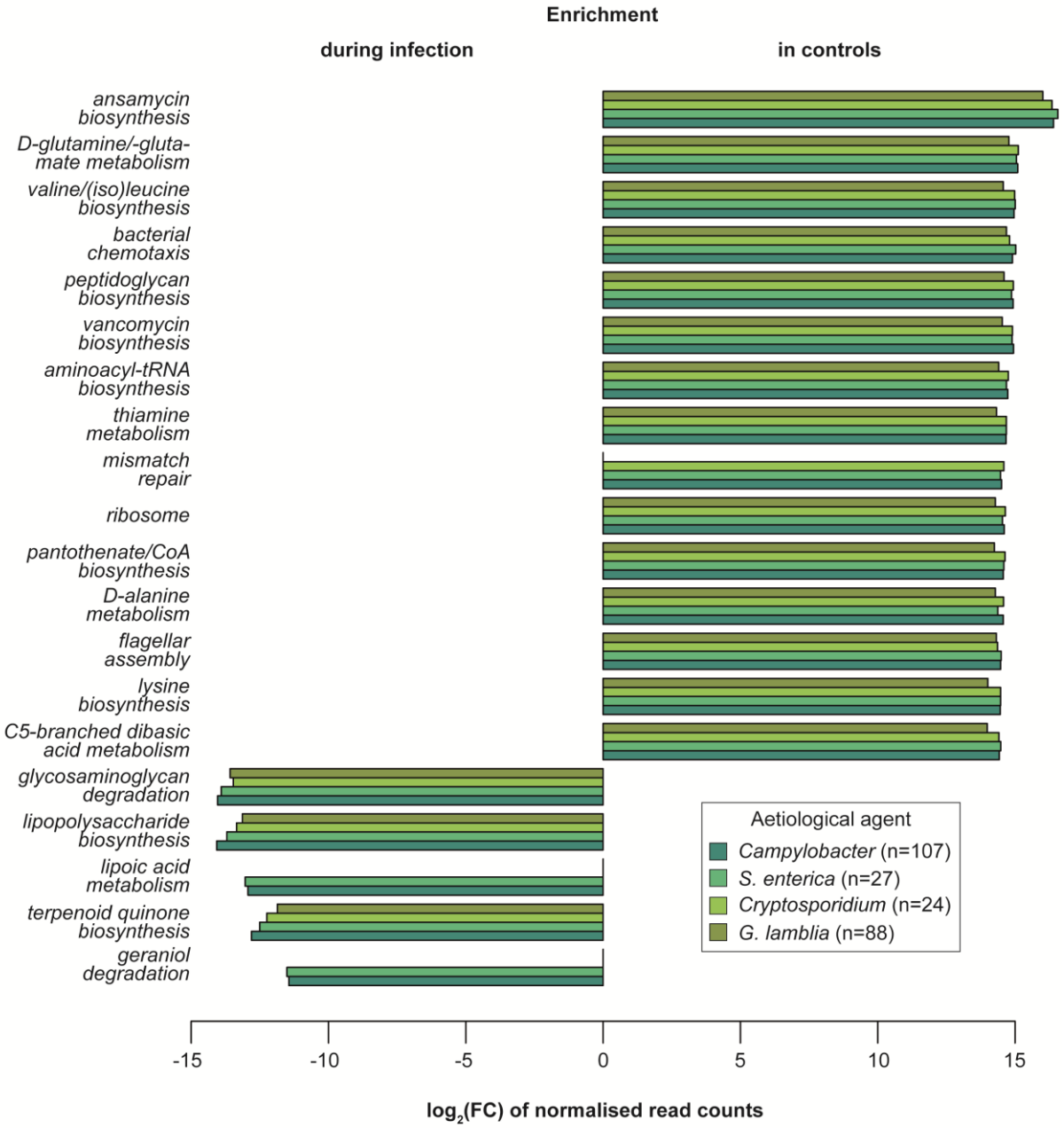

Figure 2.7: Changes in predicted metabolic functions during gastrointestinal infection. Log_2-transformed fold changes (FCs) of normalised read counts assigned to the metabolic pathways listed after prediction of functional metagenome content from metataxonomic abundance data, between 41 controls and samples testing positive for *Campylobacter, S. enterica*, *Cryptosporidium* and *G. lamblia*. Positive FCs represent enrichment in control samples, negative FCs enrichment during infection. Only the top 15 pathways enriched in controls and the top five pathways enriched in infection samples are depicted. All pathways shown were differentially abundant between sample groups according to empirical Bayes statistics on linear model fits (adjusted p-value ≤0.05) Where bars for an aetiological agent have been omitted, the pathway's abundance did not differ significantly from control for this pathogen. Storage time of the stool sample at 4°C, year of DNA extraction and sequencing centre were added to the analysis as covariates.

include cell motility, glycan biosynthesis, replication and repair, translation and the metabolism of amino acids, carbohydrates, cofactors, vitamins, terpenoids and polyketides. Only glycosaminoglycan degradation, lipopolysaccharide biosynthesis and the synthesis of steroids, their hormones and terpenoid precursors were enriched during all types of infection. Lipoic acid metabolism and geraniol degradation were enriched during bacterial but not during protozoal infection. Enrichment of polycyclic aromatic hydrocarbon degradation was observed solely during *Campylobacter* infection (FC=8.22).

2.5.3 Metabolic signatures of gastrointestinal infections

2.5.3.1 Effect of gastrointestinal infections on gut microbiota metabolic profiles

The predicted functional changes in the gut microbiota during infection suggest that some faecal metabolites could act as biomarkers for different aetiological agents. To assess whether variations in taxonomic composition translate into real functional changes, differences in the microbiota's metabolite profile during infection were identified using NMR spectroscopy. A total of 93 faecal metabolites were studied initially. This list of compounds was based on metabolites commonly identified in human faeces during previous studies at QIB (e.g. Kellingray et al.[192]). Only the 56 metabolites detected at a concentration >0.1 mM in at least ten samples were included in the following analyses to ensure that potential biomarkers are present in a large proportion of stool samples tested.

To detect overall metabolome differences between sample groups, multivariate analysis in the form of PCA was carried out. PC1 and PC2 explained 45.87% of the total variance. Most control samples formed a tight cluster, indicating similarities in functional profiles, while infection samples were more dispersed (**Figure 2.8**). As for taxonomic profiles, this suggests that infection causes a shift away from the healthy state with the exact nature of the shift varying between individuals. However, differences in metabolite profiles appeared less pronounced than differences in community composition (cf. **Figure 2.2**).

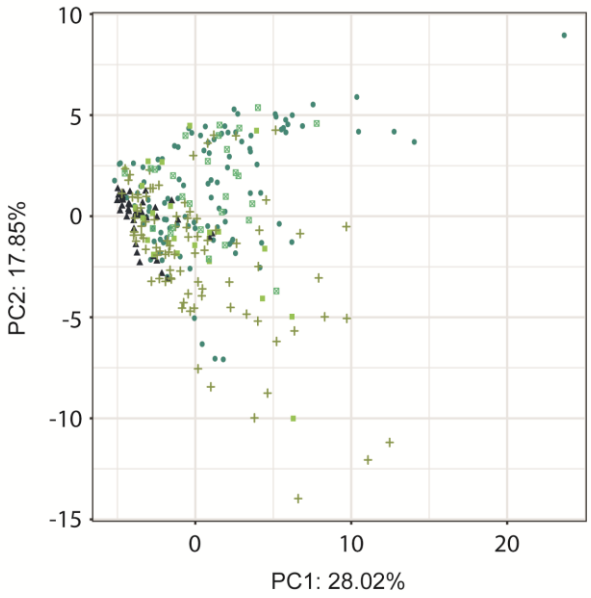

Figure 2.8: Gut microbiota metabolic profiles during gastrointestinal (GI) infection. Principal component analysis of the concentrations of 56 metabolites identified by nuclear magnetic resonance spectroscopy of 287 faecal waters from healthy controls and patients with GI infections. Samples are stratified by aetiological agent identified during routine diagnostics.

Investigation of loading weights on the two axes revealed which metabolites primarily drove the separation of samples: On PC1, the carboxylic acid acetate, the nucleobase uracil and several proteinogenic as well as the non-proteinogenic amino acids citrulline, ornithine and taurine were the main contributors to sample variance (**Figure 2.9**). Methylamine, ribose, phenylacetate, 2-methylbutyrate, the amino acids aspartate, leucine and proline, the nucleobases hypoxanthine and uracil and several carboxylic acids as well as short and medium chain fatty acids were responsible for separation along PC2.

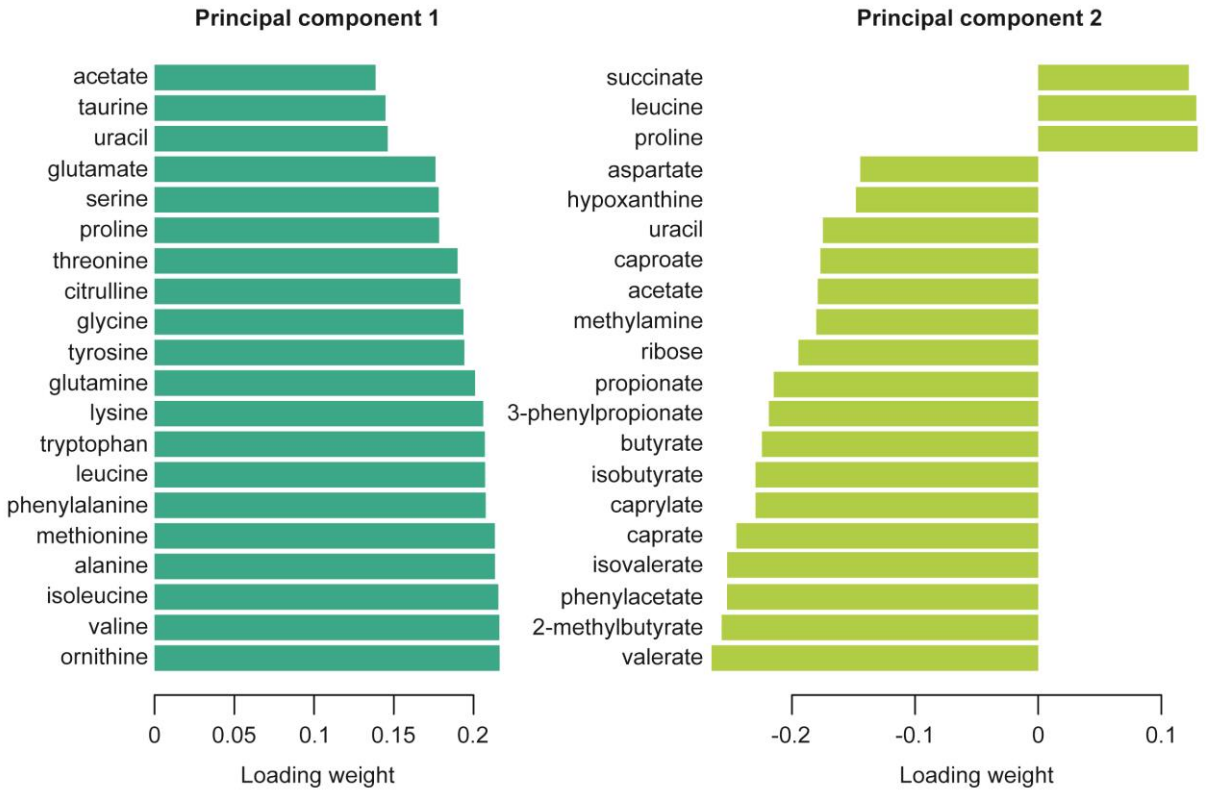

Figure 2.9: Contribution of individual compounds to differences in gut microbiota metabolite profiles. Loading weights of metabolites on principal component (PC) 1 (left) and PC2 (right) determined by PC analysis of their concentrations in 287 faecal waters from healthy controls and patients with gastrointestinal infections. Only the top 20 metabolites for each PC are shown.

2.5.3.2 Changes in concentration of specific metabolites during gastrointestinal infections

The PCA loading weights give an indication of which metabolites might be suitable biomarkers for different types of GI infections. Of the 56 metabolites analysed in total, the concentrations of 49 differed significantly, either between control and infection samples or between different aetiological agents (**Figure 2.10)**. Mean concentrations of all metabolites in the different sample groups can be found in **Appendix II**. Hierarchical clustering based on the concentrations of the 49 metabolites grouped most control samples together. Samples from patients with bacterial infections, on the other hand, showed high variation within the group. Several subgroups were observed for samples from patients suffering from protozoal gastroenteritis: While

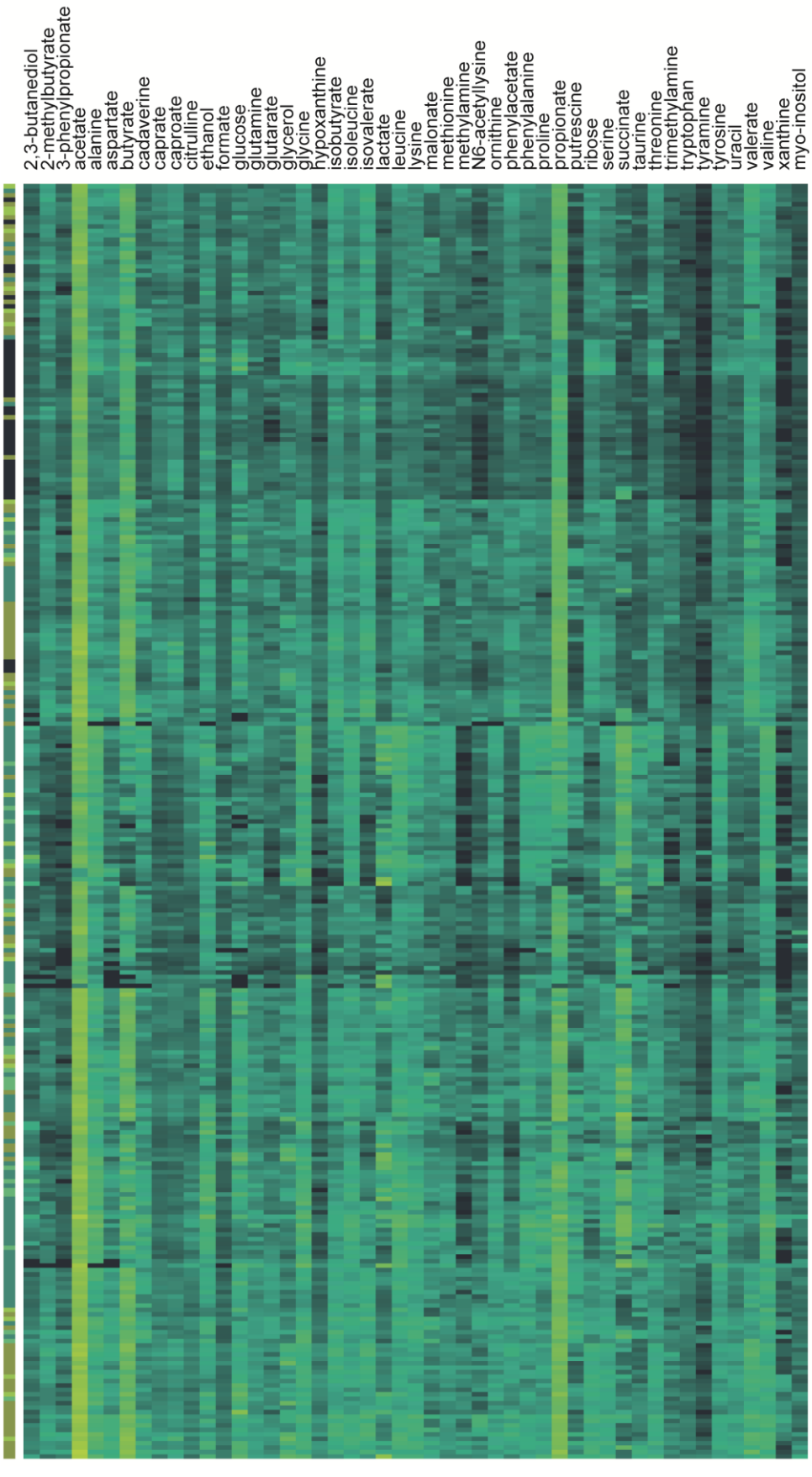

Figure 2.10: Gut microbiota metabolite concentrations during gastrointestinal (GI) infections. Log_{10}-normalised concentrations of specific metabolites determined by nuclear magnetic resonance spectroscopy of 287 faecal waters from healthy controls and patients with GI infections. Heatmap columns were reordered based on the similarity of samples to one another. Coloured bars above the heatmap indicate the aetiological agent identified in the sample during routine diagnostics. Only metabolites showing significant differences between sample groups according to Dunn's tests (adjusted p-value ≤0.05) and present at a concentration >0.1 mM in at least ten samples are depicted.

some formed two separate distinct clusters and some were grouped with control samples, others showed a variation similar to that caused by bacterial infection.

Most metabolite differences were based on increased concentrations during infection. Examples include the levels of acetate rising from 3,339.79±1,659.66 mM in controls to 6,778±3,998.90 mM during bacterial and 7,993±5,362.13 mM during protozoal infection, an increase in lactate concentration from 20.39±16.89 mM to 572.29±1,607.60 mM and 256.02±1,163.91 mM and putrescine levels changing from 3.89±2.74 mM to 144.44±141.71 mM and 64.57±97.27 mM. Butyrate, hypoxanthine, isovalerate and phenylacetate were revealed as candidate biomarkers for protozoal infections, with concentrations increasing in this sample group only (**Figure 2.11**).

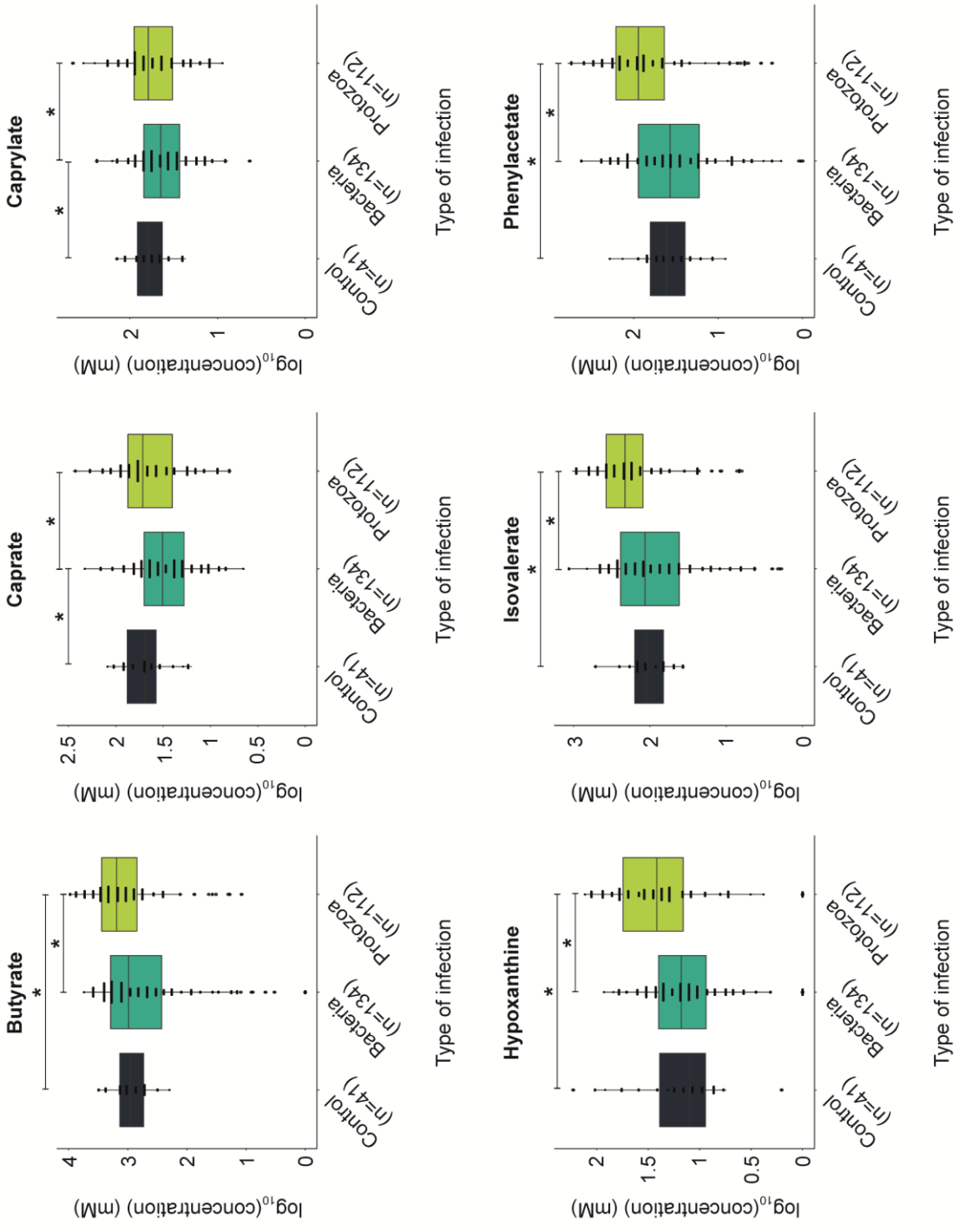

Figure 2.11: Potential gut microbiota metabolite biomarkers for distinction of gastrointestinal (GI) infection types. Log_{10}-normalised concentrations of specific metabolites, determined by nuclear magnetic resonance spectroscopy of 287 faecal waters from healthy controls and patients with GI infections. Samples are stratified by the type of infection identified during routine diagnostics. Boxes show medians and upper and lower quartiles, dots represent individual measurements. Asterisks denote a significant difference in metabolite concentration between groups, as determined by Dunn's test (adjusted p-value ≤0.05).

During bacterial infection, the levels of caprate and caprylate were decreased compared to controls and samples from patients with protozoal gastroenteritis. Aspartate was identified as a potential biomarker for campylobacteriosis with the levels of this metabolite increasing during infection with all other aetiological agents but sustaining control levels during *Campylobacter* infection (**Figure 2.12**). For distinction of pathogens causing the same type of infection, 3-phenylpropionate and serine were identified, with the concentrations of the former differing during cryptosporidiosis and giardiasis (22.96±38.04 mM vs. 43.82±53.35 mM), and the concentrations of the latter differing during campylobacteriosis and salmonellosis (129.06±101.10 mM vs. 191.74±174.79 mM).

Stool sample storage time at 4°C before faecal water preparation also had an effect on metabolite concentrations. In most cases, metabolite levels increased with increasing storage time before freezing and most significant differences found were between samples that had been frozen on the day of production and those that had not, regardless of the exact number of days at 4°C.

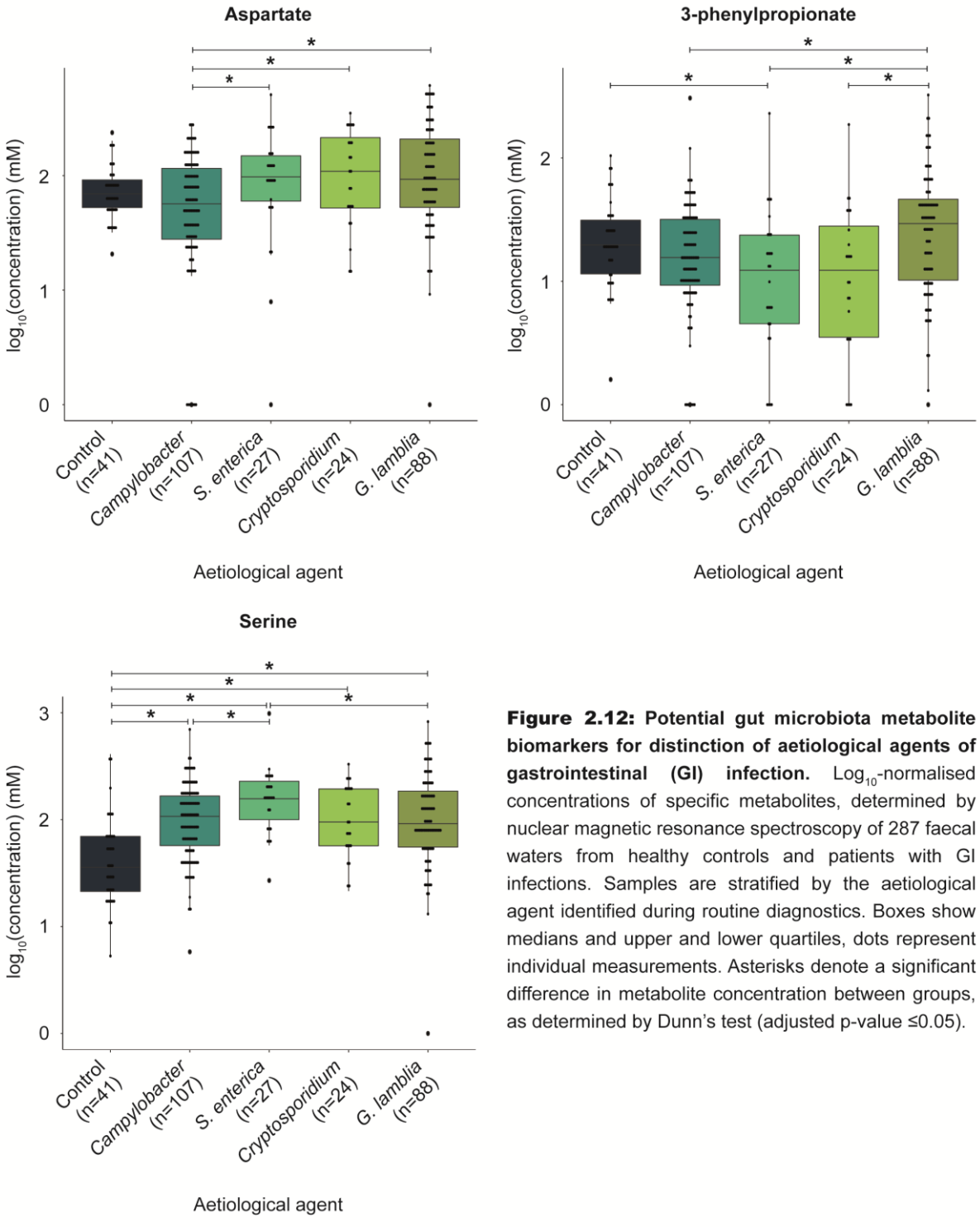

Figure 2.12: Potential gut microbiota metabolite biomarkers for distinction of aetiological agents of gastrointestinal (GI) infection. Log_{10}-normalised concentrations of specific metabolites, determined by nuclear magnetic resonance spectroscopy of 287 faecal waters from healthy controls and patients with GI infections. Samples are stratified by the aetiological agent identified during routine diagnostics. Boxes show medians and upper and lower quartiles, dots represent individual measurements. Asterisks denote a significant difference in metabolite concentration between groups, as determined by Dunn's test (adjusted p-value ≤0.05).

2.5.4 Correlation of taxonomic and metabolic gut microbiota features

So far, metataxonomic and metabolomic features have been analysed as separate entities. In the environment of the gut microbiota, however, they are interconnected with alterations in the abundance of certain microbes resulting in alterations in the levels of metabolites they produce or utilise, which can in turn inhibit or favour the growth of other taxa. To assess whether the two datasets are statistically correlated, sPLS regression was performed on clr-transformed relative abundance data for genera making up >0.1% of the community in at least ten samples and log_{10}-transformed concentrations of metabolites present at >0.1 mM in at least ten samples. The coefficient of determination (R^2) of the resulting linear model was 0.51 on canonical axis 1 (**Figure 2.13**) suggesting that the datasets correlated well for a subset of samples only. When repeating the analysis for all sample groups separately, R^2 ranged from 0.36 for giardiasis samples to 0.81 for controls.

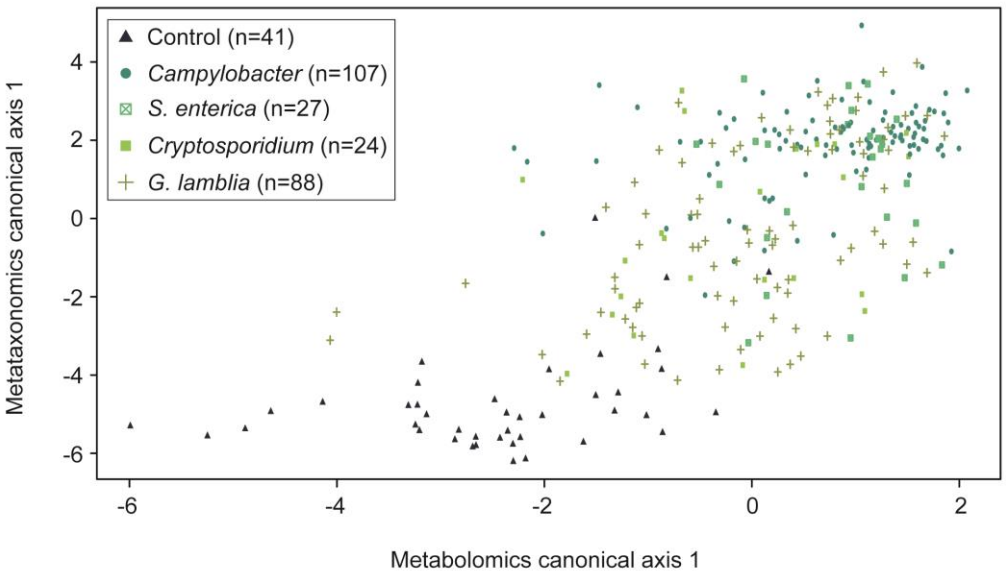

Figure 2.13: Correlation between metataxonomic and metabolomic datasets. Scores of clr-transformed gut microbiota taxa relative abundances, determined by 16S rRNA gene sequencing, and log_{10}-transformed gut microbiota metabolite concentrations, determined by nuclear magnetic resonance spectroscopy, on canonical axes 1 after sparse partial least squares regression. Data was acquired from 287 faecal samples of healthy controls and patients with gastrointestinal infections. Samples are stratified by aetiological agent identified during routine diagnostics.

Evaluation of correlation coefficients for individual features revealed three distinct clusters (**Figure 2.14**). Relative abundances of *Enterorhabdus* and several Bacteroidetes, Clostridiales and Erysipelotrichaceae genera, were positively correlated with the concentrations of the carbohydrate myo-inositol, the carboxylic

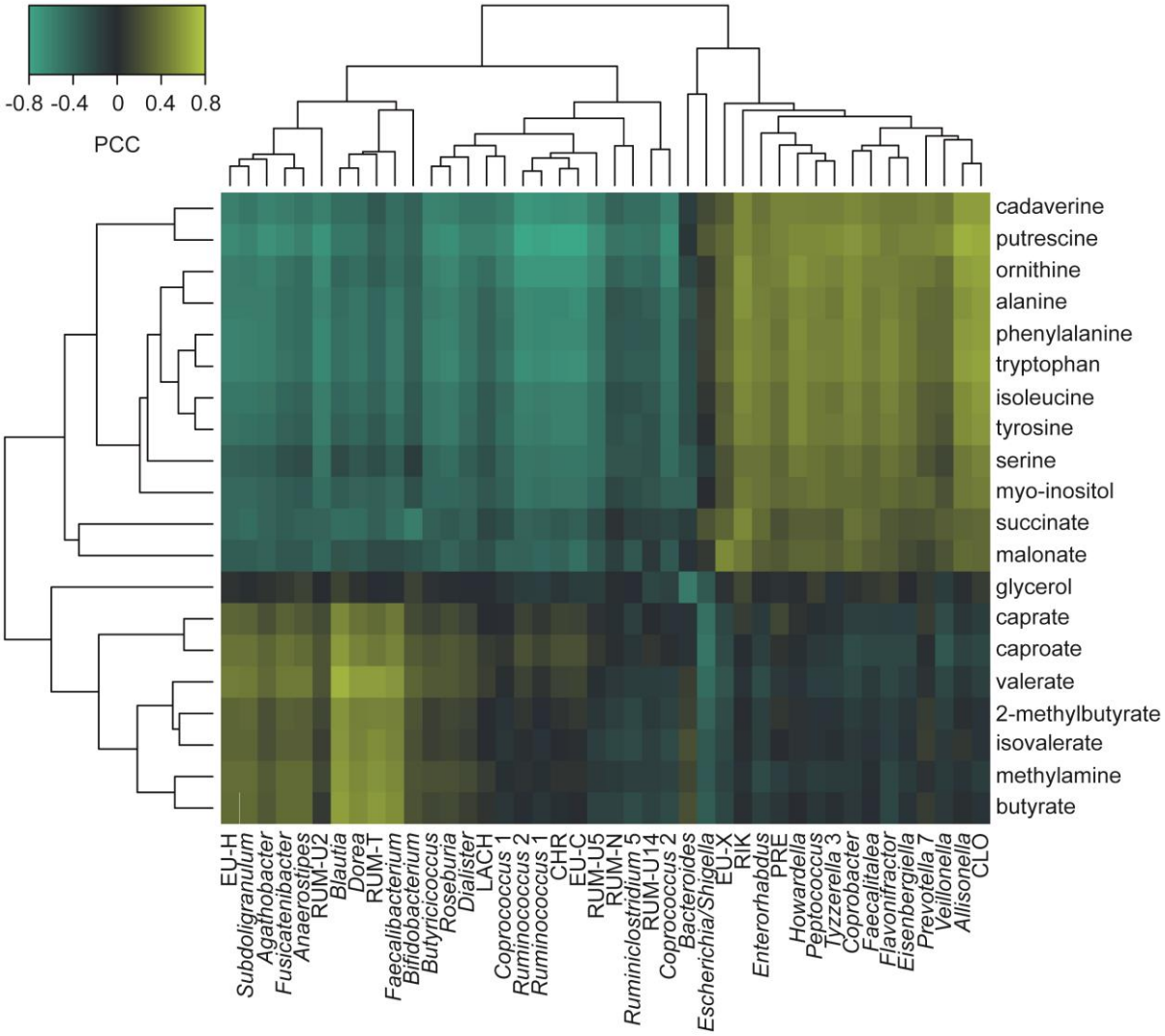

Figure 2.14: Correlation between specific gut microbiota features. Clustered image map of Pearson's correlation coefficients (PCC), determined by sparse partial least squares analysis, for clr-transformed relative abundances of selected gut microbiota genera (columns) and $\log_{10}$-transformed concentrations of selected gut microbiota metabolites (rows) after 16S rRNA gene sequencing and nuclear magnetic resonance spectroscopy. Data was acquired from 287 faecal samples of healthy controls and patients with gastrointestinal infections. Negative values (turquoise) indicate a negative, positive values (green) a positive correlation between the features. Only correlations with an absolute value ≥ 0.5 are shown. EU-H, *Eubacterium hallii* group; RUM-U2, Ruminococcaceae UCG-002; RUM-T, *Ruminococcus torques* group; LACH, Lachnospiraceae ND3007 group; CHR, Christensenellaceae R-7 group; EU-C, *Eubacterium coprostanoligenes* group; RUM-U5, Ruminococcaceae UCG-005; RUM-N, Ruminococcaceae NK4A214 group; RUM-U14, Ruminococcaceae UCG-014; EU-X, *Eubacterium xylanophilum* group; RIK, Rikenellaceae RC-9 gut group; PRE, Prevotellaceae NK3B31 group; CLO, *Clostridium inocuum* group.

acids malonate and succinate, and of amino acids and their breakdown products. Abundances of several Lachnospiraceae and Ruminococcaceae genera, on the other hand, were negatively correlated with the levels of these metabolites and positively correlated with the concentrations of fatty acids and their derivatives.

2.5.5 Prediction of aetiological agents of gastrointestinal infections by Random Forest classification

The previous analyses suggest that some gut microbiota changes introduced by infection might be pathogen-specific. However, even when there are statistical differences between sample groups, individual values within the groups for any given feature vary widely. Thus, none of the candidate biomarkers mentioned, when measured in isolation, will provide sufficient discriminatory power to guarantee unambiguous identification of the aetiological agent. Combining measurements of several features from both the metataxonomic and metabolomics dataset, on the other hand, could improve discriminatory power. To evaluate which features should form part of this set of measurements and how well they perform in distinguishing between different infections, a Random Forest classifier was built based on Faith's PD, relative abundance data for genera making up >0.1% of the community and concentrations of metabolites present at >0.1 mM in at least ten samples.

Overall mean accuracy of distinguishing samples derived from patients suffering from two different types of GI infection and controls was 81.61%. Within the validation dataset, all control samples were correctly identified as controls and none of the infection samples were misidentified as controls (**Table 2.2**). Of the

Table 2.2: Random Forest prediction accuracy for the type of gastrointestinal infection

		Diagnosis		
		Control	**Bacteria**	**Protozoa**
Prediction	**Control**	13 (100%)	0	0
	Bacteria	0	29 (80.56%)	9 (23.68%)
	Protozoa	0	7 (19.44%)	29 (76.32%)

Values are based on a training dataset comprising 200 samples and a validation dataset comprising 87 samples.

gastroenteritis patients, however, 19.44% with bacterial gastroenteritis were incorrectly predicted to suffer from protozoal infection and 23.86% diagnosed with protozoal infection were predicted to suffer from bacterial gastroenteritis. Relative abundance of the *Ruminococcus* 1 genus was the biggest contributor to the model's discriminatory power (**Figure 2.15A**). Other taxonomic features important for prediction accuracy included members of the Christensenellaceae R-7 group, *Coprococcus* 2, *Romboutsia*, *Campylobacter* and *Turicibacter*. Ribose, lysine, uracil and succinate were the main metabolomic contributors.

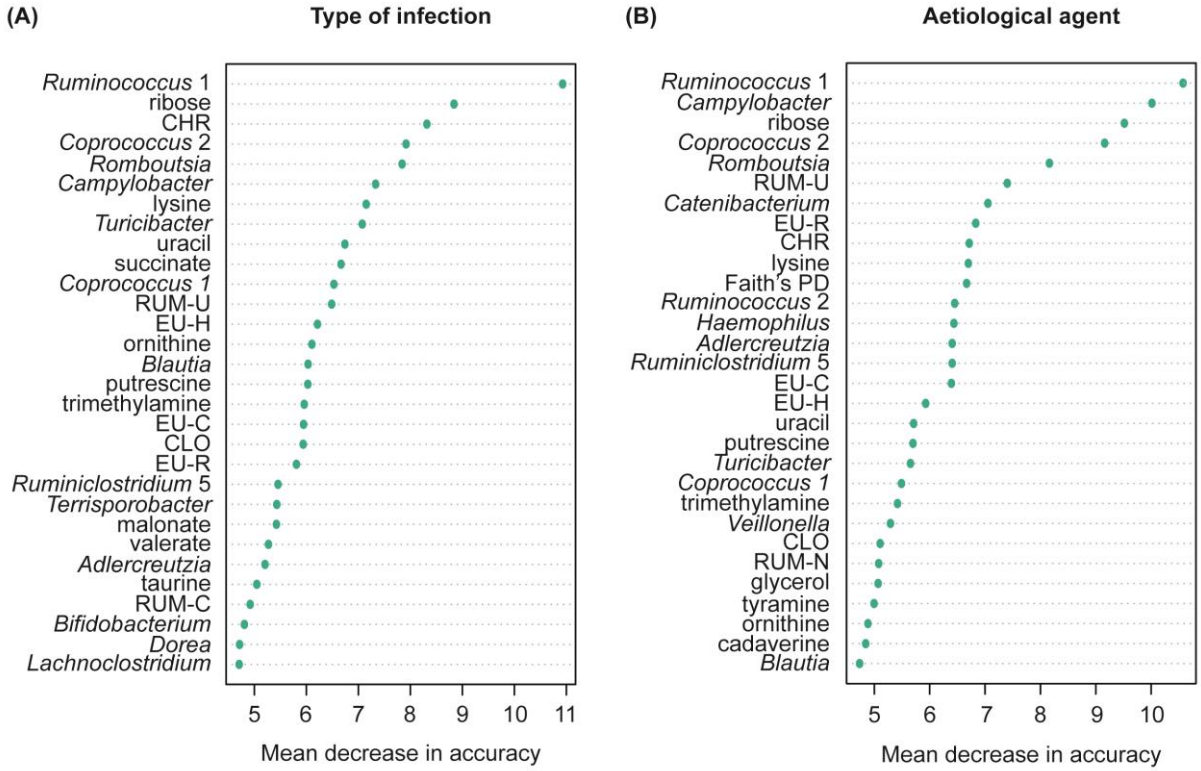

Figure 2.15: Contribution of gut microbiota features to the discriminatory power of a gastrointestinal (GI) infection aetiology classifier. Mean decrease in prediction accuracy for (A) type of infection and (B) aetiological agent identified in routine diagnostics when excluding the features listed during Random Forest classification of 287 stool samples from controls and patients with GI infections of different aetiologies. Only the 30 most important features for accuracy of prediction are shown. CHR, Christensenellaceae R-7 group; RUM-U, Ruminococcaceae UCG-014; EU-H, *Eubacterium halli* group; EU-C, *Eubacterium coprostalonigenes* group; CLO, *Clostridium sensu stricto* 1; EU-R, *Eubacterium ruminantium* group; RUM-C, Ruminococcaceae CAG-352; RUM-N, Ruminococcaceae NK4A214 group.

For identification of the exact aetiological agent, overall mean prediction accuracy dropped to 69%. While all controls were again identified correctly, 14.29% of *Campylobacter*-positive samples were misclassified as *G. lamblia*-positive and 17.86% of giardiasis samples were wrongly predicted to be positive for

Campylobacter (**Table 2.3**). None of the samples testing positive for *S. enterica* or *Cryptosporidium* were classified correctly, being identified as campylobacteriosis (62.5% and 17.86%, respectively) or giardiasis samples (37.5% and 82.14%, respectively). Although their individual weights differed slightly, many of the main contributors to prediction accuracy overlapped for the type of infection and the aetiological agent with the *Ruminococcus* 1 genus being the most important taxonomic and ribose the most important metabolomic feature (**Figure 2.15B**). Interestingly, Faith's PD was amongst the most important features for accurate prediction of the aetiological agent but not for the type of infection.

Table 2.3: Random Forest prediction accuracy for the aetiological agent of gastrointestinal infection

		Diagnosis				
		Control	***Campylo-bacter***	***S. enterica***	***Crypto-sporidium***	***G. lamblia***
Prediction	**Control**	13 (100%)	0	0	0	0
	Campylobacter	0	24 (85.71%)	5 (62.5%)	3 (30%)	5 (17.86%)
	S. enterica	0	0	0	0	0
	Cryptosporidium	0	0	0	0	0
	G. lamblia	0	4 (14.29%)	3 (37.5%)	7 (70%)	23 (82.14%)

Values are based on a training dataset comprising 200 samples and a validation dataset comprising 87 samples.

2.6 Discussion

2.6.1 Gut microbiota taxonomic and metabolite profiles in health and disease

As reported previously[163-168], acute gastroenteritis induced a shift of the gut microbiota away from the healthy state in the present study. The variation introduced by the diseased state appeared to be larger than expected interindividual differences, as indicated by the fact that control samples from 41 different healthy individuals formed a tight cluster, clearly distinct from the infection population during multivariate analysis of both 16S rRNA gene sequencing-derived taxonomic and NMR-derived metabolic profiles. A major characteristic of infection-induced dysbiosis is a decrease in α-diversity[163,167], which could be reproduced for all aetiological agents. Given that diarrhoea can lead to a loss of 200 g of stool and fluids per day[191], many microorganisms will be lost along with this matter. Especially in the case of BSC Type 7 stools the remaining microbes will be diluted in higher than usual volumes of liquid, which likely explains why bacterial diversity was lowest in loose faeces. In line with this, both the number of loose stools and the reduction in α-diversity was greater in patients suffering from bacterial infection than for protozoal infection.

The PCoA biplot suggests that samples at a greater distance from healthy controls, i.e. those patients experiencing the most severe infection-induced dysbiosis, are characterised by a high abundance of *Escherichia*. This genus seems adept at occupying niches which open up in a dysbiotic microbiota as a bloom of *Escherichia* has, for example, been observed in patients suffering from recurrent *C. difficile* infection prior to faecal microbiota transplants[192]. A decrease in Firmicutes abundance seems to be another characteristic of GI infection, identified in this as well as multiple other studies[166,168,193]. Firmicutes are one of the major phyla constituting a typical, healthy microbiota[1,26] and a reduction might indicate that infection opens up niches to be occupied by other, less common organisms. Similarly, at higher taxonomic resolution, amongst the genera decreased during infection were *Faecalibacterium*, *Ruminococci*, *Coprococcus*, *Clostridium sensu stricto* 1 and *Roseburia*, all major commensal constituents of the microbiota[1,194,195]. A reduction in

several of these genera was previously described during STEC infection[196]. The oxygen hypothesis suggests that dysbiosis can result in increased oxygen availability in the normally anaerobic gut, favouring the growth of facultative over strict anaerobes[197]. The largest change between healthy and diseased state was detected for *Akkermansia*. *A. muciniphila* is a mucin degrader and changes to the normal thickness and constitution of the GI tract's mucous layer might facilitate colonisation by pathogens[198,199].

In line with this, prediction of gut microbiota function from taxonomic composition suggested an increased abundance of pathways promoting colonisation by new microorganisms: both the degradation of glycosaminoglycans and the presence of lipopolysaccharide play a part in surface adhesion[200,201]. Predicted functions support some of the observed metabolic differences: A decrease in glutamine/glutamate and alanine metabolism during infection is reflected in higher concentrations of the compounds in disease samples. Correlation analysis of metataxonomic and metabolomics datasets revealed further relationships between the composition of the microbiota and its function. The concentrations of ketogenic amino acids such as isoleucine, phenylalanine and tryptophan were inversely correlated with abundances of Lachnospiraceae and Ruminococcaceae genera. Members of these bacterial families were found to thrive on media containing proteins and peptones as the sole fermentable substrates[202] indicating they use amino acids as energy sources. Degradation of ketogenic amino acids yields acetyl coenzyme A, a precursor of fatty acids, concentrations of which were positively correlated with Lachnospiraceae and Ruminococcaceae genera. Succinate, levels of which were positively correlated with Bacteroidetes and Clostridiales genera, was reported to enhance colonisation of germfree mice seeded with these microorganisms[203]. Furthermore, members of these taxa are major contributors to proteolytic activity in the gut[204], which could explain their positive correlation with the levels of amino acids.

The pathogenesis of gastroenteritis and associated damage to the host's enteric cells should also be considered when interpreting changes in the gut microbiota's taxonomic and metabolic profile during disease. Controlled absorption of nutrients released by the breakdown of food molecules is reliant on intact enterocytes. Enteric pathogens can cause electrolyte imbalance either by directly modulating ion transport

processes and gut barrier function or indirectly by promoting inflammation and loss of absorptive surfaces[439]. This in turn results in impaired transport of small molecules across the gut epithelium. Giardiasis, for example, is associated with shortening of microvilli due to cytoskeletal reorganisation and induction of enterocyte apoptosis, leading to decreased absorption of glucose[440,441].

2.6.2 16S rRNA gene sequencing and nuclear magnetic resonance spectroscopy as diagnostic tools for gastrointestinal infections

As a direct diagnostic tool, 16S rRNA gene sequencing is inherently limited by the fact that the technology targets a region unique to the prokaryotic genome. It can therefore not be used for identification of eukaryotic and viral pathogens. The 18S rRNA gene and the internal transcribed spacer region are similarly conserved sequences in eukaryotic genomes, widely used for profiling of fungal communities[205,206]. Sequencing of these regions could be used to supplement the data presented with information about non-bacterial taxa. Due to their high genetic diversity and mutation rates, no equivalent marker gene is available for viruses. The 16S rRNA gene sequencing approach completely failed to identify the bacterial pathogens detected during routine diagnostics at species level and sequences assigned to the genus *Campylobacter* were found in only a subset of campylobacteriosis samples. The PCoA biplot showed that many salmonellosis samples were characterised by changes in the abundance of Enterobacteriaceae, indicating that *Salmonella* might have been identified at family level only. Taxon classification is based on the extent of sequence identity in the 16S rRNA gene. For more closely related organisms, which have only undergone evolutionary diversification relatively recently, differences in one hypervariable region of the 16S rRNA gene sequence will be few, making unambiguous distinction difficult. Confident identification at higher taxonomic resolution can therefore not always be achieved, further limiting the usefulness of the method for direct pathogen diagnostics. Recent research suggests that sequencing of the entire 16S rRNA gene, on the other hand, can provide strain-level taxonomic resolution[207].

Due to the aforementioned issues with using 16S rRNA gene sequencing for diagnostics directly, the focus of the study was shifted towards identifying taxonomic

and metabolic biomarkers for indirect diagnostics. Previous studies reported universal rather than pathogen-specific gastroenteritis-induced dysbiosis[191,208], i.e. microbiota changes could not be used for distinction between different aetiological agents. Indeed, many of the significant differences in abundance of taxa and concentrations of metabolites found in this study differentiated all the infection from the control samples rather than pathogens from one another. Exceptions potentially allowing distinction between bacterial and protozoal gastroenteritis were the Firmicute *Blautia*, phenylacetate, the fatty acids butyrate, isovalerate, caprate and caprylate and the purine derivative hypoxanthine. Singh et al.[191] observed a similar decrease in *Blautia* in patients suffering from bacterial infection compared to healthy family members. *Blautia*, like other Lachnospiraceae, metabolise dietary polysaccharides that humans cannot degrade themselves[209], thus contributing to the symbiotic microbiota-host relationship. The decreased abundance of this genus during bacterial but not protozoal infection suggests that microbial activity beneficial to the host is reduced to varying extents by different types of infection. Butyrate is essential for differentiation of regulatory T-cells[210], which have an immunosuppressive function. Increased levels of butyrate during protozoal infection could therefore be part of the pathogen's immune system evasion strategy. Ng Hublin et al.[211] also found increased hypoxanthine concentrations in mice infected with *C. parvum*. Parasitic protozoa are incapable of synthesising purine nucleotides de novo and thus depend on salvage pathways from compounds such as hypoxanthine[212]. Three metabolites were identified that might allow distinction between pathogens from the same domain of life: Aspartate levels were increased during salmonellosis but not campylobacteriosis, serine levels were higher in *Salmonella*- than *Campylobacter*-positive samples and 3-phenylpropionate concentrations were elevated during giardiasis compared to cryptosporidiosis.

Although some of these potential biomarkers appear promising, the abundances/concentrations still varied widely between individuals suffering from the same infection making it impossible to define a reference range that did not overlap with another sample group. On top of this, despite filtering out very low abundance taxa, some taxa included in the analyses were not detected at all in a large number of samples. *Morganella*, for example, found to be differentially abundant between

cryptosporidiosis and giardiasis samples, would not be a suitable biomarker for distinction of the two pathogens as the genus was only present at detectable levels in 25% and 10.23% of the samples, respectively. The high variation between infection profiles might indicate that different strains of the same pathogen, which could not be distinguished by the technology used, have different effects on the microbiota. The prediction accuracies of 81.61% for the type of infection and 69% for the aetiological agent when classifying samples with a Random Forest model show that the problems imposed by this variation can partly be overcome using a range of markers instead of individual taxa/metabolites. Interestingly, the genus *Blautia* was the only feature suggested as an individual biomarker that was also of high importance for accurate prediction during Random Forest classification. This indicates that the interplay between several features is more characteristic to disease aetiology than the presence or absence of specific ones.

2.6.2.1 Study limitations

Several flaws of the present study have been identified, which have the potential to negatively affect the significance of the findings. Stool samples were stored at 4°C for a varying number of days before freezing. At this temperature, metabolism slows down but does not cease completely, leading to alterations in both taxonomic and metabolic profiles[213,214]. In line with this, prolonged storage at 4°C decreased bacterial α-diversity and increased the concentrations of many metabolites. Furthermore, previous studies found that differences in DNA extraction protocols introduce bias with regards to taxonomic profiles[215,216]. Although the same extraction protocol was used for all samples, processing was carried out over a time period of four years and by three people in total so that batch variation between extraction kits and operator bias cannot be excluded entirely. On top of this, sample batches were sequenced by different external providers and while the hypervariable region targeted and the sequencing platform used were identical for all three, thus eliminating variation introduced by these factors[217,218], the exact protocols of sample handling and processing after send out could not be controlled. During differential abundance testing, potential bias was accounted for by adding storage time at 4°C, year of DNA extraction, and sequencing centre to the analysis as covariates.

However, while the control samples obtained from healthy volunteers were frozen immediately upon receipt, the infection samples were stored at 4°C for up to five days before freezing. Most *Campylobacter*-positive samples were collected during the early stages of the study and sequenced at one centre while the controls were obtained towards the end of the study and sequenced at a different centre (cf. **Table 2.1**). Distinction of sample processing artefacts from true differences introduced by infection is therefore difficult and the chosen analysis method might conceal changes.

Apart from different pathogen strains, the wide range of both taxonomic and metabolic profiles observed for gastroenteritis samples could also be due to different reactions to colonisation by the pathogen based on pre-infection interindividual diversity of microbiota composition in the patients. Unfortunately the present study is not a longitudinal one, i.e. data is only available from the acute phase of infection but not from before and after. Additionally, no patient demographics or lifestyle information were recorded so that the effect of factors such as gender, age, comorbidities and diet[34,219-221] could not be accounted for.

Finally, while the effects of four different pathogens were studied overall, *Campylobacter*-positive samples dominated the data for bacterial and *G. lamblia*-positive samples the data for protozoal infection. Increasing the number of controls, salmonellosis and cryptosporidiosis samples to match those of the campylobacteriosis and giardiasis samples might potentially lead to the discovery of more pathogen-specific biomarkers and increase prediction accuracy of the Random Forest classifier.

2.7 Conclusions

Although this study shows that, without modifications to the protocol, 16S rRNA gene sequencing is not suitable for direct pathogen identification - at least in an environment as complex as stool – and that the use of individual pathogen-specific biomarkers for diagnostics is unlikely to be feasible, prediction of aetiological agent with the help of several taxonomic and metabolic features yielded more promising results. Further investigation of interconnected changes in the gut microbiota during gastroenteritis with a larger number of samples, encompassing additional pathogens

and adjusting for patient demographics and lifestyle could lead to the development of alternative, culture-free diagnostic tests targeting changes in the ecosystem induced by the pathogen instead of the pathogen directly. Increasing the proportion of specific diagnoses would contribute to a better understanding of the prevalence of certain aetiological agents, which can in turn inform public health interventions aimed at decreasing the burden of gastroenteritis.

However, if successful, one major issue remains: A culture-free diagnostic approach in this form does not yield a pathogen isolate for further characterisation, which is crucial for antimicrobial susceptibility testing and outbreak investigations. How high-throughput sequencing technologies could be exploited to circumvent this problem whilst still eliminating the need for culture, will be explored in the following chapter.

CHAPTER THREE

---

3. Culture-independent, metagenome-based detection and characterisation of non-typhoidal *Salmonella enterica*

3.1 Summary

Currently, initial confirmation of infection caused by a gastrointestinal pathogen and further characterisation of the detected pathogen are carried out by separate laboratories and involve multiple, culture-dependent steps. Data obtained from a single metagenomic sequencing reaction could allow both detection and characterisation of enteropathogens simultaneously, thus decreasing diagnostic turnaround times and leading to a faster availability of surveillance data. This chapter describes metagenomic sequencing of twenty stool samples from patients diagnosed with salmonellosis and comparison of the resulting data with the corresponding isolate whole genome sequencing (WGS) data generated during routine surveillance by Public Health England (PHE). Taxonomic classification of metagenomic sequencing reads identified *Salmonella enterica* in 70% of the samples. Coverage of assembled *S. enterica* metagenomes was insufficient to establish phylogenetic relatedness with their isolate genome counterparts or to determine antimicrobial resistance (AMR) profiles. This shows that, without modifications leading to an enrichment of the targeted pathogen, the protocol is not suitable to provide the resolution achieved by current surveillance methods.

3.2 Introduction

3.2.1 Salmonellosis

After *Campylobacter*, *Salmonellae* are the second most common bacterial cause of foodborne infections in Europe[222]. Salmonellosis generally manifests in the form of diarrhoea and vomiting, which is usually self-limiting and does not require treatment, or, in the case of a systemic infection, as a febrile disease. The genus *Salmonella* comprises two species, *S. bongori* and *S. enterica*. The latter can be divided into the six subspecies *enterica*, *salamae*, *arizonae*, *diarizonae*, *houtenae* and *indica* [223]. Based on specific agglutination reactions, over 2,500 *S. enterica* serovars have been identified [224].

For *S. enterica* subsp. *enterica*, responsible for 99% of human and animal infections, a further distinction is made between typhoidal and non-typhoidal *Salmonellae* (NTS). While typhoidal *Salmonellae* are host-restricted[225], NTS tend to be either generalist or host-adapted[226] and can be transferred from animals to humans causing zoonotic infections. As such, they fall under the realm of the World Health Organisation's One Health Approach. Traditionally, only the typhoidal serovars *S*. Typhi and Paratyphi were thought to be the causative agents of enteric fever[227]. However, in recent years, invasive NTS strains, capable of colonising sites outside the gastrointestinal (GI) tract, have been observed in low-income settings[228,229]. In fact, invasive NTS are among the most common causes of bloodstream infections in parts of Africa. Here, prevalence is particularly high in children and human immunodeficiency virus-infected adults[228]. Worldwide, NTS cause an estimated 93.8 million cases of gastroenteritis annually, resulting in 155,000 deaths[230]. In the UK, they are the third most common cause of bacterial gastroenteritis[162]. About 3.4 million cases and 680,000 deaths are thought to be attributable to invasive NTS infections[231] compared to about 27 million cases of typhoid fever and 200,000 associated fatalities[232]. In the UK in 2014, international travel was linked to 26% of laboratory-confirmed NTS infections[233]. Given the high prevalence of systemic disease in Southeast Asia[225], these rates are much higher

when looking at typhoid fever: In 2017, more than 90% of *Salmonella* bloodstream infections referred to PHE were acquired abroad[234].

3.2.2 Metagenomic sequencing as a diagnostic tool

Following initial diagnosis, salmonellosis cases in the UK are investigated further by a reference laboratory for surveillance purposes. Surveillance allows compilation of case numbers from across the country to evaluate the pathogen's prevalence in the population but is not limited to mere confirmation of infection. Other tasks of the reference laboratory require a detailed characterisation of the organism: establishing phylogenetic relationships between isolates aids detection and tracing of outbreaks, and generation of AMR profiles furthers the understanding of resistance dissemination (discussed further in Chapter 4). Rapid diagnostic approaches generally do not provide this level of detail, thus laboratories need to perform additional work to isolate the pathogen prior to referral. A diagnostic tool, which omits the culturing step but allows both detection and characterisation of the pathogen, would reduce the burden on laboratories and ensure faster availability of surveillance data.

Metagenomic sequencing of stool samples from patients with acute gastroenteritis holds promise for becoming this tool as it can be used for identification of enteropathogenic taxa and, after metagenome assembly, provide insights into functional characteristics of the detected pathogens. Initial research into the use of metagenomic sequencing for diagnostic purposes focussed on infections of the nervous system. Over the years, a range of pathogens, including viruses[235-239], fungi[239], bacteria[240] and parasites[241], were successfully identified from cerebrospinal fluid (CSF) and brain biopsies using a metagenomic approach. Based on this research, an assay for the diagnosis of meningitis and encephalitis of unknown aetiology from CSF was validated and is now in use at the clinical laboratories of the University of California San Francisco[242,243].

Further research into metagenomic diagnostics mainly targeted respiratory, ocular and bloodstream infections: Graf et al.[244] demonstrated good agreement between untargeted metagenomics and a commercial respiratory virus PCR panel. In the case of eye infections, the scope of diagnostic testing is restricted by the limited volume of

ocular fluid that can be collected at a time. This issue could be overcome by an untargeted sequencing approach, as demonstrated by Doan et al.[245,246] who were able to detect fungal, parasitic and viral pathogens in low-volume samples from uveitis patients. During sepsis, rapid, targeted treatment is crucial for a positive patient outcome[247] but diagnostic blood culture has a high turnaround time and might yield negative results due to initial therapy with broad-spectrum antimicrobials. Several studies suggest that metagenomic sequencing could be used for detection of circulating pathogens and even cell-free DNA from non-circulating microorganisms when specimens are culture-negative, and additionally provide information about AMR profiles to prevent treatment failures[248,249].

3.2.3 Metagenome-based diagnostics of gastrointestinal infections

Most of the research into metagenome-based diagnostics has been conducted on tissues that are generally thought to be sterile, like blood and those derived from the central nervous system, or harbour microbiotas of relatively low complexity, like specimens from the respiratory tract[250]. Only a few studies have investigated the potential of metagenomic sequencing for detection and characterisation of infections in the more complex environment of faecal material where sequences from the pathogen could be outcompeted by those from commensals resident in the GI tract. Joensen et al.[87] performed metagenomic sequencing on 38 stool samples which tested positive for *Clostridioides difficile*, *Salmonella*, *Campylobacter*, *Yersina enterocolitica* or diarrhoeagenic *Escherichia coli* by conventional diagnostics. The same pathogen was detected in 89.47% of samples by the sequencing-based approach. Additionally, putative aetiological agents could be identified in diarrhoea samples that were conventionally negative. Zhou et al.[251] were able to detect *C. difficile* in 86.3% of samples from patients with confirmed *C. difficile* infection and retrospective metagenomic analysis of eight *Campylobacter*-positive faecal samples by Andersen et al.[252] identified *Campylobacter* in all samples.

Even fewer studies do not merely focus on detection of enteropathogens but aim for a more in-depth characterisation: Gigliucci et al.[165] were able to demonstrate a match between virulence factors found in the metagenome and the whole genome sequences of Shiga toxin-producing *Escherichia coli* (STEC) isolates from an

outbreak in Italy. Providing higher resolution still, Huang et al.[169] successfully used metagenomic sequencing to subtype *Salmonella* from two foodborne outbreaks in the United States to serovar level and showed that the outbreaks were unrelated by phylogenetic analysis of assembled metagenomes, which was in agreement with the results of culture-dependent methods. For culture-independent sequencing to be considered as a replacement for the current system of *Salmonella* surveillance in the UK, the approach would need to reliably provide information equivalent to that derived from WGS data at PHE.

3.3 Objectives

The purpose of the work presented in this chapter was to compare the resolution of WGS performed on *S. enterica* isolates to that of metagenomic sequencing performed on the corresponding stool samples. It was hypothesised that *S. enterica* can be identified and placed in a phylogenetic context to other infection isolates in a culture-independent manner and that AMR profiles derived from assembled metagenomes match those determined by WGS.

3.4 Materials and Methods

Routine diagnostics were carried out by staff at the Norfolk and Norwich University Hospital's (NNUH) enteric laboratory. Processing of bacterial isolates and bioinformatics were carried out by staff at PHE's Gastrointestinal Bacteria Reference Unit. Removal of host DNA and co-assembly generation were performed by Dr. Andrea Telatin at Quadram Institute Bioscience (QIB). Unless otherwise stated, default software parameters were used.

3.4.1 Collection and processing of faecal samples

Twenty *S. enterica*-positive stool samples, for which pathogen isolates were sent to PHE for WGS-based surveillance, were collected as part of the study described in Chapter 2. Total genomic DNA was extracted from the samples as outlined in section 2.4.1.3.

3.4.2 Processing of *Salmonella enterica* isolates

3.4.2.1 Genomic DNA extraction

S. enterica cultures were inoculated in 750 µl of nutrient broth immediately after receipt at PHE and incubated at 37°C overnight. Automated genomic DNA extraction from overnight cultures was carried out using a QIAsymphony instrument (Qiagen, Hilden, Germany). DNA was quantified with a GloMax® microplate reader (Promega, Fitchburg, WI, USA).

3.4.2.2 Whole genome sequencing

Extracted genomic DNA was submitted to the Genomic Services and Development Unit at PHE for automated preparation of sequencing libraries. Concentrations were normalised with the help of Biomek liquid handling robots (Beckman Coulter, Brea, CA, USA). Libraries were generated with a Sciclone® G3 workstation (PerkinElmer, Waltham, MA, USA) using the Nextera™ XT DNA sample preparation kit (Illumina®, San Diego, CA, USA). Sizing of resulting DNA fragments was carried out using a

LabChip® GX nucleic acid analyser (PerkinElmer) and library concentrations were assessed with a ViiA7 real-time PCR thermocycler (Thermo Fisher Scientific, Waltham, MA, USA) using the KAPA library quantification kit (Roche, Basel, Switzerland).

Clusters of DNA templates were generated using the cBot system (Illumina®). Short-read sequence fragments of 100 bp were produced by paired-end sequencing on a HiSeq 2500 platform (Illumina®) in fast mode. FASTQ sequences were deposited in the National Centre for Biotechnology Information's (NCBI) Sequence Read Archive under the BioProject PRJNA315192. Accession numbers (ANs) can be found in **Table 3.1**.

Table 3.1: Accession numbers of *S. enterica* whole genome sequences and serovars determined by Public Health England.

	Isolate genome accession number	Serovar
S01	SRR5216415	*S*. Enteritidis
S02	SRR5193498	*S*. Enteritidis
S03	SRR5219518	*S*. Enteritidis
S04	SRR6920479	*S*. Enteritidis
S05	SRR5193510	*S*. Enteritidis
S06	SRR5216200	*S*. Enteritidis
S07	SRR6920142	*S*. Enteritidis
S08	SRR6900889	*S*. Enteritidis
S09	SRR7251115	*S*. Agona
S10	SRR7590102	*S*. Agona
S11	SRR7457907	*S*. Typhimurium
S12	SRR6191122	*S*. Infantis
S13	SRR5583818	*S*. Typhimurium
S14	SRR7291831	*S*. Typhimurium
S15	SRR7402418	*S*. Saintpaul
S16	SRR7458790	*S*. Java
S17	SRR7292878	*S*. Typhimurium
S18	SRR7351386	*S*. Poona
S19	SRR5583132	*S. enterica* subsp. *diarizonae*
S20	SRR7471084	*S*. Mikawasima

3.4.2.3 Quality trimming of sequencing reads

Demultiplexing of samples was carried out with the Illumina CASAVA software, version 1.8. Trimmomatic[253] v.0.36 was used to remove any bases with a Phred

quality score below 30 from both ends of the sequencing reads. Reads less than 50 bp in length after trimming were discarded along with their paired read.

3.4.2.4 Initial species identification

Initial taxonomic classification of isolates to species and subspecies level was performed using the kmerID pipeline developed by PHE (https://github.com/phe-bioinformatics/kmerid)[254]. Sequencing k-mers were compared against the k-mers of 1,769 reference genomes, representing 59 pathogenic genera, obtained from RefSeq, the NCBI reference sequence database.

3.4.2.5 Serovar prediction

Salmonella sequence type and eBurst group[255] were determined using MOST, a pipeline for identifying multilocus sequence types (MLSTs) from short-read sequencing data[256]. Based on the correlation between sequence types and serovars described by Achtman et al.[255], serovar designations were inferred from sequence types according to the PHE database containing matched MLST and serovar data for >12,000 *Salmonella* isolates[254]. Predicted serovars for the twenty *S. enterica* isolates are shown in **Table 3.1**.

3.4.3 Metagenomics

3.4.3.1 Shotgun sequencing

Extracted DNA was quantified using a Qubit fluorometer with broad-range reagents (Thermo Fisher Scientific, Waltham, MA, USA) and sent to Novogene (Beijing, China) for further processing. Sequencing libraries were generated with the NEBNext® Ultra™ DNA Library Prep Kit for Illumina® (New England Biolabs, Ipswich, MA, USA). Libraries were purified using the AMPure XP system (Beckman Coulter, Brea, CA, USA), analysed for size distribution using a 2100 Bioanalyser instrument (Agilent Technologies, Santa Clara, CA, USA) and quantified by real-time PCR. Cluster generation was performed on a cBot System (Illumina®, San Diego, CA, USA). Paired-end 150 bp sequencing was performed on a HiSeq 2500 instrument (Illumina®).

3.4.3.2 Trimming and quality filtering of sequencing reads

Quality trimming and filtering of sequencing reads was carried out by the sequencing provider. Reads were demultiplexed and adapters as well as index sequences were removed. Reads containing >10% of ambiguous base calls as well as those for which >50% of total base calls had a quality score ≤5 were filtered out.

Contaminant host DNA was removed by alignment of the remaining reads to a human reference genome using BBMap[257] v. 38.08. Only the resulting high-quality, microbial sequences were included in subsequent analyses.

3.4.3.3 Taxonomic analysis

Taxonomic analysis was carried out from the resulting FASTA files for each sample with two different taxonomic classifiers: MetaPhlAn2[258,259] v. 2.7.8 was used for analysis of forward reads only. The tool infers the presence and read coverage of taxa based on a set of approximately 1 million clade-specific markers from >7,500 species, yielding relative abundance tables.

Paired-end analysis was performed with Kraken v. 1.0 and the default Kraken database[260]. The database, built from completed microbial genomes in the NCBI RefSeq database, contains a record of k-mers (k=31) and the lowest common ancestor (LCA) of organisms whose genomes share this k-mer. The k-mers contained within a sequencing read are queried against the database and a taxonomic label is assigned based on the determined set of LCA taxa. The resulting output files were converted to sample report files providing read counts associated with each taxon using the kraken-report command. Relative abundances were determined by dividing the read counts associated with specific taxa by the total number of reads assigned to the root for this sample.

3.4.3.4 Co-assembly and metagenome binning

A metagenome co-assembly was generated from the sequencing reads of all twenty samples. Forward sequencing reads from individual samples were aligned against the co-assembly using BBMap[257] v. 38.08. The resulting BAM files were sorted with SAMtools[261] v. 0.1.19. Metagenomic binning was carried out using MetaBAT 2[262] v.

2.12.1: A depth file was produced from the previously generated BAM files and binning of the co-assembly contigs was performed with this depth file. Taxonomic assignment of the resulting 179 bins was carried out by aligning contigs against the NCBI nt database with the blob_annotate_mod.pl command, modified from the blobology package[263] for implementation with gbtools[264] v. 2.6.0. For each sample individually, forward and reverse metagenomic sequencing reads were mapped against the bin containing contigs annotated as *S. enterica* and the mapped reads were assembled with metaSPAdes[265] v. 3.11.1.

3.4.4 Phylogenetic analysis

Phylogenetic relationships between reads from both shotgun and isolate whole genome sequencing were established with the help of StrainSifter[266] based on Snakemake v. 5.4.2 using *S*. Typhimurium str. LT2 (AN: AE006468) as reference genome, which is also used as a reference genome in PHE's WGS analysis pipeline. The minimum mapping quality score was set to 60, the maximum number of allowed mismatches to 5 and the minimum read depth to 5. The minimum fraction of bases covered at minimum read depth for a sample to be included in the analysis was fixed at 0.5 and the minimum frequency of a nucleotide to call a base at any position to 0.8. The resulting tree file was visualised using iTOL (Interactive Tree of Life)[267] v. 5.

S. enterica metagenomes assembled as described in section 3.4.3.4 and WGS reads from the corresponding isolates were uploaded to PATRIC v. 3.6.2[268] along with the aforementioned *S*. Typhimurium reference genome and two further reference genomes used in PHE's WGS analysis pipeline, *S*. Agona str. SL483 (AN: CP001138) and *S*. Enteritidis str. P125109 (AN: AM933172). All uploaded genomes were annotated with PATRIC's genome annotation service using the RAST tool kit[269]. Protein family predictions were based on the *Salmonella enterica* taxon (Taxonomy ID: 28901). A phylogenetic tree was built with the Codon Tree method of PATRIC's Phylogenetic Tree Building Service, which uses RAxML[270] to analyse proteins and coding DNA aligned to PATRIC cross-genus families. The number of genes to analyse was set to 1,000 and the maximum allowed number of deletions and duplications within a homology group to 10.

3.4.5 Antimicrobial resistance profiling

3.4.5.1 Detection of antimicrobial resistance determinants in isolate whole genome sequences

Antimicrobial resistance determinants (ARDs) in isolate WGS were identified with the help of the 'Genefinder' algorithm developed at PHE. Sequencing reads were mapped to a set of reference sequences using Bowtie 2, version 2.2.5[271], followed by generation of an mpileup file with SAMtools, version 0.1.18[261,272]. Positive matches between the read set and the reference sequence or nucleotide variations were defined as query coverage 100%, base-call variation >85% and nucleotide identity >90%. The reference database used included acquired genes and mutations known to confer resistance to β-lactams (including penicillins, 2nd-, 3rd- and 4th-generation cephalosporins and carbapenems), phenicols, aminoglycosides, sulphonamides, tetracyclines, trimethoprim and fluoroquinolones[273,274]. Variants of β-lactamase genes were identified with 100% identity based on reference sequences downloaded from the Lahey (www.lahey.org) or NCBI β-lactamase data resources (https://www.ncbi.nlm.nih.gov/pathogens/beta-lactamase-data-resources). Further reference sequences for acquired resistance genes were obtained from CARD (The Comprehensive Antibiotic Resistance Database) (http://arpcard.mcmaster.ca) and the ResFinder datasets (https://cge.cbs.dtu.dk/services/data.php). Chromosomal mutations were limited to previously published variations within the quinolone resistance-determining region of *gyrA* and *parC*.

3.4.5.2 Resistome profiling of metagenomic sequencing reads

Resistome profiling of forward shotgun sequencing reads from each sample was performed using GROOT (Graphing Resistance out of Metagenomes)[275] v. 0.8.3 with the pre-clustered CARD database. Database clusters are collections of sequences sharing high nucleotide identity, from which variation graphs were produced. The graphs were indexed with the node window length set to 150. Sequencing reads were then aligned against the indexed variation graphs. Reports were generated from the resulting BAM files, including only those ARDs covered at ≥97%.

3.4.5.3 Detection of antimicrobial resistance determinants in assembled (meta)genomes

Isolate genomes and assembled metagenomes were annotated with PATRIC as described in section 3.4.4. Lists of ARDs detected in the genomes were derived from the antibiotic resistance specialty genes identified during annotation.

3.5 Results

3.5.1 Recovery of *Salmonella enterica* sequences

Read counts per sample ranged from 33.39 to 42.34 million with a median of 37.83 million. The Q30 value exceeded 91% for all samples. Host contamination was minimal (<7%), except for samples 10 (43%), 11 (45%) and 20 (27%). To assess whether culture-independent metagenomic sequencing can be used for direct detection of *S. enterica*, sequencing reads were analysed with two taxonomic classifiers. A sample was classed as positive when *S. enterica* was detected at a relative abundance >0.01%. Using this cut-off, *S. enterica* was identified in fourteen samples (70%) using Kraken and in twelve samples (60%) using MetaPhlAn2 (**Table 3.2**). Relative abundances for the positive samples ranged widely, from 0.02% to

Table 3.2: Metagenomics-based detection of *S. enterica* sequences by two taxonomic classifiers in stool samples testing positive for *Salmonella* during routine diagnostics.

	Relative abundance of *Salmonella enterica* (%)	
	Kraken	MetaPhlAn2
S01	0.30	0.15
S02	0	0
S03	0	0
S04	0.02	0
S05	0	0
S06	0.90	0.81
S07	0.10	0.05
S08	0	0
S09	0	0
S10	0.02	0
S11	0.80	0.55
S12	0	0
S13	0.10	0.05
S14	0.50	0.11
S15	1.30	0.31
S16	28.50	28.33
S17	7.50	4.92
S18	2.70	2.87
S19	7.70	4.79
S20	11.20	9.98

28.5% for Kraken and from 0.05% to 28.33% for MetaPhlAn2. The two additional samples identified as *S. enterica*-positive using Kraken only contained the pathogen at a low abundance of 0.02%. For all other samples, both classifiers were in good agreement, with abundance estimates differing by maximally 2.91%.

As described in the previous chapter, stool consistency and time since onset of symptoms can affect the likelihood of successful pathogen recovery. In the present study, detection of *S. enterica* was more likely in loose stools (**Figure 3.1A**). Five samples were derived from hospital inpatients, for which the time of onset of symptoms had not been recorded. For the remaining patients, there did not appear to be an obvious relationship between the duration of disease before sample collection and the relative abundance of *S. enterica* (**Figure 3.1B**) as detection was unsuccessful both for some patients with a recent onset of disease and those who had been ill for more than a month. The highest abundance detected was associated with a disease duration of eleven days followed by seven and 27 days.

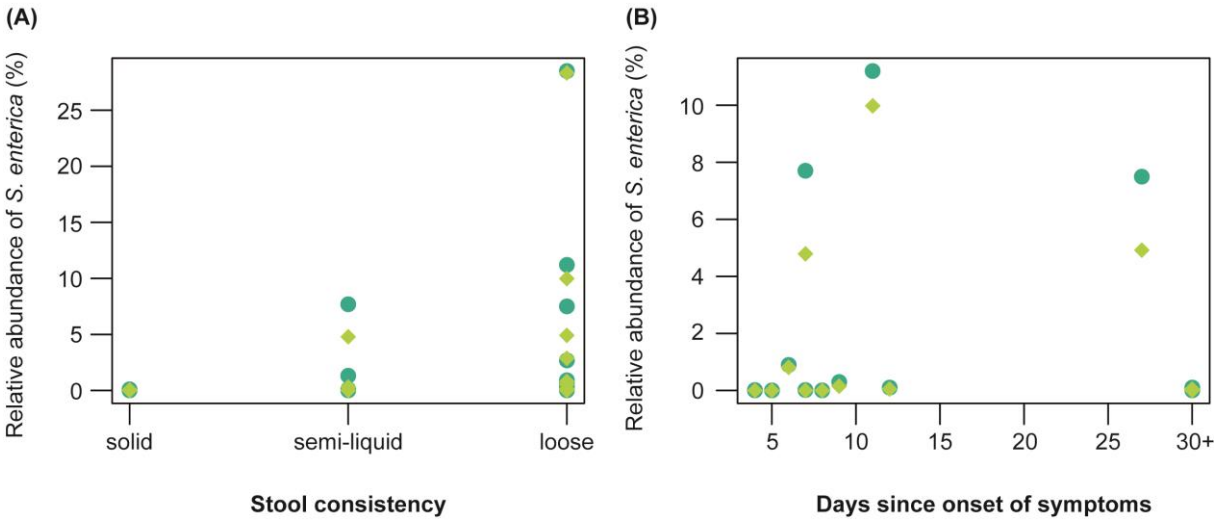

Figure 3.1: Effect of stool consistency and timepoint of sample collection on recovery of *S. enterica* sequences. Plots show the relative abundances of *S. enterica* in twenty stool samples determined using the Kraken (turquoise) and MetaPhlAn2 (green) taxonomic classifiers against (A) different stool consistencies and (B) the days that have passed between onset of symptoms and faecal sample collection for diagnostic purposes. Patients, for which the days since onset of symptoms are designated '30+', have been experiencing symptoms for more than a month. Hospital inpatients, for which the date of symptom onset was not recorded on the referral form, were omitted from the second plot.

3.5.1.1 Gut microbiota profiles during salmonellosis

In order to investigate whether metagenomic sequencing is able to identify any common features in gut microbiota composition during S. enterica infection, taxonomic profiles of the twenty samples, as determined using MetaPhlAn2, were compared (**Figure 3.2**). Sample 10 was found to contain relatively high levels of Retroviridae while sample 11 was characterised by a high abundance of

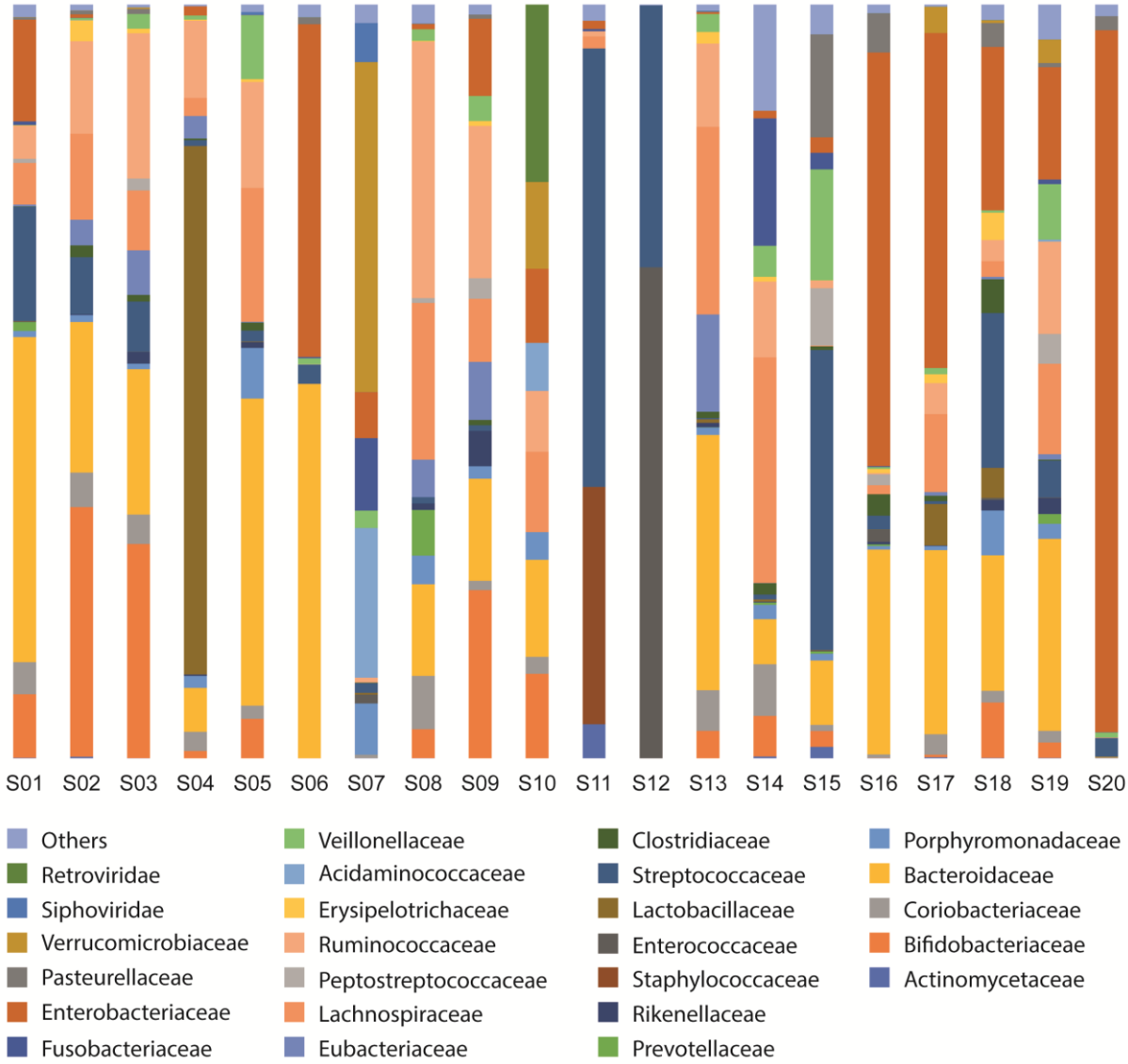

Figure 3.2: Gut microbiota composition during *S. enterica* infection. Relative abundances of gut microbiota taxa at the family level in twenty samples testing positive for *S. enterica* during routine diagnostics, determined by metagenomic sequencing and classified using MetaPhlAn2.

Streptococcaceae and Staphylococcaceae, and sample 12 only contained sequences classified as Enterococcaceae and Streptococcaceae. At such abundances, these features are unusual for the human gut microbiota and suggest either poor DNA quality, sample contamination or misclassification of sequencing reads. Noticeably, even for samples with few S. enterica reads, Enterobacteriaceae were found to be abundant. At species-level taxonomic resolution, most of the Enterobacteriaceae reads were assigned to E. coli with a relative abundance of the organism of up to 62.57% in sample 20.

3.5.2 Genome assemblies

Although further characterisation of detected *S. enterica* was attempted directly from the shotgun sequencing reads as described in the following sections, it was deemed likely that the signal from the remaining microorganisms in the gut microbiota might outcompete the signal of interest, especially in samples where the pathogen was only detected at a low abundance. To amplify the *S. enterica* signal, a co-assembly was generated from the sequencing reads of all twenty samples, followed by metagenome binning and taxonomic annotation of the resulting bins. To specifically filter out *S. enterica* reads for each patient, the sample's sequencing reads were then individually aligned to those forming the *S. enterica* bin. The mapped reads were used to assemble individual metagenomes. A comparison of assembly statistics for metagenomes and WGS-derived isolate genomes revealed that the metagenomes were of a much poorer quality (**Table 3.3**).

Contig counts ranged from 81-447 (median: 149.5) for isolate genomes and from 13-11,982 (median: 3,038) for metagenomes. Generally, a lower number of contigs indicates that more reads overlap and can be assembled into a continuous consensus sequence, thus hinting at an almost complete genome. However, while all isolate genomes were classed as 100% complete based on the percentage of universal functional roles, a lower number of contigs was not reflected in more complete metagenomes. Instead, low contig numbers appeared to be associated with a lack of *S. enterica*-specific reads. The most incomplete metagenome was obtained from sample 12 (0%) and the most complete from sample 20 (68.20%). Two additional indicators of genome quality, the N50 and the fine consistency values,

further highlighted that the assembled metagenomes were of lower quality than the isolate genomes. The N50 value describes the contig size, at which 50% of the genome are contained in contigs of equal or larger size. The higher the number, the closer it is to the size of the complete genome, indicating a good-quality assembly. N50 ranged from 28,913 to 145,067 bp (median: 75,457.50 bp) for isolate and from 245 to 5,546 bp (median: 545 bp) for metagenomes. Coarse consistency describes the number of functional roles in an annotated genome, presence or absence of which was predicted correctly based on the expected roles occurring in *S. enterica* reference genomes. Coarse consistency equalled or exceeded 99.5% for all isolate genomes but ranged from 76.30% to 96.30% for metagenomes.

Table 3.3: Assembly statistics of *S. enterica* isolate and metagenomes.

	Contig count		N50 (bp)		Completeness (100%)		Coarse consistency (%)	
	Isolate genome	Metagenome	Isolate genome	Metagenome	Isolate genome	Metagenome	Isolate genome	Metagenome
S01	160	6,038	79,256	681	100	64.30	99.60	77.20
S02	142	1,489	74,057	350	100	40.20	99.60	80
S03	146	75	74,006	249	100	2.30	99.60	94.50
S04	138	3,730	74,354	419	100	52.50	99.60	77.50
S05	81	103	141,150	253	100	3.30	99.60	94.50
S06	97	3,011	145,067	5,162	100	67.60	99.60	79.70
S07	140	11,982	80,021	484	100	66.70	99.60	78.30
S08	202	2,198	46,404	502	100	49.50	99.60	78.40
S09	142	2,381	74,477	606	100	51.50	99.90	77.80
S10	145	86	74,477	259	100	1.80	99.90	94.60
S11	162	543	102,151	270	100	12.50	99.80	91.40
S12	163	13	79,113	245	100	0	99.70	96.30
S13	189	3,065	82,251	328	100	51.40	99.90	76.30
S14	89	5,523	134,816	776	100	65.50	99.60	78.50
S15	335	5,111	28,913	2,681	100	67.60	99.60	78.60
S16	394	11,045	55,595	918	100	67.30	99.90	78.60
S17	153	2,948	80,654	5,097	100	68.10	99.90	78.90
S18	447	5,947	69,269	3,172	100	67.90	99.50	78.50
S19	265	3,297	74,345	5,546	100	67.60	99.60	79.50
S20	144	7,180	76,438	1,641	100	68.20	99.90	78.70

Metagenome assemblies were generated using MetaSPAdes from sequences aligning to the *S. enterica* bin after co-assembly of metagenomic sequencing reads from all twenty samples. Assembly statistics were determined with the help of PATRIC.

3.5.3 Phylogeny of *Salmonella enterica* strains

Individual *S. enterica* isolates received by PHE are characterised at single nucleotide polymorphism (SNP)-level resolution as described in section 4.4.2.4. This level of resolution could not be achieved for the metagenomes due to the incompleteness of the assemblies. Instead, phylogenetic analysis with the available reads and contigs was carried out to investigate whether isolates clustered together with their corresponding metagenomes.

This analysis was first performed on WGS and shotgun sequencing reads using StrainSifter, which was developed specifically for metagenomes (**Figure 3.3A**). Only seven of the twenty metagenomic sequencing samples were included in the resulting tree, indicating that reads aligning to the reference *S. enterica* strain were not detected by the pipeline in the remaining samples. The seven samples that were retained were those for which the taxonomic classifiers had identified the highest relative abundances of *S. enterica* (cf. **Table 3.2**). Whole genome and metagenomic sequencing reads clustered together for sample 6 only. On the phylogenetic tree built from assembled isolate and metagenomes, the metagenomes of both sample 6 and 11 were located close to their WGS-derived counterparts (**Figure 3.3B**). Both approaches were able to detect the close phylogenetic relatedness of isolates designated as the same serovar (cf. **Table 3.1**) and placed isolate 19, which was identified as *S. enterica* subsp. *diarizonae*, at the furthest distance from the remaining samples, which all belonged to subspecies *enterica*.

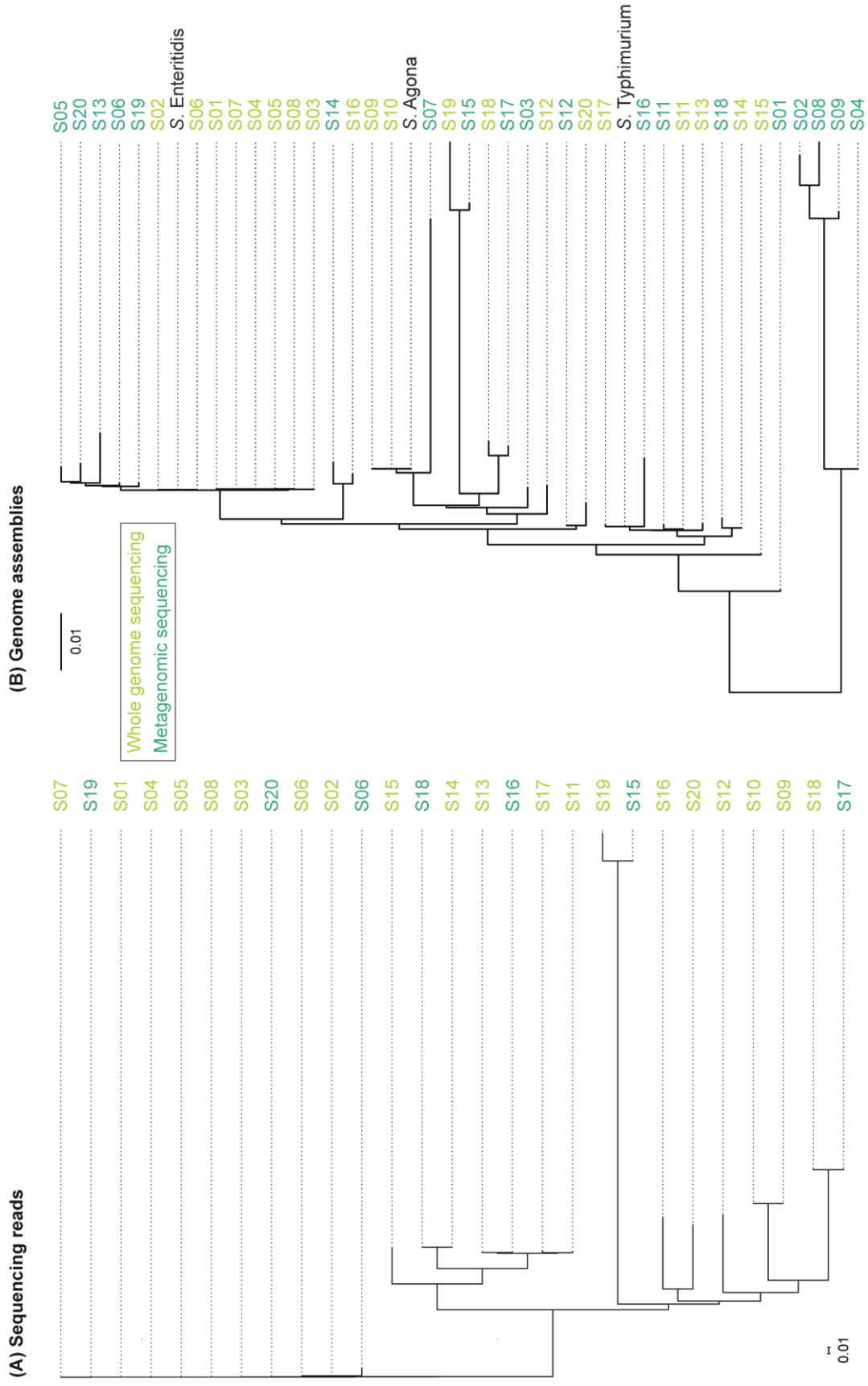

Figure 3.3: Phylogeny of *S. enterica* strains. Phylogenetic trees built from (A) WGS reads of *Salmonella enterica* isolates and shotgun sequencing reads derived from the corresponding stool samples using StrainSifter with *S.* Typhimurium str. LT2 as reference and (B) assembled isolate genomes and metagenomes using PATRIC's Phylogenetic Tree Building Service in relation to reference *S.* Agona, *S.* Enteritidis and *S.* Typhimurium genomes.

3.5.4 Antimicrobial resistance profiling in *Salmonella enterica*

The Genefinder algorithm forms part of PHE's routine WGS analysis pipeline for *S. enterica*, providing information about ARDs present in the isolate genomes. The algorithm identified an aminoglycoside acetyltransferase gene of the *aac(6')*-type, either *aac(6')-Iy* (n=14) or *aac(6')-Iaa* (n=6), in all twenty isolates included in this study (**Table 3.4**). However, these ARDs are generally silent and only become transcriptionally active in rare cases[276]. Only seven isolates carried further ARDs, namely *aadA*-type aminoglycoside adenylyltransferase genes, the aminoglycoside acetyltransferase gene *aac(3)-IVa*, *strA-strB*, the penicillinase gene $bla_{\text{TEM-1}}$, *sul* variants associated with sulphonamide resistance, the tetracycline efflux pump-encoding *tet*(A), the trimethoprim resistance gene *dfrA12*, *ermB* and *mphA* involved in resistance to macrolides, lincosamides and streptogramins, and mutations in *parC* and *gyrA*, which contribute to fluoroquinolone resistance. AMR profiles determined from assembled isolate genomes after genome annotation with PATRIC matched with few minor exceptions: *strB* but not *strA* was detected in sample 13 and *aadA3* and *aadA8b* could not be found in samples 13 and 14, respectively, while another *aadA* variant was identified successfully.

To evaluate whether direct profiling of metagenomic sequencing reads yields comparable results, GROOT was used for resistome analysis. Focussing only on ARDs detected in both the isolate genomes and metagenomic reads resulted in an exact match of AMR profiles for samples 16 and 18. In a further three samples, the pipeline detected *aac(6')-Ie-aph(2'')-Ia*, another aminoglycoside acetyltransferase gene, instead of *aac(6')-Iy*. Partially matching AMR profiles were obtained for samples 13 and 14, for which GROOT identified *strA*, *tet*(40), *sul-2* and *ermB* but none of the other ARDs present in the isolate genomes. However, since resistome profiling was carried out on the entire gut microbiota in this case it is uncertain whether the detected ARDs were carried by *S. enterica* or other organisms. To overcome this uncertainty, metagenomes assembled from sequences aligning to the *S. enterica* bin were annotated using PATRIC, and antibiotic resistance specialty genes were investigated. Only the *parC* mutations in samples 9 and 10 were found in both isolate and metagenomes.

Table 3.4: Comparison of whole genome- and metagenome-derived antimicrobial resistance profiles in *S. enterica*.

	Genefinder isolate genome	GROOT metagenomic reads	PATRIC isolate genome	PATRIC metagenome
S01	*aac(6')-Iy*	-	*aac(6')-Iy*	-
S02	*aac(6')-Iy*, *gyrA*	*aac(6')-Ie-aph(2'')-Ia*	*aac(6')-Iy*, *gyrA*	-
S03	*aac(6')-Iy*	*aac(6')-Ie-aph(2'')-Ia*	*aac(6')-Iy*	-
S04	*aac(6')-Iy*	-	*aac(6')-Iy*	-
S05	*aac(6')-Iy*	-	*aac(6')-Iy*	-
S06	*aac(6')-Iy*, *aadA1*, *sul-1*	-	*aac(6')-Iy*, *aadA*, *sul-1*	-
S07	*aac(6')-Iy*	*aac(6')-Ie-aph(2'')-Ia*	*aac(6')-Iy*	-
S08	*aac(6')-Iy*	-	*aac(6')-Iy*	-
S09	*aac(6')-Iy*, *gyrA*, *parC*	-	*aac(6')-Iy*, *gyrA*, *parC*	*parC*
S10	*aac(6')-Iy*, *gyrA*, *parC*	-	*aac(6')-Iy*, *gyrA*, *parC*	*parC*
S11	*aac(6')-Iaa*	-	*aac(6')-Iaa*	-
S12	*aac(6')-Iy*	-	*aac(6')-Iy*	-
S13	*bla* $_{\text{TEM-1}}$, *aac(6')-Iaa*, *aac(3)-IVa*, *aadA3*, *aadA22*, *strA-strB*, *tet* (A), *sul-1*, *sul-2*	*strA*, *tet* (40), *sul-2*	*bla* $_{\text{TEM-1}}$, *aac(6')-Iaa*, *strB*, *aac(3)-IV*, *aadA22*, *tet* (A), *sul-1*, *sul-2*	-
S14	*aac(6')-Iaa*, *aadA2*, *aadA8b*, *ermB*, *mphA*, *dfrA12*, *sul-1*	*ermB*	*aac(6')-Iaa*, *aadA2*, *ermB*, *mphA*, *dfrA12*, *sul-1*	-
S15	*aac(6')-Iy*	-	*aac(6')-Iy*	-
S16	*aac(6')-Iaa*, *gyrA*	*aac(6')-Iaa*, *gyrA*	*aac(6')-Iy*, *gyrA*	-
S17	*aac(6')-Iaa*	-	*aac(6')-Iaa*	-
S18	*aac(6')-Iaa*	*aac(6')-Iaa*	*aac(6')-Iy*	-
S19	*aac(6')-Iy*	-	*aac(6')-Iy*	-
S20	*aac(6')-Iy*	-	*aac(6')-Iy*	-

Genefinder was used for antimicrobial resistance (AMR) profiling of WGS-derived isolate genomes. Profiling with GROOT was performed on all metagenomic sequencing reads. AMR profiles of assembled isolate and metagenomes were derived from genome annotations generated using PATRIC.

3.6 Discussion

3.6.1 Direct detection and characterisation of *Salmonella enterica* by metagenomic sequencing of stool samples

Of the twenty salmonellosis cases used in this study, confirmed by routine diagnostics and WGS-based surveillance, only 60-70% - depending on the taxonomic classifier used – would have been classed as *S. enterica*-positive by metagenomic sequencing. This highlights that, without any modifications to the protocol, the approach would be prone to a high number of false negatives. Successful recovery of pathogen sequences can be influenced by a variety of factors, one of which is the amount of time that has passed between onset of gastroenteritis symptoms and sample collection for diagnostic purposes. Pathogen loads during viral gastroenteritis, for example, are thought to peak 24-48 hours after symptoms develop[277]. In the present study, no apparent relationship was found between time since disease onset and successful detection of *S. enterica* sequences. Stool samples were stored at 4°C for a varying number of days at the NNUH enteric laboratory before being transferred to QIB and frozen, which could have further effects on the detection of the target organism as described in chapter 2. The likelihood of recovering *S. enterica* sequences seemed to be higher for loose than solid stools. Vandeputte et al.[278] showed that variations in stool consistency were associated with differences in microbiota composition and suggested that shorter colonic transit times could increase the abundance of fast-growing species and organisms that adhere to host tissue, such as *Salmonella*. Conventional diagnostic methods like culture and PCR enrich for the pathogen of interest by using selective media or targeting and amplifying a genomic region specific to a particular pathogen. Metagenomic sequencing, on the other hand, is unbiased and heavily affected by the ratio of pathogen to commensal DNA. In samples classed as negative in this study, *S. enterica* might therefore be present at an abundance high enough to cause symptoms and be detected by selective methods but the background noise of the remaining gut microbiota, for example the high levels of *E. coli* observed in several patients, could make it impossible to pick up the low abundance signal during

unbiased sequencing at this depth. Since no longitudinal sampling was carried out, it is unclear whether high *E. coli* levels were also present in the gut microbiota of these patients prior to disease but an increase in, most likely commensal, members of this species has been detected previously during infection with *Salmonella*, *Campylobacter* and *Shigella*[169,191].

While detection of *S. enterica* by metagenomic sequencing was partially successful, further characterisation in the form of phylogenetic analysis and determination of AMR profiles failed using both sequencing reads directly and metagenome assemblies. This was still true when samples 10, 11 and 12, which showed high host contamination and/or unusual taxonomic profiles indicating poor quality, were disregarded. Only seven samples contained sufficient species-specific information to be retained in a phylogenetic tree built directly from the sequencing reads and the majority of samples were located at a distance from their isolate genome counterparts. Overall, this analysis indicates that the resolution obtained from metagenomic sequencing is not sufficiently high to examine phylogenetic relatedness between different strains of the same species, even when the species is present at high abundance as is the case for sample 16. Similarly, only few AMR profiles partly matched those derived from WGS. The apparent discrepancy between the ability to simply detect *S. enterica*, which was successful for more than half of the samples, and the failure to characterise the pathogen further can be explained by the different approaches employed by taxonomic classifiers and tools used to determine phylogeny and AMR profiles: the former make use of short taxon-specific marker sequences. The latter, on the other hand, require longer, uninterrupted consensus sequences for reliable results. The quality metrics obtained for the metagenome assemblies, however, indicate that large parts of the genome are not covered so that a lot of data will not pass the analysis tools' built-in quality control features. Rodriguez & Konstantinidis[279] estimated that datasets with >60% average coverage perform best for assemblies and SNP-level strain identification requires deep coverage of at least 10X[259], values not achieved in this study. Huang et al.[169], who successfully performed phylogenetic analysis on *S*. Heidelberg metagenomes, reported coverage ranging from 7.8X-120.4X.

3.6.2 Potential improvements to culture-free diagnostics and surveillance of salmonellosis

Although analysis at a level required for *S. enterica* surveillance was not achieved in the current study, changes to the protocol could likely improve results. The DNA extraction protocol was not developed specifically for metagenomic sequencing. Rather, the same extracts as used for 16S rRNA gene sequencing in the study described in chapter 2 were sequenced. Extensive bead beating can alter the apparent taxonomic composition of a sample[280] and host depletion was not carried out. The ratio of human to microbial DNA in stool is generally low, with up to 75% of the faecal mass estimated to be of bacterial origin[281]. In line with this, only three study samples showed significant human contamination and this could easily be removed bioinformatically. However, exclusively sequencing microbial DNA in the first place would increase the total number of reads derived from the target organism. An example of a DNA extraction protocol for metagenomic sequencing with a host depletion step is described in section 5.4.2.1.

Validation of a metagenomics-based diagnostic test for *S. enterica* infection would require establishment of a limit of detection. This could be achieved by spiking faecal samples with different known numbers of colony-forming units (CFU)/ml of the organism and investigating at which point the taxonomic classifier chosen cannot reliably pick up the pathogen signal anymore. Andersen et al.[252] performed this experiment for *Campylobacter* and found $7.75x10^4$ CFU/ml to be the lowest detectable spiking level. Furthermore, they showed that higher spiking levels were associated with higher relative abundances but this relationship was not proportional indicating that abundance data cannot easily be translated into the number of viable organisms present in a sample. Conducting a similar validation for *S. enterica* might reveal that bacterial loads were below the limit of detection for patients in which *S. enterica* sequences could not be recovered in this study. Limits of detection determined in this way will be specific for the pathogen and the DNA extraction protocol, library preparation method, sequencing platform and analysis pipeline used so they would have to be re-evaluated should any changes be made to the process.

To address the problem of insufficient coverage for detailed characterisation of the detected *S. enterica*, enrichment of the target organism prior to sequencing could be

carried out, thus decreasing the signal from other constituents of the gut microbiota. Selective capture has been described for viral sequences[282,283] and STEC using whole-genome RNA baits[284]. Development of a similar bait-capture system for *Salmonella* would increase coverage and sequencing depth of the pathogen genome and most likely allow for phylogenetic analysis and AMR profiling. However, it has to be kept in mind that a pathogen enrichment approach is only feasible when the target organism is known. While this eliminates one of the main benefits of choosing metagenomics for diagnostic purposes, its ability to identify a multitude of potential pathogens in an unbiased manner, initial pathogen identification by sequencing followed by capture-based enrichment and further characterisation of the resulting sequences might still be faster than culture-based enrichment followed by WGS.

3.7 Conclusions

This small-scale study of twenty stool samples from patients with acute gastroenteritis testing positive for *S. enterica* infection by conventional diagnostics illustrates that, while culture-free metagenomic sequencing cannot yet replace established diagnostic approaches and WGS-based surveillance of salmonellosis, there is scope for improvement of the methodology presented here. This might eventually allow reliable culture-independent detection and characterisation of *S. enterica* and lead to a decrease in current turnaround times from confirmation of infection to availability of surveillance data.

How high-throughput sequencing has already allowed reference laboratories to replace laborious wet-lab work with computational methods, albeit retaining the need for an initial culturing step, will be discussed in the following chapter.

CHAPTER FOUR

---

4. Antimicrobial resistance in UK isolates of non-typhoidal *Salmonella enterica* and the use of whole genome sequencing-derived resistance profiles for surveillance purposes

Parts of this chapter have previously been published in Neuert et al.[285].

4.1 Summary

The use of whole genome sequencing (WGS) for surveillance of enteric pathogens, such as *Salmonella*, has revolutionised the work of public health laboratories. It allows inference of a multitude of pathogen characteristics from a single sequencing experiment, including identification of species, serovar and even strain at single-base resolution. Additionally, it enables detection of functionally important genes, such as antimicrobial resistance determinants (ARDs). In this chapter, antimicrobial resistance (AMR) profiles of 3,491 non-typhoidal *Salmonella enterica* (NTS) isolates received by Public Health England's (PHE) Gastrointestinal Bacteria Reference Unit (GBRU) were characterised. Pansusceptibility was observed in 68.98% of isolates. Of the resistant isolates, 74.42% showed multidrug resistance (MDR). Co-resistance to ampicillin, streptomycin, sulphonamides and tetracycline, mostly mediated by carriage of $bla_{\text{TEM-1}}$, *strA-strB*, *sul2* and *tet*(A), was the most common MDR profile. Extendend spectrum β-lactamase (ESBL) genes were detected in 1.23% of isolates and associated with travel to North Africa, while multiple mutations in chromosomal genes underlying ciprofloxacin resistance were found in 2.35% of isolates and linked to travel to South and Southeast Asia. Prediction of phenotypic resistance based on ARDs identified in WGS resulted in complete agreement of phenotypic and genotypic profiles in 97.82% of cases. Only 0.17% of all possible isolate/antimicrobial combinations were discordant, with the largest number of mismatches associated with streptomycin resistance. This highlights the suitability of using WGS-based prediction of AMR profiles for NTS surveillance.

4.2 Introduction

4.2.1 History of antimicrobial resistance in *Salmonella enterica*

Without the use of antimicrobial agents to treat invasive and severe gastrointestinal (GI) cases of *S. enterica* infection, associated mortality rates would likely be even higher. Additionally, veterinary antimicrobial therapy has decreased the incidence of *Salmonella* in animal reservoirs, reducing the risk of zoonoses. However, these initially successful interventions have not come without a price: *S. enterica* rapidly developed resistance against antimicrobials used to control the organism in both humans and animals. Resistance dissemination was aided by the fact that many ARDs are located on mobile genetic elements and can therefore be transferred horizontally between organisms.

In humans, chloramphenicol was the drug initially recommended for the treatment of enteric fever caused by *S*. Typhi and Paratyphi. The first outbreaks resistant to chloramphenicol were reported in the early 1970s and by the 1990s strains with additional resistance to ampicillin, sulphonamides and trimethoprim were widespread[286]. The introduction of fluoroquinolones as an alternative to chloramphenicol in the 1980s was followed by a larger number of isolates displaying decreased susceptibility to ciprofloxacin[287,288]. Consequently, 3rd-generation cephalosporins and azithromycin, intended as antimicrobials of last resort when other treatments fail, are increasingly being used as first-line therapies for systemic infections[225]. Resistance to these drugs has been reported in European isolates and is often associated with travel to Asia[289].

Resistance in NTS is heavily linked to veterinary application of antimicrobials and subsequent zoonotic transfer to humans. The increased use of ampicillin, chloramphenicol, streptomycin, sulphonamides and tetracycline resulted in the emergence of *S*. Typhimurium strains displaying the so-called ACSSuT phenotype, resistance to exactly these antimicrobials, in the 1980s[290,291]. By the early 1990s, the phenotype had disseminated globally with some strains acquiring additional resistance to cephalosporins, aminoglycosides, trimethoprim and even fluoroquinolones, the antimicrobials initially introduced to circumvent the treatment

challenges imposed by the MDR organism[291]. More recently, the spread of an extensively drug-resistant strain of *S*. Kentucky, which originated in Egypt, has sparked concern. This strain is non-susceptible to extended-spectrum cephalosporins, cabapenems, most aminoglycosides, trimethoprim-sulfamethoxazole, fluoroquinolones and azithromycin[292]. Azithromycin resistance has also been reported in other NTS serovars[293,294]. By 2015, 29.3% of all NTS isolates in the EU were classed as MDR[295]. Resistance to colistin, considered the antimicrobial of last resort for the management of many MDR Gram-negative pathogens, was believed to be solely encoded chromosomally[296]. However, plasmid-mediated colistin resistance via *mcr* genes, first described in China in 2015[297], has since been detected in UK NTS isolates, most of which were associated with international travel[298].

4.2.2 Surveillance of antimicrobial resistance in *Salmonella enterica*

Surveillance of NTS infections by public health agencies is essential for monitoring transmission of the organism through the food chain to humans, and for establishing effective treatment guidance. Since infections caused by MDR NTS are associated with increased mortality and higher costs to the healthcare system[299,300], determination of AMR profiles is an integral part of this surveillance.

Traditionally, surveillance in reference laboratories has required a series of independent tests: species and subspecies identity were confirmed by biochemical tests and PCR[301,302] and agglutination reactions formed the basis of assigning isolates to serovars following the White-Kauffman-Le Minor scheme[224,303].

Despite international acceptance of the scheme, serotyping has its disadvantages: firstly, it cannot be automated and requires costly antibodies derived from rabbits. With the number of NTS isolates referred to PHE in the last years ranging from 7,250 in 2014 to 12,094 in 2007[304], this puts an enormous strain on the reference laboratory. Secondly, antigenic concordance does not always reflect the pathotype of the organism. The serological formulae of *S*. Paratyphi B and *S*. Java, for example, are identical but while the former causes typhoid fever, the latter is responsible for non-invasive gastroenteritis[305]. To overcome these issues, a sequencing-based approach for serotyping *S. enterica* has been developed: multilocis sequence typing

(MLST) relies on grouping isolates according to the sequence similarity of several housekeeping gene fragments[306]. Since first proposed, this scheme has been shown to correlate well with phenotypic serovar designations in many cases[255].

Advances in sequencing technology have allowed reference laboratories to investigate isolates in even greater detail[307]. At PHE's GBRU, traditional phenotypic serotyping was replaced by the routine implementation of WGS for identification and surveillance of *S. enterica* in April 2015[254]. Typing at single-nucleotide resolution allows inference of phylogenetic relationships between isolates and has proven invaluable for outbreak detection[308-312]. Ashton et al.[254] showed that serovar designations based on MLST derived from WGS data correlated with the results of phenotypic serotyping in 96% of cases, thus WGS even permits maintenance of serovar nomenclature for backwards compatibility and data exchange with countries that still rely on traditional serotyping methods. At PHE, most of these methods have been discontinued as routine services in favour of WGS-derived methodologies.

The gold standard for generation of AMR profiles in *S. enterica* is the determination of minimum inhibitory concentrations (MICs) and the classification into resistant and susceptible isolates based on clinical breakpoints as recommended by the European Committee on Antimicrobial Susceptibility Testing (EUCAST). Annotation of WGS, however, also allows the detection of ARDs, from which AMR profiles can be inferred, and therefore provides the means to rapidly monitor emerging trends in AMR patterns of NTS without the requirement for additional laboratory testing.

4.3 Objectives

The purpose of the work presented in this chapter was to characterise the prevalence of phenotypic AMR and genetic ARDs in a subset of UK NTS isolates referred to PHE's reference laboratory. It was hypothesised that the detection of ARDs could be used to infer AMR profiles from WGS data, providing evidence that genome-derived prediction can replace phenotypic antimicrobial susceptibility testing (AST).

4.4 Materials and Methods

Processing of bacterial isolates, bioinformatics and antimicrobial susceptibility testing were carried out by staff at PHE's GBRU. Further analysis of the resulting data was performed by the author.

4.4.1 Processing of bacterial isolates

Processing and sequencing of bacterial isolates was carried out as described in sections 3.4.2.1 and 3.4.2.2. All 7,009 isolates received by the reference laboratory between April 2014 and March 2015 were included in this study. WGS accession numbers can be found in supplementary table 1 of Neuert et al.[285].

4.4.2 Bioinformatics workflow

Quality trimming of sequencing reads, initial species identification and serovar prediction were performed as described in sections 3.4.2.3-3.4.2.5. Detection of ARDs was carried out as detailed in section 3.4.5.1.

4.4.2.1 Single nucleotide polymorphism typing

For single nucleotide polymorphism (SNP) typing, sequence reads were mapped to a reference genome from the same eBurst group using the BWA-MEM algorithm from the Burrows-Wheeler Aligner package[313] and SNPs were identified with GATK, version 2.6.5[314]. Isolates were grouped together into clusters of increasing similarity and assigned a hierarchical 'SNP address' as described in Dallman et al.[315].

4.4.3 Antimicrobial susceptibility testing

Isolates were recovered from the PHE archive and retrospective AST was performed and interpreted using EUCAST breakpoints and screening concentrations as outlined in **Table 4.1**. Temocillin and cefoxitin were included in the panel to aid the detection of OXA-48-like carbapenemase and acquired AmpC genes, respectively. While isolates with an MIC greater than the lower of the two screening concentrations were

already classed as resistant for cefotaxime, ceftazidime and chloramphenicol, decreased susceptibility (MIC 0.06-0.25 mg/L) and resistance (MIC >0.5 mg/L) were distinguished for ciprofloxacin. Breakpoint values were determined by agar dilution with Müller-Hinton agar (2 g/L beef extract, 17.5 g/L acid hydrosylate of casein, 1.5 g/L starch). If required, MICs were confirmed by Etest® (bioMérieux, Marcy-l'Étoile, France) or by agar dilution.

Table 4.1: Antimicrobials included in the testing panel and associated screening concentrations.

	Antimicrobial class	Screening concentrations (mg/L)
Ampicillin	β-lactam (penicillin)	8
Temocillin	β-lactam (penicillin)	128
Cefoxitin	β-lactam (2 nd-generation cephalosporin)	8
Cefotaxime	β-lactam (3 rd-generation cephalosporin)	0.5 & 1
Ceftazidime	β-lactam (3 rd-generation cephalosporin)	1 & 2
Cefpirome	β-lactam (4 th-generation cephalosporin)	1
Ertapenem	β-lactam (carbapenem)	0.064
Chloramphenicol	phenicol	8 & 16
Gentamicin	aminoglycoside	2
Streptomycin	aminoglycoside	16
Tobramycin	aminoglycoside	2
Sulphonamides	sulphonamide	16
Tetracycline	tetracycline	8
Trimethoprim	trimethoprim	2
Ciprofloxacin	quinolone	0.064 & 0.5

4.4.4 Statistical analysis

Chi-square tests of association were carried out using the chisq.test function in R. A p-value ≤0.05 was considered statistically significant. Bonferroni correction for multiple testing was performed by dividing this threshold p-value by the number of tests.

4.5 Results

4.5.1 Distribution of serovars in the dataset

From the 7,009 isolates that underwent WGS at PHE in the study period, outbreak cases were identified based on SNP addresses identical down to the five SNP level and deduplicated. Furthermore, isolates with an insufficient number of high-quality reads left after trimming and those for which the KmerID pipeline indicated contamination with another strain were excluded. This resulted in inclusion of 3,491 (49.81%) of the total isolates in the study. Of these, the majority (n=3,487) were of human origin, three were derived from a food and one from an unknown source.

Amongst the 3,491 isolates included, 3,471 (99.43%) were identified as *S. enterica* subsp. *enterica*, comprising 222 different serovars and 58 isolates that could not successfully be subtyped to serovar level. Eleven isolates (0.32%) were classified as *S. enterica* subsp. *houtenae* and nine (0.26%) as *S. enterica* subsp. *salamae*. The ten most common subsp. *enterica* serovars were *S*. Typhimurium (23.69%, n=827), *S*. Enteritidis (8.42%, n=294), *S*. Virchow (4.01%, n=140), *S*. Stanley (3.78%, n=132), *S*. Newport (3.75%, n=131), *S*. Infantis (3.47%, n=121), *S*. Kentucky (3.12%, n=109), *S*. Oranienburg (2.06%, n=72), *S*. Java (2.03%, n=71) and *S*. Saint-paul (1.86%, n=65) (**Fig. 4.1**).

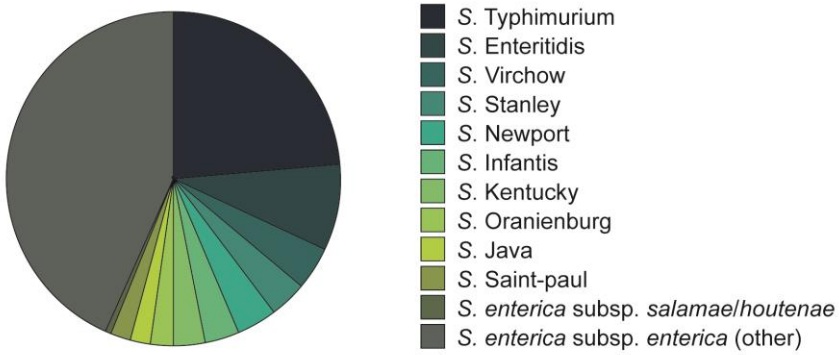

Figure 4.1: Distribution of serovars amongst 3,491 non-typhoidal *S. enterica* isolates. The ten most common serovars plus those belonging to a subspecies other than *enterica* are shown. The *S. enterica* subsp. *enterica* (other) group comprises 212 additional serovars and 58 isolates that could not successfully be subtyped to serovar level.

4.5.2 Prevalence of phenotypic antimicrobial resistance

Phenotypic resistance to at least one antimicrobial of the testing panel (cf. **Table 4.1**) (excluding decreased susceptibility to ciprofloxacin) was detected in 1,083 (31.02%) of the 3,491 total isolates (**Fig. 4.2**). Focussing on the ten most common serovars, the percentage of resistant isolates was highest in *S*. Kentucky (73.39%, n=80), *S*. Typhimurium (66.51%, n=550) and *S*. Infantis (50.41%, n=61). The lowest number of resistant isolates was observed for *S*. Oranienburg (1.39%, n=1), *S*. Java (9.86%, n=7) and *S*. Enteritidis (13.61%, n=40). All *S. enterica* subsp. *salamae* and *houtenae* isolates were pansusceptible to the testing panel.

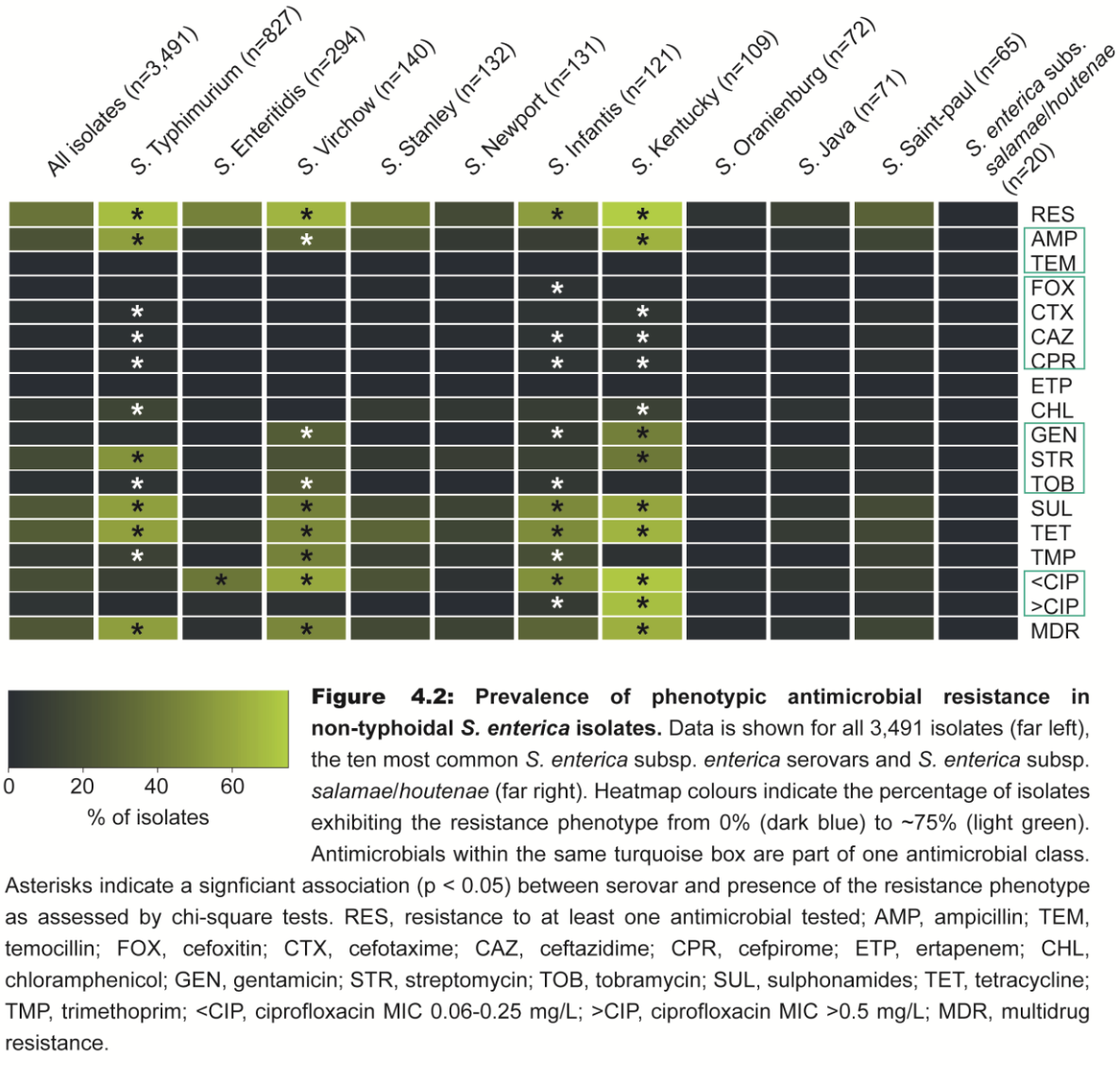

Figure 4.2: Prevalence of phenotypic antimicrobial resistance in non-typhoidal *S. enterica* isolates. Data is shown for all 3,491 isolates (far left), the ten most common *S. enterica* subsp. *enterica* serovars and *S. enterica* subsp. *salamae*/*houtenae* (far right). Heatmap colours indicate the percentage of isolates exhibiting the resistance phenotype from 0% (dark blue) to ~75% (light green). Antimicrobials within the same turquoise box are part of one antimicrobial class. Asterisks indicate a signficiant association ($p < 0.05$) between serovar and presence of the resistance phenotype as assessed by chi-square tests. RES, resistance to at least one antimicrobial tested; AMP, ampicillin; TEM, temocillin; FOX, cefoxitin; CTX, cefotaxime; CAZ, ceftazidime; CPR, cefpirome; ETP, ertapenem; CHL, chloramphenicol; GEN, gentamicin; STR, streptomycin; TOB, tobramycin; SUL, sulphonamides; TET, tetracycline; TMP, trimethoprim; <CIP, ciprofloxacin MIC 0.06-0.25 mg/L; >CIP, ciprofloxacin MIC >0.5 mg/L; MDR, multidrug resistance.

Tetracycline resistance was the most prevalent phenotype (26.27% of all isolates, n=917), followed by resistance to sulphonamides (23.72%, n=828) and ampicillin (21.43%, n=748). In comparison, resistance to cephalosporins was low: 0.57% (n=20) for cefoxitin, 1.63% (n=57) for cefotaxime and 1.35% for both ceftazidime and cefpirome (n=47). Resistance to the carbapenem ertapenem was observed in ten isolates (0.29%), resistance to temocillin in a single isolate (0.03%). Amongst the aminoglycosides, streptomycin resistance was most common (17.79%, n=621, compared to 3.98%, n=139, for gentamicin and 2.78%, n=97, for tobramycin). While only 3.95% (n=138) of isolates exhibited full resistance to ciprofloxacin, decreased susceptibility to the fluoroquinolone was detected in 18.50% (n=646) of NTS in this study. MDR was observed in 23.09% (n=806) of isolates.

Chi-square tests of association between resistance phenotypes and the ten most common *S. enterica* serovars revealed that serovars Typhimurium ($p=3.24x10^{-140}$), Virchow ($p=1.17x10^{-6}$), Infantis ($p=4.37x10^{-6}$) and Kentucky ($p=7.2x10^{-22}$) were significantly associated with resistance to at least one antimicrobial when compared to all other isolates. Specifically, an association with aminoglycoside ($p<0.03$), sulphonamide ($p\leq9.57x10^{-9}$) and tetracycline resistance ($p\leq1.94x10^{-8}$) was observed in all four serovars. With the exception of *S*. Infantis, these serovars were also more likely to exhibit MDR ($p\leq4.49x10^{-11}$). Ampicillin resistance was more prevalent in *S*. Typhimurium ($p=1.62x10^{-174}$), *S*. Virchow ($p=0.02$) and *S*. Kentucky ($p=7.08x10^{-28}$) while an increased frequency of cephalosporin resistance was detected in *S*. Typhimurium ($p\leq0.01$), *S*. Infantis ($p=0.02$) and S. Kentucky ($p\leq0.01$). Trimethoprim resistance was more common in *S*. Typhimurium ($p=2.01x10^{-5}$), *S*. Virchow ($p=3.39x10^{-51}$) and *S*. Infantis ($p=9.28x10^{-5}$). Chloramphenicol resistance was more frequently detected in *S*. Typhimurium ($p=1.37x10^{-29}$) and *S*. Kentucky ($p=0.01$) and ciprofloxacin resistance in *S*. Infantis ($p=0.03$) and *S*. Kentucky ($p=7.08x10^{-277}$). While most *S*. Enteritidis isolates (86.39%, n=254) were pansusceptible to the testing panel and only nine (3.06%) displayed full ciprofloxacin resistance, decreased susceptibility to the antimicrobial, observed in 114 *S*. Enteritidis (38.78%), was positively associated with this serovar ($p=1.79x10^{-20}$). Similarly, only five *S*. Virchow isolates (3.57%) were fully resistant to ciprofloxacin but 61.43% (n=86) showed

decreased susceptibility. The Bonferroni-corrected p-value for these tests was $2.31x10^{-4}$.

4.5.2.1 Co-occurrence of phenotypic antimicrobial resistance

Among the 1,083 resistant isolates, 154 (14.22%) showed resistance to a single antimicrobial class and 123 (11.36%) were resistant to two antimicrobial classes. The majority (74.42%, n=806) exhibited MDR, resistance to three or more antimicrobial classes (**Figure 4.3**). MDR was especially common in *S*. Virchow (92.86% of resistant isolates, n=65), *S*. Kentucky (86.25%, n=69) and *S*. Typhimurium (83.82%, n=461). The majority of resistant *S*. Enteritidis (62.5%, n=25), on the other hand, only displayed resistance to a single antimicrobial class, either to ampicillins or quinolones. None of the isolates comprised in this study were resistant to all nine antimicrobial classes tested but one *S*. Typhimurium exhibited resistance to eight classes, as well as decreased susceptibility to ciprofloxacin.

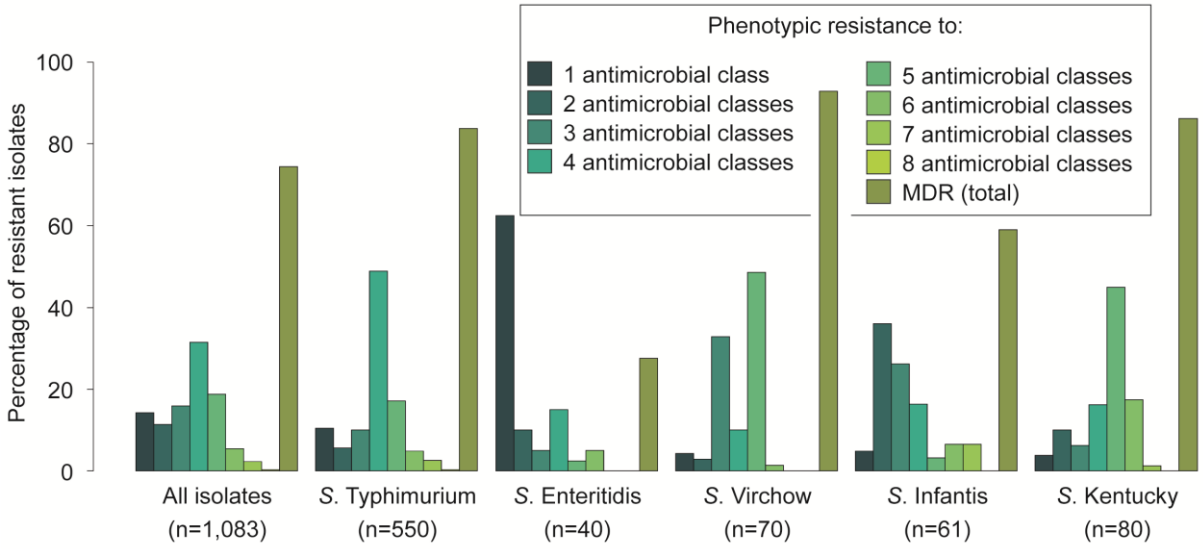

Figure 4.3: Prevalence of resistance to multiple antimicrobial classes in non-typhoidal *S. enterica*. Data is shown for all 1,083 resistant isolates (far left) and for resistant isolates belonging to serovars Typhimurium, Enteritidis, Virchow, Infantis and Kentucky. Bars on the right of each group depict the sum of MDR isolates.

The high frequency of resistance to more than one antimicrobial class prompted an investigation of resistance co-occurrence within isolates. Chi-square tests of association between individual resistance phenotypes were carried out (**Figure 4.4**).

All 1,083 resistant isolates, as well as 218 isolates showing decreased ciprofloxacin susceptibility, were included in the analysis. Co-occurrence of temocillin resistance with the other antimicrobials was not evaluated as resistance to temocillin was only detected in a single isolate. The Bonferroni-corrected p-value for the tests was 2.22×10^{-4}. There was a very strong association ($p \leq 7.23 \times 10^{-84}$) between resistance to different antimicrobials of the same class. This was the case for the cephalosporins as well as gentamicin and tobramycin. Between classes, ampicillin, streptomycin, sulphonamide and tetracycline resistance were strongly associated ($p \leq 1.89 \times 10^{-30}$).

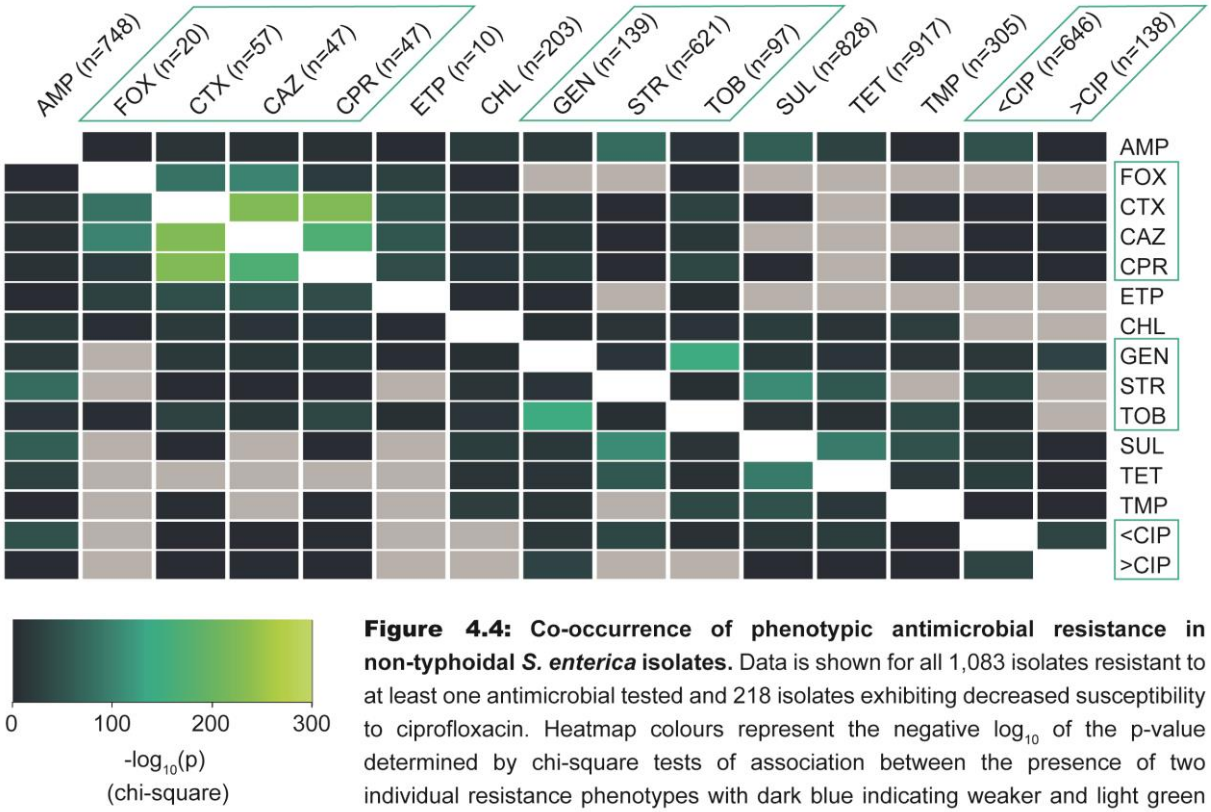

Figure 4.4: Co-occurrence of phenotypic antimicrobial resistance in non-typhoidal *S. enterica* isolates. Data is shown for all 1,083 isolates resistant to at least one antimicrobial tested and 218 isolates exhibiting decreased susceptibility to ciprofloxacin. Heatmap colours represent the negative $\log_{10}$ of the p-value determined by chi-square tests of association between the presence of two individual resistance phenotypes with dark blue indicating weaker and light green stronger association. Grey cells denote no association. Antimicrobials within the same turquoise box are part of one antimicrobial class. AMP, ampicillin; FOX, cefoxitin; CTX, cefotaxime; CAZ, ceftazidime; CPR, cefpirome; ETP, ertapenem; CHL, chloramphenicol; GEN, gentamicin; STR, streptomycin; TOB, tobramycin; SUL, sulphonamides; TET, tetracycline; TMP, trimethoprim; <CIP, ciprofloxacin MIC 0.06-0.25 mg/L; >CIP, ciprofloxacin MIC >0.5 mg/L.

Decreased ciprofloxacin susceptibility was associated with resistance to ampicillin ($p = 4.24 \times 10^{-48}$). Conversely, some resistance phenotypes were less likely to be found within the same isolate: there was no association between tetracycline and cephalosporin resistance; chloramphenicol resistance and decreased susceptibility or resistance to ciprofloxacin; streptomycin resistance and resistance to trimethoprim or

ciprofloxacin. The results of chi-square tests have to be interpreted with caution when resistance combinations occur at a low frequency. Cefoxitin and ertapenem resistance, for example, were only detected in few isolates decreasing the chance of these phenotypes co-occurring with others.

In line with these findings, co-resistance to ampicillin, streptomycin, sulphonamides and tetracycline was the most frequently observed MDR as well as overall resistance profile (22.53% of resistant isolates, n=244). However, almost all isolates displaying this profile (n=238) were *S*. Typhimurium. In *S*. Virchow co-resistance to sulphonamides, tetracycline and trimethoprim was most common (n=23) while the most frequent MDR profile in *S*. Kentucky comprised resistance to ampicillin, gentamicin, streptomycin, sulphonamides, tetracycline and ciprofloxacin (n=24).

4.5.3 Prevalence of genetic antimicrobial resistance determinants

To identify the genetic determinants underlying the AMR phenotypes observed, WGS of the isolates were analysed using the 'Genefinder' algorithm. Overall, 112 different genes and chromosomal mutations known to be associated with phenotypic resistance to the antimicrobials included in the testing panel were detected. Some ARDs were more prevalent in and significantly associated with specific *S. enterica* serovars (**Figure 4.5**), which will be described in more detail in the following sections. The Bonferroni-corrected p-value for these tests was 2.31×10^{-4}.

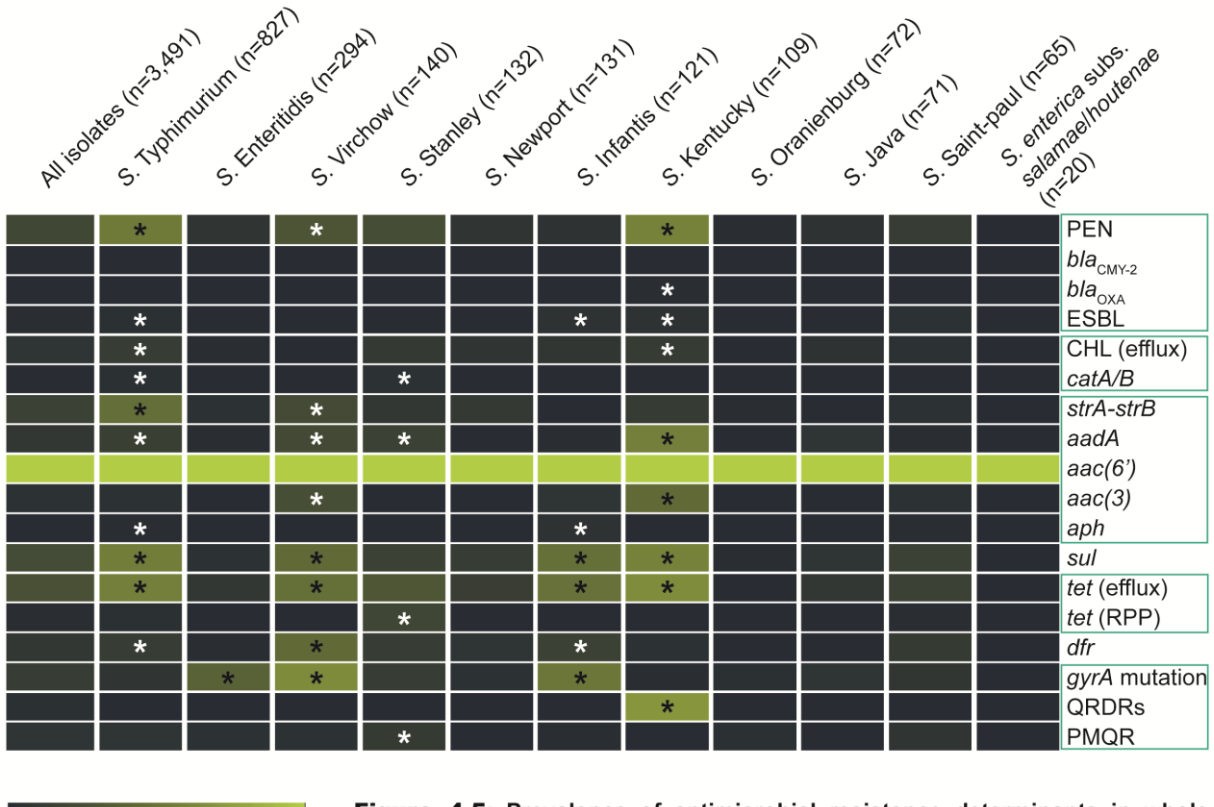

Figure 4.5: Prevalence of antimicrobial resistance determinants in whole genome sequences of non-typhoidal *S. enterica* isolates. Data is shown for all 3,491 isolates (far left), the ten most common *S. enterica* subsp. *enterica* serovars and *S. enterica* subsp. *salamae*/*houtenae* (far right). Heatmap colours indicate the percentage of isolates carrying the resistance determinant from 0% (dark blue) to 100% (light green). Determinants within one turquoise box confer resistance to the same antimicrobial class with all β-lactamase genes grouped together. Asterisks indicate a signficiant association ($p < 0.05$) between serovar and presence of the resistance determinant as assessed by chi-square tests. PEN, penicillinase genes; ESBL, extended-spectrum β-lactamase genes; CHL, chloramphenicol resistance genes; RPP, ribosomal protection protein; QRDRs, multiple mutations in the quinolone resistance-determining regions of *gyrA* and *parC*; PMQR, plasmid-mediated quinolone resistance genes.

4.5.3.1 β-lactam resistance determinants

Of the 3,491 isolates in this study, 749 (21.46%) carried genes conferring resistance to β-lactam antibiotics (**Table 4.2**). Carbapenemase genes were not detected. Penicillin resistance determinants were detected in 709 isolates (20.31%), most commonly $bla_{\text{TEM-1}}$ (n=603) and $bla_{\text{CARB-2}}$ (also known as $bla_{\text{PSE-1}}$) (n=75). Additionally, a further nine different TEM-type β-lactamase genes were identified. A single isolate carried the $bla_{\text{HERA-3}}$ gene. Serovars Typhimurium (n=451; $p=3x10^{-171}$),

Table 4.2: Prevalence of β-lactam resistance determinants in non-typhoidal *S. enterica* and selected serovars.

	Product	All isolates (n=3,491)	*S.* Typhimurium (n=827)	*S.* Virchow (n=140)	*S.* Infantis (n=121)	*S.* Kentucky (n=109)
bla $_{CARB-2}$	PEN	75 (2.15%)	59 (7.13%)	0	0	0
bla $_{CMY-2}$	AmpC	16 (0.46%)	7 (0.85%)	0	0	1 (0.92%)
bla $_{CTX-M-1}$	ESBL	3 (0.09%)	3 (0.36%)	0	0	
bla $_{CTX-M-14}$	ESBL	8 (0.23%)	2 (0.24%)	0	0	5 (4.59%)
bla $_{CTX-M-15}$	ESBL	4 (0.11%)	1 (0.12%)	0	0	0
bla $_{CTX-M-3}$	ESBL	1 (0.03%)	0	1 (0.71%)	0	0
bla $_{CTX-M-55}$	ESBL	9 (0.26%)	7 (0.85%)	0	0	0
bla $_{CTX-M-65}$	ESBL	6 (0.17%)	0	0	5 (4.13%)	0
bla $_{CTX-M-9}$	ESBL	10 (0.29%)	10 (1.21%)	0	0	0
bla $_{HERA-3}$	PEN	1 (0.03%)	1 (0.12%)	0	0	0
bla $_{OXA-1}$	class D	4 (0.11%)	3 (0.36%)	0	0	0
bla $_{OXA-10}$	class D	1 (0.03%)	0	0	0	1 (0.92%)
bla $_{OXA-114}$	class D	1 (0.03%)	0	0	0	1 (0.92%)
bla $_{OXA-134}$	class D	1 (0.03%)	1 (0.12%)	0	0	0
bla $_{SHV-12}$	ESBL	4 (0.11%)	0	0	0	0
bla $_{TEM-1}$	PEN	603 (17.27%)	392 (47.40%)	38 (27.14%)	6 (4.96%)	63 (57.80%)
bla $_{TEM-117}$	PEN	12 (0.34%)	1 (0.12%)	0	0	0
bla $_{TEM-135}$	PEN	7 (0.20%)	0	0	0	0
bla $_{TEM-159}$	PEN	4 (0.11%)	0	4 (2.86%)	0	0
bla $_{TEM-176}$	PEN	1 (0.03%)	0	0	0	0
bla $_{TEM-191}$	PEN	3 (0.09%)	1 (0.12%)	0	0	0
bla $_{TEM-215}$	PEN	1 (0.03%)	0	0	0	0
bla $_{TEM-217}$	PEN	1 (0.03%)	1 (0.12%)	0	0	0
bla $_{TEM-33}$	PEN	1 (0.03%)	1 (0.12%)	0	0	0
bla $_{TEM-57}$	PEN	6 (0.17%)	0	0	0	2 (1.83%)

Values denote the number/percentage of isolates carrying the resistance determinant. bla_{CARB-2} is also known as bla_{PSE-1}. The product column describes the type of β-lactamase encoded by the gene. PEN, penicillinase; ESBL, extended-spectrum β-lactamase.

Virchow (n=42; $p=4.92x10^{-3}$) and Kentucky (n=65; $p=1.06x10^{-24}$) were significantly associated with carriage of penicillinase genes.

The acquired AmpC resistance gene *bla*$_{CMY-2}$ was found in 16 isolates (0.46%). Seven isolates (0.20%), four serovar Typhimurium and two serovar Kentucky, carried OXA-type class D β-lactamase genes.

ESBL-encoding genes were present in 43 isolates (1.23%), with most carrying CTX-M type resistance determinants (n=41). The *bla*$_{SHV-12}$ ESBL gene was identified in four isolates. More than half (n=23) of the isolates with ESBL genes were serovar Typhimurium and a further five each were serovars Infantis and Kentucky. Combinations of penicillinase and ESBL genes occurred in 16 isolates.

4.5.3.2 Chloramphenicol resistance determinants

Genes linked to chloramphenicol resistance were identified in 215 isolates (6.16%) (**Table 4.3**). Efflux pump genes occurred in 194 isolates (5.56%), either *floR* (n=147) or *cmlA1* (n=67), and their carriage was significantly associated with serovars Typhimurium (n=107; $p=7.01x10^{-26}$) and Kentucky (n=13; $p=6.21x10^{-3}$). Chloramphenicol acetyltransferase genes of the *catA*- or *catB*-type were detected in 32 isolates (0.92%). Of these, 19 were *S*. Typhimurium and four *S*. Stanley. Eleven isolates harboured genes encoding both an efflux pump and an acetyltransferase.

Table 4.3: Prevalence of chloramphenicol resistance determinants in non-typhoidal *S. enterica* and selected serovars.

	Product	All isolates (n=3,491)	*S*. Typhimurium (n=827)	*S*. Stanley (n=132)	*S*. Kentucky (n=109)
catA1	AcT	19 (0.54%)	16 (1.93%)	0	1 (0.92%)
catA2	AcT	12 (0.34%)	3 (0.36%)	4 (3.03%)	0
catB3	AcT	1 (0.03%)	0	0	0
cmlA1	efflux pump	67 (1.92%)	34 (4.11%)	4 (3.03%)	5 (4.59%)
floR	efflux pump	147 (4.21%)	84 (10.16%)	11 (8.33%)	10 (9.17%)

Values denote the number/percentage of isolates carrying the resistance determinant. AcT, acetyltransferase.

4.5.3.3 Aminoglycoside resistance determinants

All but eight of the total 3,491 isolates carried an aminoglycoside acetyltransferase *aac(6')*-type gene, either *aac(6')-Iy* (n=1,997) or *aac(6')-Iaa* (n=1,486) (**Table 4.4**). Among the ten most common *S. enterica* serovars, the former was more frequent in *S*. Enteritidis, *S*. Virchow, *S*. Stanley, *S*. Newport, *S*. Infantis, *S*. Oranienburg, *S*. Java and *S*. Saint-paul, while the latter was more often found in *S*. Typhimurium and *S*. Kentucky. The aforementioned genes were the only resistance determinants identified in the 20 *S. enterica* subsp. *salamae*/*houtenae* isolates. Furthermore, *aac(6')-Ii* and *aac(6')-IIc* were detected in three and two isolates, respectively.

Streptomycin resistance determinants were detected in 728 isolates (20.85%): 537 carried *strA-strB* and 292 carried at least one of 13 *aadA*-type aminoglycoside adenylyltransferase genes, most commonly *aadA2* (n=189) and *aadA17* (n=107). Both *strA-strB* and an *aadA* variant occurred in 101 isolates. Carriage of *strA-strB* was significantly associated with serovars Typhimurium (n=381; $p=7.29x10^{-172}$) and Virchow (n=32; p=0.02), carriage of adenylyltransferase genes with serovars Typhimurium (n=124; $p=5.66x10^{-15}$), Virchow (n=30; $p=2.97x10^{-8}$), Stanley (n=22; $p=8.01x10^{-4}$) and Kentucky (n=63; $p=1.48x10^{-78}$).

Variants of aminoglycoside acetyltransferase genes of the *aac(3)*-type, associated with resistance to gentamicin and tobramycin, were present in 130 isolates (n=3.72%), most notably *aac(3)-Id* (n=50) and *aac(3)-IIa* (n=36). Serovars Virchow (n=36; $p=2.62x10^{-43}$) and Kentucky (n=47; $p=1.77x10^{-105}$) were significantly associated with carriage of *aac(3)* variants. The aminoglycoside adenylyltransferase gene *ant(2'')-Ia* was identified in 12 isolates (0.34%), nine of which were *S*. Typhimurium and three *S*. Infantis. Aminoglycoside phosphotransferase genes of the *aph*-type occurred in 27 isolates (0.77%), 16 *S*. Typhimurium and five *S*. Infantis. Both streptomycin and gentamicin/tobramycin resistance determinants were found in 122 isolates. No 16S rRNA methyltransferase genes were detected.

Table 4.4: Prevalence of aminoglycoside resistance determinants in non-typhoidal *S. enterica* and selected serovars.

	Product	All isolates (n=3,491)	*S.* Typhimurium (n=827)	*S.* Virchow (n=140)	*S.* Stanley (n=132)	*S.* Infantis (n=121)	*S.* Kentucky (n=109)
aac(3)-Id	AcT	50 (1.43%)	0	0	0	1 (0.83%)	47 (43.12%)
aac(3)-IIa	AcT	36 (1.03%)	1 (0.12%)	35 (25%)	0	0	0
aac(3)-IId	AcT	18 (0.52%)	11 (1.33%)	1 (0.71%)	1 (0.76%)	0	0
aac(3)-IVa	AcT	24 (0.69%)	16 (1.93%)	0	0	5 (4.13%)	0
aac(3)-VIa	AcT	2 (0.06%)	1 (0.12%)	0	0	1 (0.83%)	0
aac(6')-Iaa	AcT	1486 (42.57%)	821 (99.27%)	0	2 (1.52%)	0	81 (74.31%)
aac(6')-Ii	AcT	3 (0.09%)	1 (0.12%)	0	0	0	0
aac(6')-IIc	AcT	2 (0.06%)	0	0	0	0	0
aac(6')-Iy	AcT	1997 (57.20%)	4 (0.48%)	140 (100%)	130 (98.48%)	121 (100%)	28 (25.69%)
aadA1	AdT	20 (0.57%)	10 (1.21%)	0	4 (3.03%)	1 (0.83%)	1 (0.92%)
aadA12	AdT	10 (0.29%)	4 (0.48%)	1 (0.71%)	1 (0.76%)	0	2 (1.83%)
aadA13	AdT	3 (0.09%)	2 (0.24%)	0	0	1 (0.83%)	0
aadA15	AdT	18 (0.52%)	9 (1.09%)	0	1 (0.76%)	0	1 (0.92%)
aadA16	AdT	5 (0.14%)	0	0	0	0	0
aadA17	AdT	107 (3.07%)	72 (8.71%)	25 (17.86%)	1 (0.76%)	0	0
aadA2	AdT	189 (5.41%)	106 (12.82%)	30 (21.43%)	17 (12.88%)	0	5 (4.59%)
aadA22	AdT	9 (0.26%)	4 (0.48%)	0	1 (0.76%)	0	1 (0.92%)
aadA23	AdT	11 (0.32%)	8 (0.97%)	0	2 (1.52%)	0	0
aadA3	AdT	8 (0.23%)	4 (0.48%)	0	1 (0.76%)	0	0
aadA5	AdT	5 (0.14%)	2 (0.24%)	0	0	0	0
aadA7	AdT	64 (1.83%)	0	0	0	1 (0.83%)	60 (55.04%)
aadA8b	AdT	39 (1.12%)	12 (1.45%)	4 (2.86%)	10 (7.58%)	0	1 (0.92%)
ant(2'')-Ia	AdT	12 (0.34%)	9 (1.09%)	0	0	3 (2.48%)	0
aph(3'')-Ib	PhT	1 (0.03%)	0	0	0	0	0
aph(3')-IIa	PhT	10 (0.29%)	5 (0.60%)	0	1 (0.76%)	0	0
aph(4)-Ia	PhT	23 (0.66%)	15 (1.81%)	0	0	5 (4.13%)	0
strA-strB	PhT	537 (15.38%)	381 (46.07%)	32 (22.86%)	11 (8.33%)	1 (0.83%)	13 (11.93%)

Values denote the number/percentage of isolates carrying the resistance determinant. AcT, acetyltransferase; AdT, adenylyltransferase; PhT, phosphotransferase.

4.5.3.4 Sulphonamide resistance determinants

Genes associated with resistance to sulphonamides occurred in 830 isolates (23.78%): 490 carried *sul2*, 350 *sul1* and 75 *sul3* (**Table 4.5**). Combinations of two different *sul* genes were detected in 77 isolates and four isolates harboured all three variants. There was a significant association between carriage of sulphonamide resistance determinants and serovars Typhimurium (n=468; $p=1.54x10^{-141}$), Virchow (n=62; $p=1.08x10^{-8}$), Infantis (n=58; $p=4.24x10^{-10}$) and Kentucky (n=65; $p=1.14x10^{-18}$).

Table 4.5: Prevalence of sulphonamide resistance determinants in non-typhoidal *S. enterica* and selected serovars.

	Product	All isolates (n=3,491)	*S*. Typhimurium (n=827)	*S*. Virchow (n=140)	*S*. Infantis (n=121)	*S*. Kentucky (n=109)
sul1	DhpS	350 (10.03%)	99 (11.97%)	62 (44.29%)	56 (46.28%)	62 (56.88%)
sul2	DhpS	490 (14.04%)	390 (47.16%)	0	2 (1.65%)	7 (6.42%)
sul3	DhpS	75 (2.15%)	31 (3.75%)	1 (0.71%)	0	3 (2.75%)

Values denote the number/percentage of isolates carrying the resistance determinant. DhpS, dihydropteroate synthase.

4.5.3.5 Tetracycline resistance determinants

Tetracycline resistance determinants were identified in 927 isolates (26.55%) (**Table 4.6**). The majority carried at least one of four efflux pump-encoding genes, mainly *tet*(A) (n=843). In 57 isolates, the ribosomal protection protein gene *tet*(M) was found. A combination of two different tetracycline resistance determinants was detected in 56 isolates. Carriage of efflux pump genes was significantly associated with serovars Typhimurium (n=478; $p=2.49x10^{-121}$), Virchow (n=67; $p=7.16x10^{-9}$), Infantis (n=59; $p=2.4x10^{-8}$) and Kentucky (n=71; $p=3.04x10^{-20}$), carriage of *tet*(M) with *S*. Stanley (n=26; $p=4.72x10^{-60}$).

Table 4.6: Prevalence of tetracycline resistance determinants in non-typhoidal *S. enterica* and selected serovars.

	Product	All isolates (n=3,491)	*S.* Typhimurium (n=827)	*S.* Virchow (n=140)	*S.* Stanley (n=132)	*S.* Infantis (n=121)	*S.* Kentucky (n=109)
tet (A)	efflux pump	843 (24.15)	426 (51.51%)	67 (47.86%)	40 (30.30%)	57 (47.11%)	71 (65.14%)
tet (C)	efflux pump	10 (0.29%)	2 (0.24%)	0	0	0	0
tet (D)	efflux pump	5 (0.14%)	0	0	0	2 (1.65%)	0
tet (G)	efflux pump	65 (1.95%)	55 (6.65%)	0	0	0	0
tet (M)	RPP	57 (1.63%)	17 (2.06%)	0	26 (19.70%)	0	0

Values denote the number/percentage of isolates carrying the resistance determinant. RPP, ribosomal protection protein.

4.5.3.6 Trimethoprim resistance determinants

Dihydrofolate reductase genes of the *dfrA*-type conferring resistance to trimethoprim occurred in 302 isolates (8.65%) (**Table 4.7**). The most common genes were *dfrA12* (n=84), *dfrA1* (n=81), *dfrA14* (n=65) and *dfrA5* (n=41). Additionally, a further seven other *dfrA* variants were detected. Only one *S*. Typhimurium harboured a combination of two different genes, namely *dfrA1* and *dfrA12*. There was a significant association between carriage of trimethoprim resistance genes and serovars Typhimurium (n=103; $p=1.17x10^{-5}$), Virchow (n=62; $p=6.96x10^{-52}$) and Infantis (n=20; $p=2.95x10^{-3}$).

Table 4.7: Prevalence of trimethoprim resistance determinants in non-typhoidal *S. enterica* and selected serovars.

	Product	All isolates (n=3,491)	*S.* Typhimurium (n=827)	*S.* Virchow (n=140)	*S.* Infantis (n=121)
dfrA1	DhfR	81 (2.32%)	17 (2.06%)	39 (27.86%)	3 (2.48%)
dfrA12	DhfR	84 (2.41%)	38 (4.59%)	4 (2.86%)	0
dfrA14	DhfR	65 (1.86%)	26 (2.78%)	0	15 (12.40%)
dfrA15	DhfR	6 (0.17%)	0	0	0
dfrA16	DhfR	9 (0.26%)	9 (1.09%)	0	0
dfrA17	DhfR	5 (0.14%)	2 (0.24%)	0	0
dfrA18	DhfR	1 (0.03%)	0	0	0
dfrA27	DhfR	5 (0.14%)	0	0	0
dfrA5	DhfR	41 (1.17%)	14 (1.69%)	19 (13.57%)	0
dfrA7	DhfR	3 (0.09%)	1 (0.12%)	0	0
dfrA8	DhfR	3 (0.09%)	0	0	2 (1.65%)

Values denote the number/percentage of isolates carrying the resistance determinant. DhfR, dihydrofolate reductase.

4.5.3.7 Fluoroquinolone resistance determinants

Acquired genes and chromosomal mutations associated with decreased susceptibility and resistance to ciprofloxacin were detected in 2,241 isolates (64.19%) (**Table 4.8**). However, the majority of these isolates (n=1,554) only carried a mutation in the quinolone resistance-determining region (QRDR) of the DNA topoisomerase gene *parC* causing a threonine to serine substitution at position 57 of the encoded protein. Seven isolates harboured the fluoroquinolone- and aminoglycoside-modifying N-acetyltransferase gene *aac(6')-Ib-cr*, six of these in combination with other fluoroquinolone resistance determinants.

Single mutations in the QRDR of the DNA gyrase subunit gene *gyrA* were identified in 430 isolates (12.32%), most commonly *gyrA*[87:D-Y] (n=155) and

Table 4.8: Prevalence of fluoroquinolone resistance determinants including chromosomal mutations in non-typhoidal *S. enterica* and selected serovars.

	Product	All isolates (n=3,491)	*S*. Enteritidis (n=294)	*S*. Virchow (n=140)	*S*. Stanley (n=132)	*S*. Infantis (n=121)	*S*. Kentucky (n=109)
aac(6')-Ib-cr	AcT	7 (0.20%)	1 (0.34%)	0	0	0	1 (0.92%)
gyrA [100:Y-H]	gyrase	4 (0.11%)	4 (1.36%)	0	0	0	0
gyrA [100:Y-X]	gyrase	1 (0.03%)	1 (0.34%)	0	0	0	0
gyrA [69:V-X]	gyrase	1 (0.03%)	0	0	0	0	0
gyrA [72:D-X; 82:D-X]	gyrase	1 (0.03%)	0	0	0	0	0
gyrA [83:S-F; 87:D-G]	gyrase	14 (0.40%)	0	0	0	0	11 (10.09%)
gyrA [83:S-F; 87:D-N]	gyrase	26 (0.74%)	0	0	0	0	24 (22.02%)
gyrA [83:S-F; 87:D-Y]	gyrase	41 (1.17%)	0	0	0	0	41 (37.61%)
gyrA [83:S-F]	gyrase	52 (1.49%)	10 (3.40%)	2 (1.43%)	0	0	2 (1.83%)
gyrA [83:S-X; 87:D-X;100:Y-X]	gyrase	1 (0.03%)	0	0	0	0	1 (0.92%)
gyrA [83:S-Y]	gyrase	111 (3.18%)	21 (7.14%)	18 (12.86%)	6 (4.55%)	38 (31.40%)	0
gyrA [87:D-B]	gyrase	6 (0.17%)	0	0	0	0	0
gyrA [87:D-G; 100:Y-H]	gyrase	1 (0.03%)	1 (0.34%)	0	0	0	0
gyrA [87:D-G]	gyrase	36 (1.03%)	16 (5.44%)	0	0	12 (9.92%)	0
gyrA [87:D-N]	gyrase	64 (1.83%)	27 (9.18%)	5 (3.57%)	0	1 (0.83%)	0
gyrA [87:D-Y]	gyrase	155 (4.44%)	35 (11.90%)	66 (11.90%)	10 (7.58%)	11 (9.09%)	0
oqxA	PMQR	1 (0.03%)	0	0	0	0	0
oqxB	PMQR	1 (0.03%)	0	0	0	0	0
parC [56:A-V; 57:T-S]	TI	1 (0.03%)	0	0	0	0	0
parC [57:T-N]	TI	1 (0.03%)	0	0	0	0	0
parC [57:T-S; 80:S-I]	TI	77 (2.21%)	0	0	0	0	76 (69.72%)
parC [57:T-S; 80:S-R]	TI	4 (0.11%)	0	0	0	0	0
parC [57:T-S; 98:P-A]	TI	1 (0.03%)	0	0	0	0	0
parC [57:T-S]	TI	1808 (51.79%)	2 (0.68%)	23 (16.43%)	132 (100%)	121 (100%)	32 (29.36%)
parC [57:T-X; 80:S-X]	TI	1 (0.03%)	0	0	0	0	1 (0.92%)
parC [57:T-X]	TI	5 (0.14%)	0	1 (0.71%)	0	0	0
qepA1	PMQR	1 (0.03%)	0	0	0	0	0

	Product	All isolates (n=3,491)	*S*. Enteritidis (n=294)	*S*. Virchow (n=140)	*S*. Stanley (n=132)	*S*. Infantis (n=121)	*S*. Kentucky (n=109)
qnrA1	PMQR	10 (0.29%)	0	0	2 (1.52%)	0	0
qnrB1	PMQR	3 (0.09%)	0	0	0	0	0
qnrB19	PMQR	49 (1.40%)	8 (2.72%)	0	0	2 (1.65%)	0
qnrB6	PMQR	4 (0.11%)	1 (0.34%)	0	0	0	0
qnrB7	PMQR	1 (0.03%)	0	0	0	0	0
qnrB9	PMQR	4 (0.11%)	0	0	0	0	0
qnrD	PMQR	19 (0.54%)	0	0	0	0	0
qnrS1	PMQR	95 (2.72%)	1 (0.34%)	5 (3.57%)	9 (6.82%)	0	1 (0.92%)
qnrS2	PMQR	4 (0.11%)	0	0	3 (2.27%)	0	0

Values denote the number/percentage of isolates carrying the resistance determinant. AcT, acetyltransferase; PMQR, plasmid-mediated quinolone resistance determinant; TI, topoisomerase.

gyrA[83:S-Y] (n=111). In 152 isolates, this *gyrA* mutation was accompanied by a single mutation in *parC*. Single *gyrA* mutations were significantly associated with serovars Enteritidis (n=114; $p=1.37x10^{-46}$), Virchow (n=91; $p=2.13x10^{-82}$) and Infantis (n=62; $p=2.56x10^{-39}$).

Plasmid-mediated quinolone resistance (PMQR) genes were found in 189 isolates (5.41%), most frequently *qnrS1* (n=95) and *qnrB19* (n=49). In 20 isolates, PMQR genes occurred together with a single *gyrA* mutation, with or without a single mutation in *parC*, and in 103 isolates with a single *parC* mutation only. Carriage of PMQR genes was significantly associated with *S*. Stanley (n=14; p=0.01).

Multiple mutations in the QRDR of both *gyrA* and *parC* were detected in 82 isolates (2.35%). The most common combinations were either *gyrA*[83:S-F;87:D-Y] (n=41) or *gyrA*[83:S-F;87:D-N] (n=25) with *parC*[57:T-S;80:S-I]. There was a significant association between carriage of multiple *gyrA* and *parC* mutations and serovar Kentucky (n=77; p=0). Of the isolates carrying both multiple *gyrA* and *parC* mutations, only one *S*. Indiana had additional PMQR genes, namely *oqxA* and *oqxB*.

4.5.3.8 Co-occurrence of genetic resistance determinants

In the previous sections, the co-occurrence of genes associated with resistance to antimicrobials from the same class was described. However, the frequently observed

phenotypic co-resistance to antimicrobials from different classes (cf. **Figure 4.4**) suggests that some of their genetic determinants must also co-occur. Indeed, chi-square tests of association between individual genotypes indicated that penicillinase genes, *strA-strB* and *aadA* variants, *sul* genes and tetracycline efflux pump genes, which underlie the common ampicillin, streptomycin, sulphonamides and tetracycline resistance phenotype (ASSuT), were strongly associated ($p \leq 4.5 \times 10^{-146}$) (**Figure 4.6**). Furthermore, chloramphenicol efflux genes commonly co-occurred with penicillinase genes ($p=1.02 \times 10^{-120}$), *aadA* variants ($p=6.64 \times 10^{-249}$) and determinants of sulphonamide ($p=1.99 \times 10^{-127}$), tetracycline ($p=5.19 \times 10^{-109}$) and trimethoprim resistance ($p=4.16 \times 10^{-88}$). Association between phenotypic chloramphenicol resistance and resistance to ampicillin, streptomycin, sulphonamide, tetracycline and trimethoprim was less strong than for the resistance profile described previously, yet still significant ($p \geq 1.03 \times 10^{-24}$ compared to $p \leq 1.89 \times 10^{-30}$). Multiple mutations in the QRDR of *gyrA* and *parC* were associated with *aadA* and *aac(3)* variants ($p=1.08 \times 10^{-115}$ and 1.14×10^{-151}, respectively). The Bonferroni-corrected p-value for these tests was 1.54×10^{-4}. It should be kept in mind that for the co-occurrence analysis of phenotypic resistance, the 2,190 isolates showing neither resistance to any of the antimicrobials tested nor reduced susceptibility to ciprofloxacin were excluded while analysis of the genetic determinants included all 3,491 isolates, and therefore also genes that did not result in phenotypic consequences.

In line with the findings described above and disregarding the isolates that carried an *aac(6')* variant with or without *parC*[57:T-S], ARDs normally considered silent[273,276], as the only genetic resistance determinant, the most commonly occurring genotype in *S*. Typhimurium was bla_{TEM-1}, *strA-strB*, *aac(6')-Iaa*, *sul2*, *tet(A)* (n=208). In *S*. Enteritidis, for which only few isolates exhibited phenotypic resistance, 97 isolates harboured a single *gyrA* mutation and *aac(6')-Iy*. The bla_{TEM-1}, *strA-strB*, *aadA2*, *aadA17*, *aac(6')-Iy*, *aac(3)-IIa*, *sul1*, *tet(A)*, *dfrA1* genotype was detected in 17 *S*. Virchow isolates, the bla_{TEM-1}, *aadA7*, *aac(6')-Iaa*, *aac(3)-Id*, *sul1*, *tet(A)*, *gyrA*[83:S-F;87:D-Y], *parC*[57:T-S;80:S-I] genotype in 24 *S*. Kentucky.

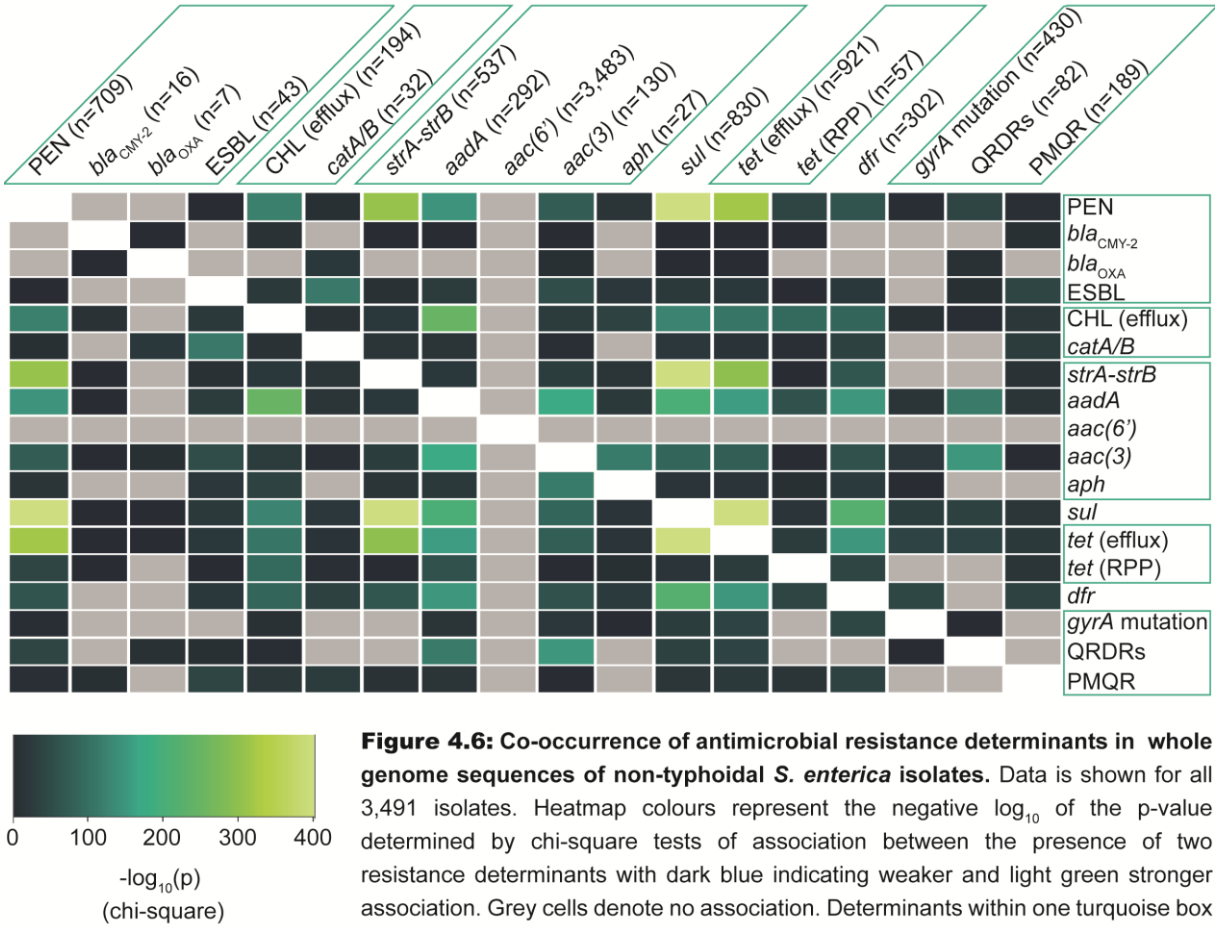

Figure 4.6: Co-occurrence of antimicrobial resistance determinants in whole genome sequences of non-typhoidal *S. enterica* isolates. Data is shown for all 3,491 isolates. Heatmap colours represent the negative $\log_{10}$ of the p-value determined by chi-square tests of association between the presence of two resistance determinants with dark blue indicating weaker and light green stronger association. Grey cells denote no association. Determinants within one turquoise box confer resistance to the same antimicrobial class with all β-lactamase genes grouped together. PEN, penicillinase genes; ESBL, extended-spectrum β-lactamase genes; CHL, chloramphenicol resistance genes; RPP, ribosomal protection protein; QRDRs, multiple mutations in the quinolone resistance-determining regions of *gyrA* and *parC*; PMQR, plasmid-mediated quinolone resistance genes.

4.5.4 Prediction of phenotypic resistance profiles from genotype

4.5.4.1 Comparison of phenotypic and genotypic resistance profiles

Prediction of phenotypic antimicrobial resistance profiles based on the genetic resistance determinants detected in the WGS was carried out for all 3,491 isolates using information from the literature about the phenotypic consequences of the ARDs in question. Results of AMR profile prediction from genotype were compared to the AMR profiles resulting from phenotypic susceptibility testing. For ciprofloxacin, in cases where the genotype suggested reduced susceptibility but the isolate exhibited full resistance phenotypically, the prediction was still classed as a match since

reduced susceptibility has been linked with clinical failure[287,288]. The phenotypic and genotypic AMR profiles of 3,415 isolates (97.82%) were entirely in agreement for both approaches for all 15 antimicrobials from nine different classes tested. For the 76 isolates with discordant results, the genotype wrongly predicted pan-susceptibility for one isolate (1.32% of wrong predictions) phenotypically resistant to ciprofloxacin. For a further 64 discrepant results (84.21%), the mismatch was based on false or missing prediction of resistance to a single antimicrobial, and for the remaining 11 (14.47%) to two antimicrobials. Overall, 52,277 (99.83%) out of a possible 52,365 predictions for each isolate and individual antimicrobial of the testing panel were correct (**Table 4.9**). Sensitivity of resistance prediction from genotype, defined as the

Table 4.9: Comparison of phenotypic antimicrobial susceptibility testing and genome-derived resistance prediction for 3,491 non-typhoidal *S. enterica* isolates.

	PT: S		PT: R	
	GT: R	GT: S	GT: R	GT: S
AMP	1 (0.04%)	2,742	747	1 (0.1%)
TEM	0	3,490	0	1 (100%)
FOX	0	3,471	19	1 (5%)
CTX	0	3,434	57	0
CAZ	0	3,444	47	0
CPR	0	3,444	47	0
ETP	0	3,481	10	0
CHL	4 (0.1%)	3,284	201	2 (1%)
GEN	1 (0.03%)	3,351	138	1 (0.7%)
STR	51 (1.78%)	2,821	613	8 (1.3%)
TOB	2 (0.1%)	3,392	97	0
SUL	2 (0.1%)	2,661	828	0
TMP	6 (0.2%)	2,568	917	0
TET	1 (0.03%)	3,185	301	4 (1.3%)
>CIP	1 (0.03%)	3,352	137	1 (0.7%)

Values denote the number of isolates. Values in brackets show the percentage of major (left) and very major prediction errors (right) associated with each phenotype. PT, phenotype; GT, genotype; S, susceptible; R, resistant; AMP, ampicillin; TEM, temocillin; FOX, cefoxitin; CTX, cefotaxime; CAZ, ceftazidime; CPR, cefpirome; CHL, chloramphenicol; GEN, gentamicin; STR, streptomycin; TOB, tobramycin; SUL, sulphonamides; TET, tetracycline; TMP, trimethoprim; >CIP, ciprofloxacin MIC >0.5 mg/L.

number of isolates correctly predicted to be resistant divided by the sum of isolates correctly predicted to be resistant and isolates falsely predicted to be susceptible, was ≥95% for all antimicrobials except temocillin (**Figure 4.7**). However, only a single isolate was found to be phenotypically resistant to temocillin. Specificity of prediction, defined as the number of isolates correctly predicted to be susceptible divided by the sum of isolates correctly predicted to be susceptible and isolates falsely predicted to be resistant, exceeded 98% for all 15 antimicrobials tested. The nature of false predictions will be discussed in more detail below.

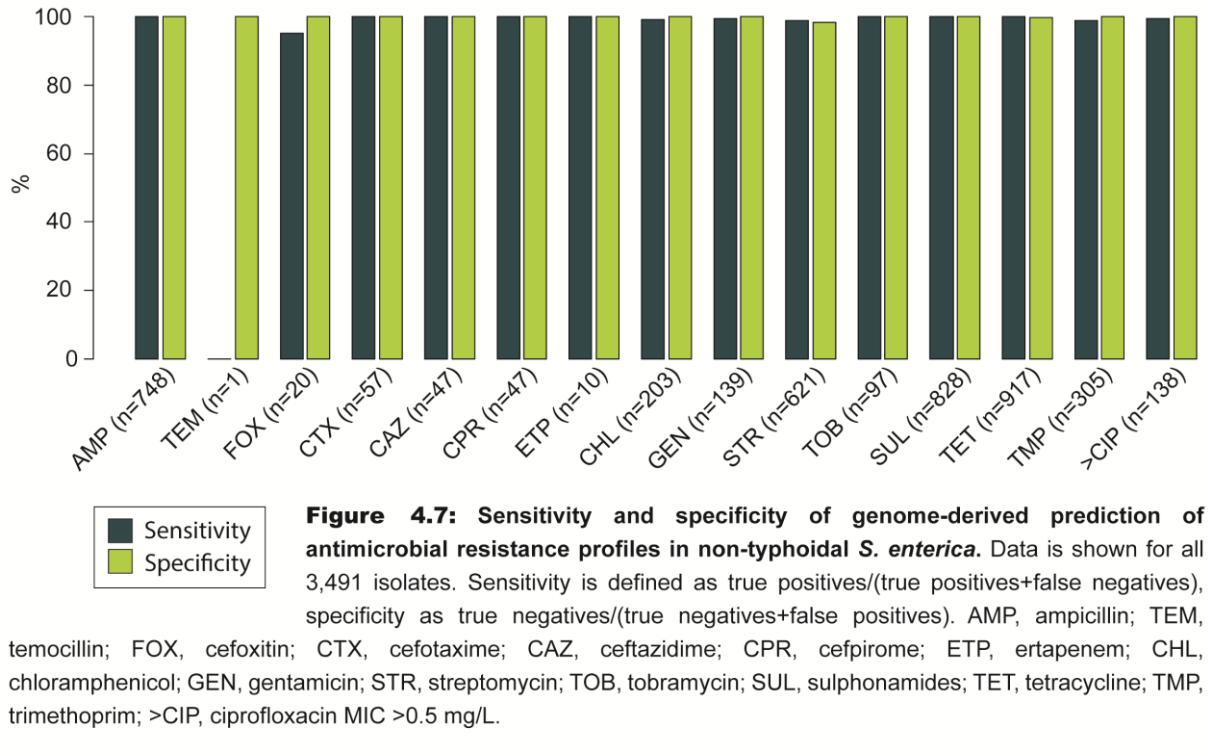

Figure 4.7: Sensitivity and specificity of genome-derived prediction of antimicrobial resistance profiles in non-typhoidal *S. enterica*. Data is shown for all 3,491 isolates. Sensitivity is defined as true positives/(true positives+false negatives), specificity as true negatives/(true negatives+false positives). AMP, ampicillin; TEM, temocillin; FOX, cefoxitin; CTX, cefotaxime; CAZ, ceftazidime; CPR, cefpirome; ETP, ertapenem; CHL, chloramphenicol; GEN, gentamicin; STR, streptomycin; TOB, tobramycin; SUL, sulphonamides; TET, tetracycline; TMP, trimethoprim; >CIP, ciprofloxacin MIC >0.5 mg/L.

4.5.4.2 Major and very major prediction errors

Of the 88 isolate/antimicrobial mismatches, 69 (78.41%) constituted major errors (MEs), i.e. isolates were genotypically predicted to be resistant but showed phenotypic susceptibility, rather than very major errors (VMEs), which were genotypically susceptible but phenotypically resistant. This equates to an overall ME rate of 0.13% and a VME rate of 0.04%.

The relationship between geno- and phenotype is relatively straightforward in some cases: isolates carrying a penicillinase gene are expected to exhibit ampicillin

resistance, *bla*$_{CMY-2}$, class D β-lactamases and ESBLs can confer resistance to cephalosporins, yet they will also be active against ampicillin. The single ME associated with predicted ampicillin resistance was based on the detection of *bla*$_{TEM-1}$ without phenotypic consequences. The presence of *sul*, *tet* and *dfrA* genes is linked to resistance to sulphonamides, tetracycline and trimethoprim, respectively. *sul1*, *sul2* and *dfrA14* were present without phenotypic consequences in one isolate each. Five of the six isolates with predicted but not phenotypic tetracycline resistance harboured *tet(M)*.

Despite *strA-strB* or *aadA* being good predictors of streptomycin resistance, the largest fraction of the 88 mismatches (n=59, 67.05%) was associated with this antimicrobial. Fifty-one were MEs, with *strA-strB* detected in 29 and an *aadA* variant in 21 genomes without phenotypic consequences. Since *aac(6')* variants are normally silent and only become transcriptionally active in rare cases[276], their presence was not considered an indicator of streptomycin resistance. However, all eight isolates falsely predicted to be susceptible to this antimicrobial carried *aac(6')-Iy* as the only aminoglycoside resistance determinant. For the other aminoglycosides, *aac(3)*-type acetyltransferase genes indicate gentamicin and tobramycin resistance. The single ME associated with gentamicin resistance was linked to the presence of *aac(3)-IId*, the VME to carriage of *aac(6')-Iy* only. Of the two isolates falsely predicted to be resistant to tobramycin, one harboured *ant(2")-Ia* and the second one *aac(3)-IIa*.

In the case of ciprofloxacin, where decreased susceptibility and full resistance can be distinguished, prediction of phenotype from genotype becomes more complex. The most common ARD profiles and resulting phenotypes are shown in **Table 4.10**. A combination of mutations in the QRDRs of *gyrA* and *parC* is expected to confer full ciprofloxacin resistance[316] while single *gyrA* mutations and/or the presence of PMQRs are thought to be the basis of decreased susceptibility. Carriage of *parC*[57:T-S] alone, without additional mutations in *gyrA*, does not generally result in altered interactions with the antimicrobial[273]. Of the 138 isolates displaying phenotypic ciprofloxacin resistance, 19 carried a single *gyrA* mutation only, 17 a single *gyrA* mutation together with a PMQR gene, and 20 had one or more PMQR genes. The single VME was associated with detection of *parC*[57:T-S] as the sole

Table 4.10: Most common genotypic determinants of reduced susceptibility and resistance to ciprofloxacin in 3,491 non-typhoidal *S. enterica* isolates.

	<CIP		>CIP	
	S	R	S	R
gyrA [87:D-Y]	6	110	114	2
gyrA [83:S-Y];*parC* [57:T-S]	1	50	42	9
gyrA [87:D-N]	4	47	51	0
gyrA [83:S-Y]	0	47	43	4
parC [57:T-S];*qnrS1*	0	44	39	5
gyrA [83:S-F;87:D-Y];*parC* [57:T-S;80:S-I]	0	41	0	41
qnrS1	1	36	28	9
gyrA [87:D-Y];*parC* [57:T-S]	2	32	34	0
parC [57:T-S];*qnrB19*	0	32	31	1
gyrA [83:S-F];*parC* [57:T-S]	1	28	28	1
gyrA [83:S-F;87:D-N];*parC* [57:T-S;80:S-I]	0	25	1	24
gyrA [83:S-F]	0	23	20	3
gyrA [87:D-G]	2	17	19	0
gyrA [87:D-G];*parC* [57:T-S]	2	15	17	0
parC [57:T-S];*qnrD*	6	6	12	0
qnrB19	0	12	9	3
g*yrA* [83:S-F;87:D-G];*parC* [57:T-S;80:S-I]	0	11	0	11
gyrA [87:D-N];*parC* [57:T-S]	0	11	11	0
qnrA1	0	8	8	0
aac(6')-Ib-cr	1	6	3	4
gyrA [83:S-Y];*parC* [57:T-S];*qnrS1*	0	4	0	4
gyrA [87:D-Y];*qnrS1*	0	4	0	4
parC [57:T-S];*qnrB9*	0	4	4	0
parC [57:T-S];*qnrS2*	0	4	4	0
qnrB6	0	4	2	2
gyrA [83:S-F;87:D-G];*parC* [57:T-S;80:S-R]	0	3	0	3
gyrA [83:S-Y];*parC* [57:T-S];*qnrD*	0	3	1	2
gyrA [83:S-Y];*qnrS1*	0	3	0	3
parC [57:T-S];*qnrA1*	0	2	2	0
parC [57:T-S];*qnrB1*	0	2	2	0

Values denote the number of isolates with the genotype exhibiting the resistance phenotype. Only genotypes with phenotypic consequences are shown. <CIP, ciprofloxacin MIC 0.06-0.25 mg/L; >CIP, ciprofloxacin MIC >0.5 mg/L; S, susceptible; R, resistant.

fluoroquinolone resistance determinant, the ME with the presence of *gyrA*[83:S-F;87:D-N] and *parC*[57:T-S;80:S-I] resulting in reduced susceptibility instead of full resistance. Although the QRDR of *gyrA* is located between amino acids 67 and 106, mutations at positions 83 and 87 are most common[317] and in the present study none of the isolates with only *gyrA* mutations at other positions of the QRDR exhibited reduced ciprofloxacin susceptibility.

4.5.5 *Salmonella* pheno- and genotypes and international travel

Travel history data was available for 1,070 isolates (30.65%). Classed according to the United Nations geoscheme, four of these isolates (0.37%) were linked to travel to Central Africa, 57 (5.33%) to East Africa, 150 (14.02%) to North Africa, four (0.37%) to Southern Africa, 42 (3.93%) to West Africa, 61 (5.70%) to the Caribbean, 37 (3.46%) to Central America, 14 (1.31%) to North America, 17 (1.59%) to South America, 11 (1.03%) to East Asia, 290 (27.10%) to Southeast Asia, 97 (9.07%) to South Asia, 100 (9.35%) to Western Asia, 16 (1.50%) to Eastern Europe, one (0.09%) to Northern Europe, 74 (6.92%) to Southern Europe, 10 (0.93%) to Western Europe and seven (0.65%) to Oceania. For the remaining 78 isolates, the travel destination was either recorded as “unknown” or as an entire continent.

Chi-square tests revealed that travel was significantly associated with acquisition of serovars Enteritidis ($p=9.19x10^{-26}$), Virchow ($p=3.54x10^{-3}$), Stanley ($p=2.48x10^{-4}$), Kentucky ($p=5.59x10^{-5}$) and Saint-paul ($p=1.67x10^{-3}$) (**Table 4.11**). While resistance in general, to at least one antimicrobial of the testing panel, and MDR were not more prevalent in travel-related isolates (32.06% vs. 30.57% and 22.99% vs. 23.13%, respectively), travel was significantly associated with phenotypic resistance to gentamicin ($p=5.16x10^{-3}$) as well as reduced susceptibility ($p=2.09x10^{-21}$) and resistance to ciprofloxacin ($p=2.88x10^{-5}$). Consequently, travel was also associated with the presence of single *gyrA* mutations ($p=7.67x10^{-9}$), PMQRs ($p=1.27x10^{-7}$), multiple *gyrA* and *parC* mutations ($p=8.51x10^{-6}$) and *aac(3)* variants ($p=8.11x10^{-3}$). Furthermore, *aadA* variants (p=0.02) and *tet(M)* ($p=3.67x10^{-3}$) were more likely to be found in isolates for which travel was reported.

Table 4.11: ***Salmonella*** **pheno- and genotypes significantly associated with international travel.**

	Travel (n=1,070)	No travel (n=2,421)
***S*. Enteritidis**	170 (15.89%)	124 (5.12%)
***S*. Virchow**	59 (5.51%)	81 (3.35%)
***S*. Stanley**	60 (5.61%)	72 (2.97%)
***S*. Kentucky**	53 (4.95%)	56 (2.31%)
***S*. Saint-paul**	32 (2.99%)	33 (1.36%)
GEN	58 (5.42%)	81 (3.35%)
<CIP	299 (27.94%)	347 (14.33%)
>CIP	65 (6.07%)	73 (3.02%)
aadA	108 (10.09%)	184 (7.60%)
aac(3)	54 (5.05%)	76 (3.14%)
tet(M)	28 (2.62%)	29 (1.20%)
***gyrA* mutation**	184 (17.20%)	246 (10.16%)
QRDRs	44 (4.11%)	38 (1.57%)
PMQR	91 (8.50%)	98 (4.05%)

Values denote the number/percentage of isolates, for which travel was recorded (left), and those, for which no travel history was available (right), displaying the pheno-/genotype. GEN, gentamicin, <CIP, ciprofloxacin MIC 0.06-0.25 mg/L; >CIP, ciprofloxacin MIC >0.5 mg/L, QRDRs, multiple mutations in the quinolone resistance-determining regions of *gyrA* and *parC*, PMQR, plasmid-mediated quinolone resistance gene.

Figure 4.8 provides an overview of the association between specific travel destinations and *Salmonella* pheno- and genotypes. Only a few attributes were linked to West Africa, East and West Asia, the Americas and Europe but North Africa and South and Southeast Asia were identified as resistance hotspots. Decreased ciprofloxacin susceptibility ($p=4.1x10^{-13}$), resistance to cephalosporins ($p\leq7.64x10^{-3}$), gentamicin ($p=8.63x10^{-7}$), tobramycin ($p=3.79x10^{-10}$), trimethoprim ($p=5.42x10^{-8}$) and chloramphenicol ($p=5.54x10^{-3}$), and carriage of the corresponding genetic resistance determinants ($p\leq3.03x10^{-4}$), were significantly associated with travel to North Africa. Ertapenem (p=0.02), gentamicin ($p=4.73x10^{-4}$), trimethoprim (p=0.03) and

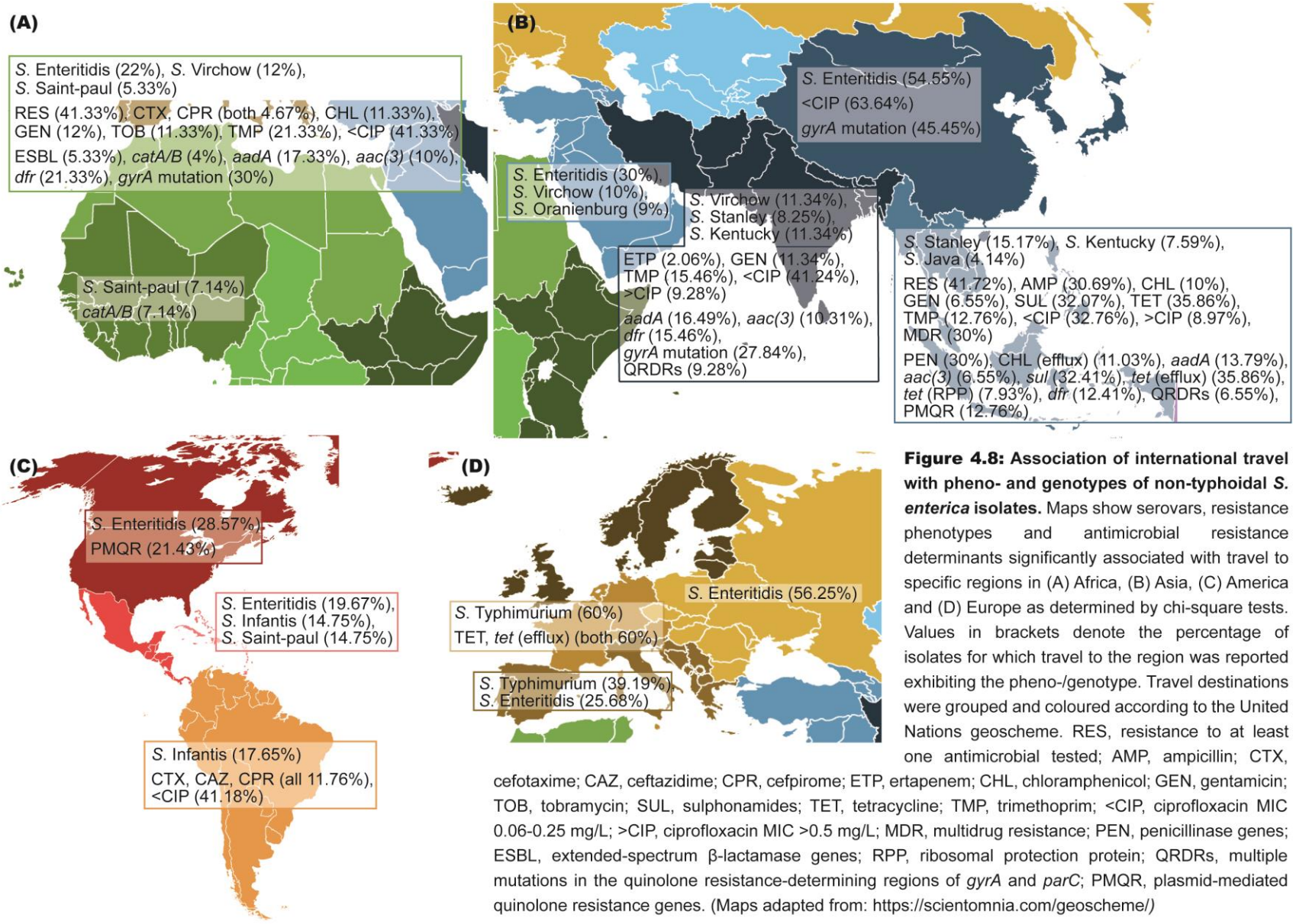

Figure 4.8: Association of international travel with pheno- and genotypes of non-typhoidal *S. enterica* isolates. Maps show serovars, resistance phenotypes and antimicrobial resistance determinants significantly associated with travel to specific regions in (A) Africa, (B) Asia, (C) America and (D) Europe as determined by chi-square tests. Values in brackets denote the percentage of isolates for which travel to the region was reported exhibiting the pheno-/genotype. Travel destinations were grouped and coloured according to the United Nations geoscheme. RES, resistance to at least one antimicrobial tested; AMP, ampicillin; CTX, cefotaxime; CAZ, ceftazidime; CPR, cefpirome; ETP, ertapenem; CHL, chloramphenicol; GEN, gentamicin; TOB, tobramycin; SUL, sulphonamides; TET, tetracycline; TMP, trimethoprim; <CIP, ciprofloxacin MIC 0.06-0.25 mg/L; >CIP, ciprofloxacin MIC >0.5 mg/L; MDR, multidrug resistance; PEN, penicillinase genes; ESBL, extended-spectrum β-lactamase genes; RPP, ribosomal protection protein; QRDRs, multiple mutations in the quinolone resistance-determining regions of *gyrA* and *parC*; PMQR, plasmid-mediated quinolone resistance genes. (Maps adapted from: https://scientomnia.com/geoscheme/)

ciprofloxacin resistance (p=0.01) were more common in isolates linked to South Asia and ampicillin ($p=8.14x10^{-5}$), chloramphenicol ($p=2.29x10^{-3}$), sulphonamide ($p=6.28x10^{-4}$), tetracycline ($p=1.4x10^{-4}$), trimethoprim (p=0.02) and ciprofloxacin resistance ($p=9.98x10^{-6}$) were frequently observed after travel to Southeast Asia. Southeast Asia was the only travel destination significantly associated with MDR ($p=4.45x10^{-3}$). The Bonferroni-corrected p-value for these tests was $5.72x10^{-5}$.

4.6 Discussion

4.6.1 Antimicrobial resistance and its genetic determinants in non-typhoidal *S. enterica*

Most *S. enterica* isolated in European countries belong to a limited number of serovars[255]. In England and Wales in 2012, *S*. Typhimurium and *S*. Enteritidis made up 48.7% of all isolates[318]. GBRU's routine phenotypic testing strategy for surveillance of NTS attempts to maximise the detection of AMR by focussing on serovars known to have high resistance rates. This leads to an underrepresentation of some, such as *S*. Enteritidis, and an overrepresentation of other serovars, such as *S*. Typhimurium and *S*. Kentucky, in the dataset presented here. Thus, it is not a true representation of the expected serovar distribution in England and Wales over the timeframe investigated, which made it impossible to meaningfully assess changes in incidence of resistance to specific antimicrobials over the years.

Pansusceptibility to the phenotypic testing panel was observed in 68.98% of isolates in this dataset. Worryingly though, of the resistant isolates, over 70% were MDR, i.e. resistant to three or more antimicrobial classes. Rates of MDR were especially high in serovars Typhimurium, Kentucky and Virchow. The most frequent resistance phenotypes were tetracycline (26.27% of isolates), sulphonamide (23.72%) and ampicillin (21.43%). These trends are comparable to those observed for the entire EU in 2014 and 2015[295,319]. The ACSSuT phenotype was identified in 1.52% of isolates making it the third most common MDR phenotype after co-resistance to ampicillin, streptomycin, sulphonamides and tetracycline (6.99%) and co-resistance to sulphonamides, tetracycline and trimethoprim (1.89%). The genetic determinants underlying these phenotypes are located on a mobile element termed *Salmonella* genomic island 1[320], thereby explaining their linked carriage and detection (cf. **Fig. 4.6**). It has been suggested that an increased use of alternative antimicrobials for treatment of typhoid fever, such as ciprofloxacin and extended-spectrum β-lactams, favoured the re-emergence of susceptibility to classical first-line drugs[321,322], thus explaining the observed worldwide decrease in the number of isolates displaying the ACSSuT phenotype[323,324].

Full resistance to ciprofloxacin was detected in 3.95% of isolates and particularly prevalent in *S*. Infantis and *S*. Kentucky. An increased number of mutations in *gyrA* and *parC*, which reduce the binding affinity of ciprofloxacin, are associated with an increase in MIC[273]. Therefore, in order to confer full resistance, multiple mutations in both *gyrA* and *parC* were thought to be required[316]. Carriage of PMQRs on their own, which prevent interactions between the antimicrobial and topoisomerases by sterical hindrance, were not considered sufficient. Indeed, in our study, the majority of isolates showing resistance carried at least two mutations in both *gyrA* and *parC*. However, 37 had a PMQR gene, alone or in conjunction with a single *gyrA* mutation, which would normally be expected to result in reduced susceptibility instead of full resistance. Ciprofloxacin MICs for isolates carrying only PMQR genes were found to range between 0.25 and 1 mg/L[325] so that some isolates with this profile would be classed as resistant and some as having reduced susceptibility during phenotypic testing.

Although still relatively low, resistance to extended-spectrum cephalosporins is on the rise[326,327]: while only 0.04% of S. enterica isolates in England and Wales showed resistance between 1992 and 2003[328], its prevalence had increased to 0.55% by 2010-2012[329]. In this study, 1.63% and 1.35% of isolates exhibited resistance to the 3rd-generation cephalosporins cefotaxime and ceftazidime, respectively, and 1.35% were resistant to the 4th-generation cephalosporin cefpirome. Resistance to these antimicrobials is encoded by AmpC and ESBL genes, detected in 0.46% and 1.23% of isolates, respectively. Since extended-spectrum cephalosporins are used as an alternative for treatment of invasive disease in cases of resistance to ciprofloxacin, the emergence of co-resistance to both antimicrobial classes is of great concern. Co-resistance was identified in 13 isolates (0.37%) in this study, which is in agreement with EU data from the year 2015[295] and slightly higher than the 0.25% prevalence observed in the UK between 2010 and 2012[329].

Due to this emerging co-resistance, the use of azithromycin for treating NTS infections has increased in recent years, also prompting the surveillance of resistance to this antimicrobial. The first EU-wide reports in 2017 identified azithromycin resistance in 2.5% of NTS isolates[330]. Although not included in the phenotypic testing panel, the ‘Genefinder’ algorithm was able to detect genetic

azithromycin resistance determinants. Twenty-two isolates carried *mph*(A). All 22 were phenotypically resistant to at least one other antimicrobial, seven showed ciprofloxacin resistance and four were resistant to extended-spectrum cephalosporins. The *mph*(E) gene was detected in seven isolates, only one of which was resistant to ciprofloxacin. Validation of predicting phenotypic azithromycin resistance based on the presence of these genes would further aid surveillance efforts for this antimicrobial.

4.6.1.1 Prevalence of antimicrobial resistance and its determinants in travel-related isolates

Based on the travel history data available, North Africa as well as South and Southeast Asia were identified as resistance hotspots (cf. **Fig. 4.8**). Phenotypes more frequently observed after travel to these regions included extended-spectrum cephalosporin, chloramphenicol, sulphonamide, tetracycline, trimethoprim and ciprofloxacin resistance. Co-resistance to ciprofloxacin and extended-spectrum cephalosporins was shown to be especially prevalent in Asia with 9.3% of NTS isolates between 2003 and 2005 exhibiting this phenotype[331]. Of the 13 co-resistant isolates in this study, one was associated with travel to Thailand and one with travel to Egypt. Travel to Asia was recorded for seven of the 43 isolates carrying ESBL genes and travel to North Africa for a further eight isolates. As observed previously by Hopkins et al.[332], PMQR genes were more likely to be found in isolates from patients who had travelled to Southeast Asia.

Furthermore, in this study, Southeast Asia was found to be the only travel destination significantly associated with MDR. Burke et al.[329] identified resistance to at least five antimicrobial classes in all NTS isolates, for which travel to Thailand was recorded. Extensively-drug resistant *S*. Typhimurium, like the one isolate in this study showing resistance to eight antimicrobial classes plus decreased susceptibility to ciprofloxacin, have been observed in Southeast Asia before[333,334]. Unfortunately, no travel history data was available for this isolate.

Since clinicians are aware of the increased likelihood of acquiring MDR NTS infections during travel, azithromycin treatment is recommended especially for returning travellers. However, travel was reported for eight of the 22 isolates in our

study carrying the *mph*(A) gene, highlighting how alternative antimicrobials might become inefficient when overused.

4.6.2 Feasibility of genome-based prediction of antimicrobial resistance profiles

WGS has previously been employed successfully for prediction of AMR profiles in a variety of enteropathogens, including *S. sonnei* [274], *E. coli* [335-337], typhoidal *Salmonellae*[273] and smaller datasets of NTS[293,338,339]. The comparison of phenotypic susceptibility testing and genotypic prediction of AMR profiles based on WGS data presented here for a much larger dataset, comprising 3,491 NTS isolates, identified 88 discordant results (0.17%) out of a possible 52,365 isolate/antimicrobial combinations, with the AMR profiles of 3,415 isolates (97.82%) completely matching for both approaches. Zankari et al.[338] observed complete agreement of the two approaches for 50 *S*. Typhimurium isolates but only when excluding ciprofloxacin from the testing panel. Similar to the results of the present study, McDermott et al.[339] found lower sensitivity and specificity for prediction of streptomycin resistance than for other antimicrobials tested.

Many MEs, where an isolate is expected to be resistant based on the presence of genetic determinants but shows phenotypic susceptibility, seem to be associated with the clinical breakpoints used for phenotypic testing. Slight technical variations of the agar dilution method for isolates with an MIC close to the recommended breakpoint can lead to their false classification as susceptible. This was found to especially be an issue when testing for streptomycin resistance[340] and could explain some of the 51 MEs observed for this antimicrobial. To overcome this, it has recently been suggested to integrate phenotypic and genetic data by adapting breakpoint values based on the MICs associated with specific ARDs[341]. A caveat of this study was that phenotypic susceptibility testing was carried out retrospectively. Most of the ARDs detected by the algorithm are plasmid-encoded and during storage and subculture of isolates plasmids may be lost. Genes detected when sequencing the original culture might therefore not be present during retrospective phenotypic testing of a different colony. Furthermore, the mere detection of a gene does not always reflect an organism's phenotype as it does not provide any information about expression levels

and enzyme activities or the interplay between different factors involved in resistance. Silent resistance genes, such as *bla*$_{CMY-2}$ and *tet* variants, have previously been described in *Salmonella*[342,343].

VMEs, where an isolate is genotypically predicted to be susceptible but exhibits phenotypic resistance, highlight the importance of active curation of the ARD reference database used for prediction. Mismatches are likely based on the presence of ARDs not included in the database or on novel resistance mechanisms, the genetic determinants of which have not yet been described. The 'Genefinder' pipeline, for example, does not detect impermeability or efflux pump genes potentially contributing to ciprofloxacin resistance[344]. Continuous scanning of the literature for newly discovered resistance mechanisms and their incorporation into the database are essential to maintain a high level of prediction sensitivity. The increasing number of genomes available for analysis can be expected to be accompanied by the discovery of new genes. Only recently, computational methods identified previously unknown *qnr*-type PMQRs, six of which could be experimentally validated as conferring reduced ciprofloxacin susceptibility[345]. At the time this study was conducted, colistin ARDs were not included in the database at all. However, considering that colistin use in England is higher than in other European countries[346], resistance to this antimicrobial has to be monitored. Since the first reports of the plasmid-encoded *mcr-1* gene, seven further *mcr* variants have been described[345] and a new homologue was recently discovered by Carroll et al.[347] through *in silico* screening of *Salmonella* genomes.

Despite the issues mentioned above, the overall ME and VME rates of 0.13% and 0.04%, respectively, obtained in this study fall below the cut-offs of 3% and 1.5% from the US Food and Drug Administration for authorising new susceptibility testing devices[348]. Nonetheless, AMR prediction based on WGS data is not yet deemed suitable to guide clinical decision making. Apart from the cost and the infrastructure necessary to implement the service routinely, which would be too much for many front-line laboratories, the high turnaround times are the biggest obstacle[349]. Faster diagnosis is associated with improved patient outcomes but at PHE it takes seven days from receipt of the isolate to completion of all bioinformatics analyses. For AMR

surveillance, however, where slightly longer turnaround times are permissible, WGS constitutes a valuable alternative to traditional methods.

4.7 Conclusions

This large-scale study demonstrates that AMR and MDR are ubiquitous in UK isolates of NTS, especially in returning travellers. This highlights the need for continuous surveillance of trends in AMR patterns to inform public health interventions aimed at minimising the dissemination of resistance in the population. The successful prediction of AMR profiles based on detection of ARDs in the genome presented here supports the suitability of WGS profiling to reliably replace phenotypic susceptibility testing for rapid monitoring of AMR in NTS. Since sequencing is routinely used in public health laboratories already, it constitutes a time-saving alternative to traditional approaches that can further our understanding of resistance mechanisms as long as constant curation of the resistance gene database used is warranted and validation of prediction for further antimicrobials, such as azithromcyin and colistin, is undertaken.

WGS-based AMR profiling, while eliminating laborious AST, still requires an initial culturing step. The use of culture-independent metagenomic sequencing for investigation of AMR dissemination networks will be explored in the next chapter.

CHAPTER FIVE

---

5. Prevalence and persistence of changes in resistome and taxonomic composition in the gut microbiota of returning travellers: the GutBack project

5.1 Summary

Antimicrobial resistance (AMR) is on the rise even in industrialised countries, not just in healthcare but also in community settings. International travel is expected to play an important part in the dissemination of resistance genes but most of the data supporting this theory is based on surveillance of infectious disease, thus ignoring the potential carriage of antimicrobial resistance determinants (ARDs) in the gut microbiota of returning travellers who do not seek healthcare. This chapter describes the results of the GutBack study, a prospective, longitudinal investigation of changes in gut microbiota resistome and taxonomic composition in 48 UK travellers visiting regions with a high prevalence of AMR. A 1.27-fold increase in the sum of all individual ARDs detected in the cohort was observed after travel. Beta-lactam, particularly *bla*$_{CTX}$ (log_2(FC)=2.86) and *bla*$_{TEM}$ genes (log_2(FC)=2.32), and sulphonamide ARDs (log_2(FC)=2.41) as well as chloramphenicol acetyltransferase genes (log_2(FC)=1.86) were more abundant in post-travel compared to pre-travel samples. Carriage rates of *Escherichia*, *Klebsiella*, *Salmonella* and *Shigella* were significantly higher after travel. Changes in sulphonamide ARDs, *bla*$_{CTX}$ genes and increased carriage rates of potentially pathogenic Enterobacteriaceae persisted up to six months after travel. Antimicrobial use was associated with a shift in overall resistome, traveller's diarrhoea (TD) with a shift in overall taxonomic composition.

5.2 Introduction

5.2.1 Impact of international travel on gastrointestinal health

The decreased cost and higher availability of airline travel has made international journeys accessible to a wide demographic[350]. In 2015, 1.2 billion people travelled internationally[351], with numbers expected to reach 1.8 billion by 2030[352]. Per year, an estimated 320 million people from industrialised countries visit developing regions[351]. Exposure to a foreign environment harbouring different microorganisms, as well as changed diet and sleeping patterns, have been shown to alter gut microbiota composition, even in the absence of ill health[353,354]. These changes can be exacerbated by acquisition of an infectious disease[355]. Varying based on destination, duration of travel and the season, 6-87% of people report travel-related illness[356].

An imbalanced gut microbiota during travel might lead to reduced colonisation resistance. At the same time, travellers are often exposed to food and water contaminated with pathogens due to lower hygiene and sanitation standards. TD is therefore unsurprisingly the most common infectious disease acquired abroad. Aetiological agents of TD are primarily bacterial. While overall, it is most frequently caused by enterotoxigenic (ETEC) and enteroaggregative *Escherichia coli*[357] (EAEC), the prevalence of TD pathogens varies between travel destinations: *Campylobacter* and *Salmonella* infections, for example, are more common in South and Southeast Asia, and *Shigella* infections are often found in travellers returning from the Middle East or Africa[357,358]. Travel to Asia has also been linked to protozoal TD caused by pathogens such as *Giardia* and *Entamoeba histolytica*[357].

5.2.2 Acquisition of antimicrobial resistance during travel

Pathogens isolated in developing regions show high rates of AMR, generally attributed to the uncontrolled use of antimicrobials in humans and animals due to the lack of prescription regulations implemented in most industrialised countries[359,360]. Multidrug-resistant (MDR) Enterobacteriaceae in particular are common in tropical regions[361]. The family comprises many of the species frequently causing TD, and the production of extended-spectrum β-lactamases (ESBLs) by these organisms is of

great concern as it enables them to hydrolyse most clinically relevant β-lactam antimicrobials. Infections caused by ESBL-producing Enterobacteriaceae (ESBL-E) are associated with a poor prognosis and higher mortality rates[362,363]. ESBL-E have been isolated from vegetables, sewage and the faeces of livestock in developing countries[364]. Consequently, community carriage of ESBL-E is high in these regions, estimated at 20-70% in Asia, 10-15% in South America and 10-40% in Africa[361,365]. Specific countries, such as India, Thailand and Egypt, are considered hotspots for ESBL-E in the community[361,365,366].

In England, the estimated community carriage of ESBL-E of 7.3% is still comparably low[367]. However, the rise in carriage rates observed even in industrialised countries is accompanied by an increased risk of community-acquired infections[361,363,366]. Colonisation abroad and subsequent import of resistant microorganisms to travellers' home countries have undoubtedly played a part in this development. Acquisition rates of 14-69% for MDR Enterobacteriaceae were reported for travellers to tropical regions[368,369]. Numerous studies from, for example, Australia[370], France[371], Germany[372], the Netherlands[373,374] and Sweden[369,375,376] demonstrated acquisition of ESBL-E or CTX-M-type ESBL genes during travel. In line with community carriage rates, the risk of acquisition was highest for individuals who visited the Indian peninsula followed by destinations in Southeast Asia[368,369,372-374,377,378]. Gut microbiota dysbiosis induced by TD and antimicrobial use were the most frequently identified additional risk factors[369,370,375,377,379].

5.2.3 Dissemination of acquired antimicrobial resistance after travel

Despite its association with TD, acquisition of resistant organisms is not restricted to individuals who experience ill health. Although most of our understanding of the link between travel and AMR acquisition is based on surveillance data from patients accessing healthcare, travellers can become asymptomatic carriers, thus constituting a reservoir for future infections of the host or other individuals, which is seldom detected before illness occurs. The potential of unsuspecting travellers spreading acquired resistance determinants was further illustrated by Nordahl Petersen et al.[350], who found a higher abundance of ARDs, including *bla*$_{\text{CTX-M}}$ and other β-lactamase

genes, in toilet waste from international planes arriving from Asia compared to those coming from North America.

Onward transmission of acquired ESBL-E to household members, both co-travellers who tested negative immediately after return and non-travellers, has been demonstrated[368,373], proving the potential of AMR dissemination in the community. The longer an individual remains colonised with the resistant organism, the higher the chance for transmission to occur. Although longitudinal studies have shown that most individuals clear acquired MDR Enterobacteriaceae within a month[368], sustained carriage of up to six months is not uncommon[369,370,373] and colonisation persisting for up to one year has been observed[368,380].

The spread of acquired resistance is further facilitated by the existence of ARDs on mobile elements, which can be horizontally transferred to other organisms. Many previous studies focussed on culture- or PCR-based detection of resistance in Enterobacteriaceae, which normally only make up a small amount of an individual's gut microbiota[9,37], and ignored ARDs harboured by commensals or non-pathogenic, environmental bacteria acquired abroad. If they persist, these ARDs can be transferred to pathogenic organisms during future infections, thus limiting treatment options. A metagenomics study of UK travellers encompassing the entirety of the gut microbiota and its resistance potential outside of a healthcare context could therefore further our understanding of the dissemination of ARDs in the community.

5.3 Objectives

The purpose of the work presented in this chapter was to conduct a prospective longitudinal study investigating changes in gut microbiota taxonomic composition and its pool of ARDs in a cohort of UK travellers visiting areas with high prevalence of AMR in the community. It was hypothesised that travel leads to an enrichment of certain ARDs and taxa, regardless of health status, and that some of the changes persist up to six months after travel.

5.4 Materials and Methods

Participant recruitment and sample collection were undertaken by staff at University College London's (UCL) Hospital for Tropical Diseases. DNA extractions were carried out by the author. Library preparation, metagenomic sequencing, quality control and taxonomic classification were performed by MicrobesNG (University of Birmingham, Birmingham, UK). The pipeline for detection of ARDs was run by Dr. Tim Dallman at Public Health England (PHE). Statistical analysis of the data was performed by the author. Unless otherwise stated, default software parameters were used.

5.4.1 Participant recruitment

Ethical approval for the study was obtained from the National Health Service Research Ethics Committee. Recruitment took place between December 2015 and July 2016. Individuals aged 18 years or over planning travel to countries outside of Europe, North America and Australia for two weeks to three months, who attended the pre-travel outpatient clinic at UCL's Hospital for Tropical Diseases, were invited to participate in the study. Additional recruitment of prospective travellers was carried out via advertisement of the study on the local university campus. Potential participants were provided with a participant information leaflet in person or by post. Written, fully informed consent was obtained from all individuals who agreed to take part and did not meet any of the following exclusion criteria:

- travel outside of Europe, North America and Australia three months prior to recruitment or planned travel within six months after return.
- antimicrobial use at the time of or in the two weeks prior to recruitment.
- age under 18 years.

Participants were assigned a unique identification number for data anonymisation.

5.4.1.1 Collection of faecal samples

Participants provided a faecal sample before travel (pre-travel), within two weeks after travel (post-travel) and around six months after travel (follow-up) to assess long-term persistence of changes. Samples were stored at -80°C in the local laboratory

and sent to PHE's Gastrointestinal Bacteria Reference Unit (Colindale, UK) in batches. Upon receipt at PHE, samples were aliquoted and stored at -20°C until further processing.

5.4.1.2 Questionnaires

Along with the provision of faecal samples, participants completed a questionnaire at the three different timepoints. Data collected included information on demographics, participant health, medication, region and duration of travel and travel-related behaviour.

5.4.2 Processing of faecal samples

5.4.2.1 Genomic DNA extraction

Total genomic DNA was extracted from participants' faecal samples using a combination of mechanical and enzymatic lysis. For negative controls, all the following steps were undertaken without addition of a sample. Incubation of 200 mg of sample with 2 ml Gibco™ phosphate-buffered saline (PBS), pH 7.4 (Thermo Fisher Scientific, Waltham, MA, USA) and 2 ml 2% 2-mercaptoethanol (Sigma-Aldrich, St. Louis, MO, USA) was carried out for one hour on a Stuart tube rotator (Cole-Parmer, Vernon Hills, IL, USA). After centrifugation at 6,000 x g for 10 min, the pellets were resuspended in 10 ml PBS by vortexing. To remove particulate matter, the samples were first sterile-filtered through a 100 µm, then through a 40 µm Steriflip centrifuge tube top filter unit (Merck Millipore, Burlington, MA, USA) with the help of a vacuum pump. Samples were centrifuged at 6,000 x g for 10 min and the pellet resuspended in 1 ml PBS.

The following steps were performed using DNA LoBind tubes (Eppendorf, Hamburg, Germany). Host DNA depletion was carried out by selective lysis of human cells and subsequent DNase treatment using the MolYsis kit (Molzym, Bremen, Germany). Unless stated otherwise, reagents were obtained from this kit. Samples were incubated with 200 µl buffer CM for 5 min at room temperature. After centrifugation at 6,000 x g for 10 min, the pellets were resuspended in 1 ml Gibco™ Hank's Balanced Salt Solution with calcium and magnesium (Thermo Fisher

Scientific) and 250 µl DNase buffer and 10 µl DNase. The mixture was incubated at 25°C for 15 min, centrifuged at 6,000 x g for 10 min and the pellets resuspended in 1 ml buffer RS. This was followed by another centrifugation at 6,000 x g for 10 min. The pellets were resuspended in 100 µl PBS and incubated at 65°C for 5 min.

Enzymatic lysis of the remaining cells was carried out using the MasterPure purification kit (Epicentre, Madison, WI, USA). Unless stated otherwise, reagents were taken from this kit for the remainder of the extraction protocol. Additional lytic enzymes were obtained from Sigma-Aldrich. The samples were incubated with 150 µl 2x tissue and cell lysis buffer, 2 µl lysozyme (10 mg/ml), 6 µl mutanolysin (20 U/µl) and 3 µl lysostaphin (8 µg/µl) at 37°C for one hour in a ThermoMixer® (Eppendorf). This was followed by a 30-minute ThermoMixer® incubation at 65°C with 2 µl proteinase K.

For mechanical lysis, the samples were transferred to tubes containing acid-washed glass beads 212-300 µm in size along with an additional 50 µl 2x tissue and cell lysis buffer. Bead beating was carried out with a FastPrep-24™ instrument (MP Biomedicals, Santa Ana, CA, USA). Samples were pulsed twice at 6.0 m/s for 20 s followed by one pulse of 10 s with 2-minute intervals between pulses. After a 10,000 x g centrifugation for 10 min at 4°C, lysates were separated from cell debris and allowed to cool on ice for 2 min. Samples were kept on ice from this step onwards whenever possible. For protein precipitation, 180 µl MPC were added. Samples were vortexed for 15 s and centrifuged at 15,000 x g and 4°C for 10 min. For DNA precipitation, the supernatants were mixed with 500 µl of isopropanol (Sigma-Aldrich) by inversion and centrifuged at 15,000 x g and 4°C for 10 min. Pellets were washed twice with 200 µl 70% ethanol (Sigma-Aldrich) and resuspended in 100 µl TE buffer.

For clean-up, samples were transferred to OneStep PCR inhibitor removal kit columns (Zymo Research, Irvine, CA, USA) and centrifuged at 8,000 x g for 1 min. Quantification of eluted DNA was carried out using a Qubit fluorometer with high-sensitivity reagents (Thermo Fisher Scientific). Extracts were stored at -20°C until sequencing.

5.4.2.2 Metagenomic sequencing

Metagenomic DNA libraries were generated using the Nextera XT kit (Illumina®, San Diego, California) at MicrobesNG. Paired-end 125 bp sequencing was performed on a HiSeq instrument (Illumina®) in high-output mode.

5.4.3 Bioinformatics workflow

5.4.3.1 Detection of antimicrobial resistance determinants

Detection of ARDs in the sequencing data was carried out using AmrPlusPlus, version 1.1 (https://github.com/cdeanj/amrplusplus)[381]. Trimming of paired-end reads and removal of adapter contamination was performed with Trimmomatic, version 0.36[253]
with the following parameters: ILLUMINACLIP 2:20:10:3:TRUE, LEADING 3, TRAILING 3, SLIDINGWINDOW 4:15, MINLEN 36. Contaminant host DNA was identified by alignment of trimmed reads to human chromosome 21 with the Burrows-Wheeler-Aligner (BWA)[313] and removed with SAMtools[261]. Remaining reads were aligned to the MEGARes database, version 1.01, a non-redundant compilation of ARD sequences from Resfinder, ARG-ANNOT, CARD (The Comprehensive Antibiotic Resistance Database) and the National Centre for Biotechnology Information's Lahey Clinic β-lactamase archive using BWA. After sorting, the resulting BAM file was converted to a SAM file with SAMtools. The threshold for positive detection of an ARD in the sequencing data was set to 80%, i.e. ≥80% of nucleotides in the reference sequence had to be covered by at least one read.

AmrPlusPlus provides a hierarchical output of the ARDs identified as outlined in **Figure 5.1**. Read counts assigned to individual ARDs are first added up as gene- or operon-level groups. ARD groups are then further classified based on the biological mechanism by which they confer resistance. Finally, resistance mechanisms are grouped together based on the antimicrobial class they confer resistance to. Groups, mechanisms or classes present in less than ten samples were removed prior to statistical analysis.

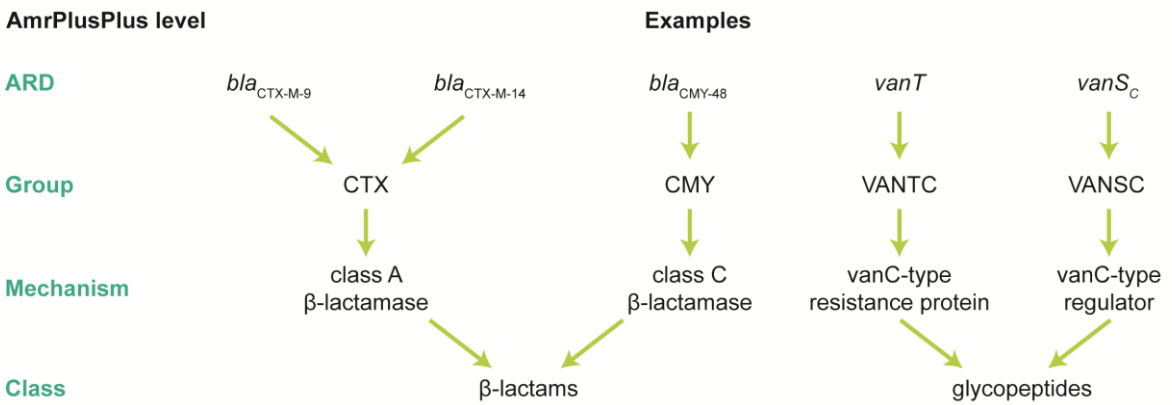

Figure 5.1: Overview of the AmrPlusPlus pipeline output. ARDs conferring resistance to β-lactams and glycopeptides are used as examples to illustrate the hierarchical output. Reads assigned to individual ARDs are first added up at the group level, then at the mechanism level and finally at the class level.

5.4.3.2 Taxonomic analysis

Taxonomic classification of sequences was performed using Kraken, v. 1.0[260] as described in section 3.4.3.3. A table with relative abundances of operational taxonomic units (OTUs) was generated using the 'feature-table relative-frequency' command of the QIIME 2 package (v. 2018.8)[173].

5.4.4 Statistical analysis

5.4.4.1 T-statistics

Differences in the number of detected ARDs or OTUs between two groups were assessed using the t.test function in R choosing the paired test option when comparing samples between timepoints. A p-value ≤0.05 was considered statistically significant.

5.4.4.2 Normalisation of sequencing reads

Sequencing reads assigned to individual ARDs or OTUs were aggregated at the desired level of the AmrPlusPlus output or taxonomy using the metagenomeSeq package v. 1.26.3 in R[187]. Read counts were normalised with the help of the package's cumNorm function, which calculates scaling factors (SFs) equal to the sum

of counts up to a specified quantile. The 0.5 quantile (i.e. median) was selected for this analysis.

5.4.4.3 Ordination

Constrained correspondence analysis (CCA) on aggregated, normalised read counts was performed using the vegan package v. 2.5-6 in R[382]. When appropriate, TD, antibiotic and antimalarial use during initial travel, antibiotic use between initial travel and provision of the third sample, repeat travel and TD during repeat travel were added to the model as covariates.

Permutation tests for the significance of constraints were carried out using the package's anova function with the number of permutations set to 999 and separate analysis of terms. A p-value ≤0.05 was considered statistically significant.

5.4.4.4 Tests of association

Differences in carriage rates between timepoints were assessed using the mcnemar.test, differences in carriage and acquisition rates between groups at the same timepoint using the chisq.test function in R. A p-value ≤0.05 was considered statistically significant. Bonferroni correction for multiple testing was performed by dividing this threshold p-value by the number of tests.

5.4.4.5 Differential abundance testing

Differences in specific features at the desired level of AmrPlusPlus output or taxonomy were estimated by creation of zero-inflated Gaussian mixture models as described in Paulson, 2016[187]. A design matrix was generated from the aggregated, normalised read count data with the model.matrix function, adding TD, antibiotic and antimalarial use during initial travel, antibiotic use between initial travel and provision of the third sample, repeat travel and TD during repeat travel as covariates when appropriate. Furthermore, normalisation factors, derived from the SFs described in section 5.4.4.2, were added to the matrix. Normalisation factors were calculated as follows:

$$\log_2(SF/\text{median}(SF) + 1)$$

The fitZig function was used to produce weighted fits for all features with the previously generated design matrix as input. The useCSSoffset argument was set to FALSE and the control parameter was defined via the zigControl function. The maximum number of iterations was set to 10 and the dfMethod to default. From the resulting linear model fits, fold change (FC) estimates and t-statistics were computed using the limma package's (v. 3.40.6) eBayes function[188]. An adjusted p-value ≤0.05 was considered statistically significant.

5.5 Results

5.5.1 Study participants

A total of 107 participants were recruited. Sample donation and questionnaires at all three timepoints were completed by 76 of the 107 initial recruits (71.03%). Stools of the first 50 participants, for which complete sample sets were received at PHE, were prepared for sequencing. All further results refer to these 50 sample sets.

5.5.1.1 Demographics and pre-travel health

Information on sex, age, birthplace, presence of chronic health conditions and use of regular medication for the 50 study participants are shown in **Table 5.1**. Median age was 41.5 years, ranging from 22 to 79 years.

The 11 extraintestinal chronic health complaints included asthma, hypertension, psoriasis, cancer, diabetes, human immunodeficiency virus and osteoporosis, and the six gastrointestinal ones gastric ulcers, inflammatory bowel disease and irritable bowel syndrome. Tumour necrosis factor α and protein pump inhibitors, steroids, statins and antiretrovirals were amongst the regular medications prescribed to the participants.

Table 5.1: Demographics and pre-travel health of study participants.

	Travellers (n=50)
Sex	
Female	34 (68%)
Male	16 (32%)
Age	
22-30	12 (24%)
31-50	17 (34%)
51-64	7 (14%)
≥65	14 (28%)
Birthplace	
UK	27 (54%)
Europe (other)	7 (14%)
North America	6 (12%)
Australia	1 (2%)
East/West Africa	4 (8%)
South/Southeast Asia	3 (6%)
South America	1 (2%)
Caribbean	1 (2%)
Medical conditions	
No	33 (66%)
Gastrointestinal	6 (12%)
Extraintestinal	11 (22%)
Pregnancy	3 (6%)
Regular medication	
No	33 (66%)
Yes	17 (34%)

5.5.1.2 Travel information and health

Information on destination and duration of travel, travel-related illness and medication used during travel can be found in **Table 5.2**. Travel destinations in East Africa were Kenya, Tanzania, South Sudan, Malawi, Zimbabwe, Zambia, Uganda, Mozambique and Ethiopia, in Southern Africa South Africa, Botswana and Namibia, in West Africa Ghana, Nigeria and The Gambia, in the Caribbean Saint Kitts and Nevis, in Central America Honduras, Nicaragua, El Salvador, Mexico, Belize, Guatemala and Costa Rica), in South America Brazil, Argentina, Ecuador, Peru and Bolivia, in Eastern Asia China, in South Asia India and Sri Lanka and in Southeast Asia Myanmar, Laos,

Table 5.2: Travel details, ill health and drug usage in study participants.

	Travellers (n=50)
Travel destination	
Africa	19 (38%)
East Africa	14 (28%)
Southern Africa	7 (14%)
West Africa	5 (10%)
America	15 (30%)
Caribbean	1 (2%)
Central America	4 (8%)
South America	10 (20%)
Asia	16 (32%)
Eastern Asia	1 (2%)
South Asia	6 (12%)
Southeast Asia	10 (20%)
Travel duration	
2 weeks	11 (22%)
>2 weeks - ≤1 month	31 (62%)
>1 month - ≤2 months	3 (6%)
>2 months	5 (10%)
Illness during travel	
No	22 (44%)
Diarrhoea	23 (46%)
Nausea & vomiting	4 (8%)
Fever	4 (8%)
Other	6 (12%)
Antimicrobial usage	
No	38 (76%)
Azithromycin	3 (6%)
Ciprofloxacin	9 (18%)
Malaria prophylaxis	
No	29 (58%)
Atovaquone/proguanil	19 (38%)
Chloroquine	1 (2%)
Mefloquine	1 (2%)

Travel destinations were grouped according to the United Nations geoscheme. Where the number of travellers visiting a continent's subregion does not add up to the number of travellers visiting the continent, participants travelled to more than one subregion. Where numbers do not add up to 50 in the "Illness during travel" section, participants suffered from more than one health complaint.

Vietnam, Thailand, the Philippines and Cambodia. Travel across several countries was undertaken by 12 participants. Three of these visited two geographical subregions but all remained on the same continent. Median travel duration was 18.5 days, ranging from 14 to 78 days.

Symptoms lasted less than a week in most individuals affected by TD (n=21) but one participant reported a duration of ten days and another one a duration of 42 days. Four participants who suffered from TD additionally experienced fever, a further four nausea and vomiting, one muscle cramps and one flu-like symptoms. Further health complaints in the absence of TD included a sore throat, constipation and coughs. Of the twelve participants using antimicrobials, nine took them for a single day, one completed a two-day, one a three-day and a further one a six-day course. All but one participant who used antimicrobials suffered from TD.

5.5.1.3 Travel-related behaviour

Reasons for travel as well as eating and drinking habits of the 50 study participants are illustrated in **Table 5.3**.

Table 5.3: Study participants' behaviour during travel.

	Travellers (n=50)
Reason for travel	
Business	5 (10%)
Leisure	29 (58%)
Healthcare work	6 (12%)
Visiting friends/relatives	10 (20%)
Drinking water	
Bottled only	43 (86%)
Various sources	7 (14%)
Alcohol consumption	
None	16 (32%)
0-2 units/d	29 (58%)
>3 units/d	5 (10%)
Main site of meals	
Household	21 (42%)
Restaurants	26 (52%)
Streetfood	3 (6%)
Consumption of uncooked meat/fish	
No	44 (88%)
Yes	6 (12%)
Probiotics	
No	43 (86%)
Yes	7 (14%)

5.5.1.4 Post-travel health and behaviour

Information on antimicrobial usage and spontaneous repeat travel before provision of the six-month follow-up sample, as well as on ill health and use of malaria prophylaxis in re-travellers, is shown in **Table 5.4**. The eight participants who took antimicrobials used azithromycin (n=1), trimethoprim (n=2), metronidazole (n=1), co-amoxyclav (n=2), nitrofurantoin (n=1), cephalexin (n=1) or amoxicillin (n=1). Ten of these repeat travellers visited countries in Africa, Asia and Central America. To prevent overall sample numbers from becoming too low, they were not excluded from the study, in spite of planned repeat travel initially being an exclusion criterion. To adjust for the effects of a second exposure to a foreign environment, repeat travel

was, however, added as a covariate to all statistical analyses involving follow-up samples.

Table 5.4: Study participants' health and behaviour after travel.

	Travellers (n=50)
Antimicrobial usage	
No	42 (84%)
Yes	8 (16%)
Repeat travel	
No	17 (34%)
Europe & North America only	22 (44%)
Africa	1 (%)
Asia	8 (16%)
Central America	2 (4%)
Illness during repeat travel	(n=33)
Diarrhoea	6 (18.18%)
Other	2 (6.06%)
Malaria prophylaxis during repeat travel	
No	31 (93.94%)
Atovaquone/proguanil	1 (3.03%)

5.5.2 Metagenomic sequencing

Per-sample read counts ranged from 27,814 to 28.04 million with a median of 3.98 million. One pre- and one post-travel sample had to be excluded. The samples were from two different participants, both of which re-travelled and one of which experienced TD and used antibiotics. All three samples from these participants were excluded from the following analyses, reducing the total number of complete datasets to 48.

5.5.3 Prevalence and persistence of resistome changes after travel

5.5.3.1 Resistome diversity

The impact of travel on resistome diversity was assessed by comparing the number of individual ARDs detected with the AmrPlusPlus pipeline at the three timepoints. The total number of different ARDs identified was 601. There was a FC of 1.27 in the

sum of all ARDs in post- compared to pre-travel samples. For follow-up compared to pre-travel samples, this FC was 1.22.

Twenty-seven participants (56.25%) carried a higher number of ARDs immediately after travel than before. For 15 of these 27 (55.56%), the number remained increased in the follow-up samples. The mean (±SD) number of individual ARDs per sample increased non-significantly from 52.67±26.47 (range: 7-127) before travel to 67.02±38.52 (range: 9-178) immediately after travel and remained elevated in the follow-up samples at 64.42±45.69 (range: 11-190) (**Figure 5.2**). Repeat travel has to be considered as a potential confounding factor in the apparently persistent increase in the number of ARDs six months after initial travel.

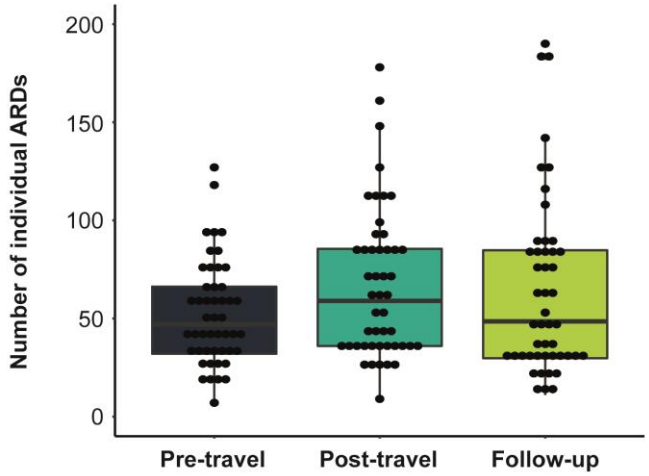

Figure 5.2: Changes in gut resistome diversity after travel. Number of individual antimicrobial resistance determinants (ARDs) detected in the stools of 48 study participants before (blue), immediately after (turquoise) and six months after travel (green). Boxes show medians and upper and lower quartiles, dots represent individual measurements.

5.5.3.2 Differences in pre- and post-travel resistomes

CCA was performed to assess whether travel causes a shift in the overall ARD pool sufficient to distinguish a pre- from a post-travel resistome. To investigate the impact of travel alone, regardless of ill health and medication, TD as well as antimicrobial and antimalarial use were added as covariates. The latter were included since, although not targeted at them, some antimalarial agents are thought to have an antibacterial effect on members of the gut microbiota community[353]. Pre- and post-travel resistomes showed significant separation starting at the highest level of the

AmrPlusPlus output, antimicrobial class ($p=1x10^{-3}$), and down to the resistance mechanism ($p=1x10^{-3}$) and ARD group levels ($p=1x10^{-3}$) (**Figure 5.3**). The differences in resistomes remained significant six months after travel at all three levels ($p=6x10^{-3}$; $p=2x10^{-3}$; $p=5x10^{-3}$, respectively).

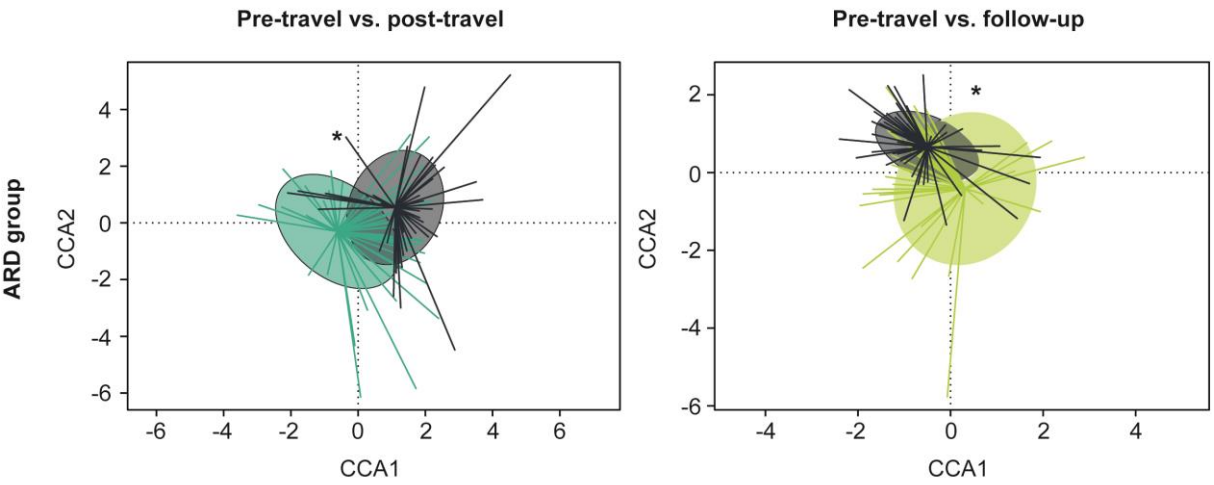

Figure 5.3: Changes in gut resistome composition after travel. Constrained correspondence analysis of resistome composition in 48 study participants before and immediately after travel (left) and before and six months after travel (right). The plot shows results for the antimicrobial resistance determinant (ARD) group level. Traveller's diarrhoea (TD) and the use of antibiotics and antimalarials were added as covariates in comparisons of pre- and post-travel samples. For comparing pre-travel and follow-up samples, antibiotic use after initial travel, repeat travel and TD during repeat travel were added as further covariates. Asterisks indicate significant separation of resistomes between timepoints as determined by ANOVA ($p \leq 0.05$).

5.5.3.3 Differences in specific antimicrobial resistance determinants after travel

To find out which ARDs drive the observed separation of post-travel and follow-up from pre-travel resistomes at the different levels of the AmrPlusPlus output, carriage rates of ARDs were compared for the three different timepoints using McNemar's test. Additionally, acquisition rates during travel were investigated. To gain quantitative information on top of mere presence/absence testing of ARDs, differences in their abundance, i.e. the number of reads assigned to an ARD at each timepoint, were assessed using zero-inflated Gaussian mixture models.

Before travel, all participants already carried ARDs conferring resistance to tetracyclines and macrolides, lincosamides and streptogramin (MLS) (**Figure 5.4A**). Trimethoprim ARDs were detected in 46 participants (95.83%) before travel, with a

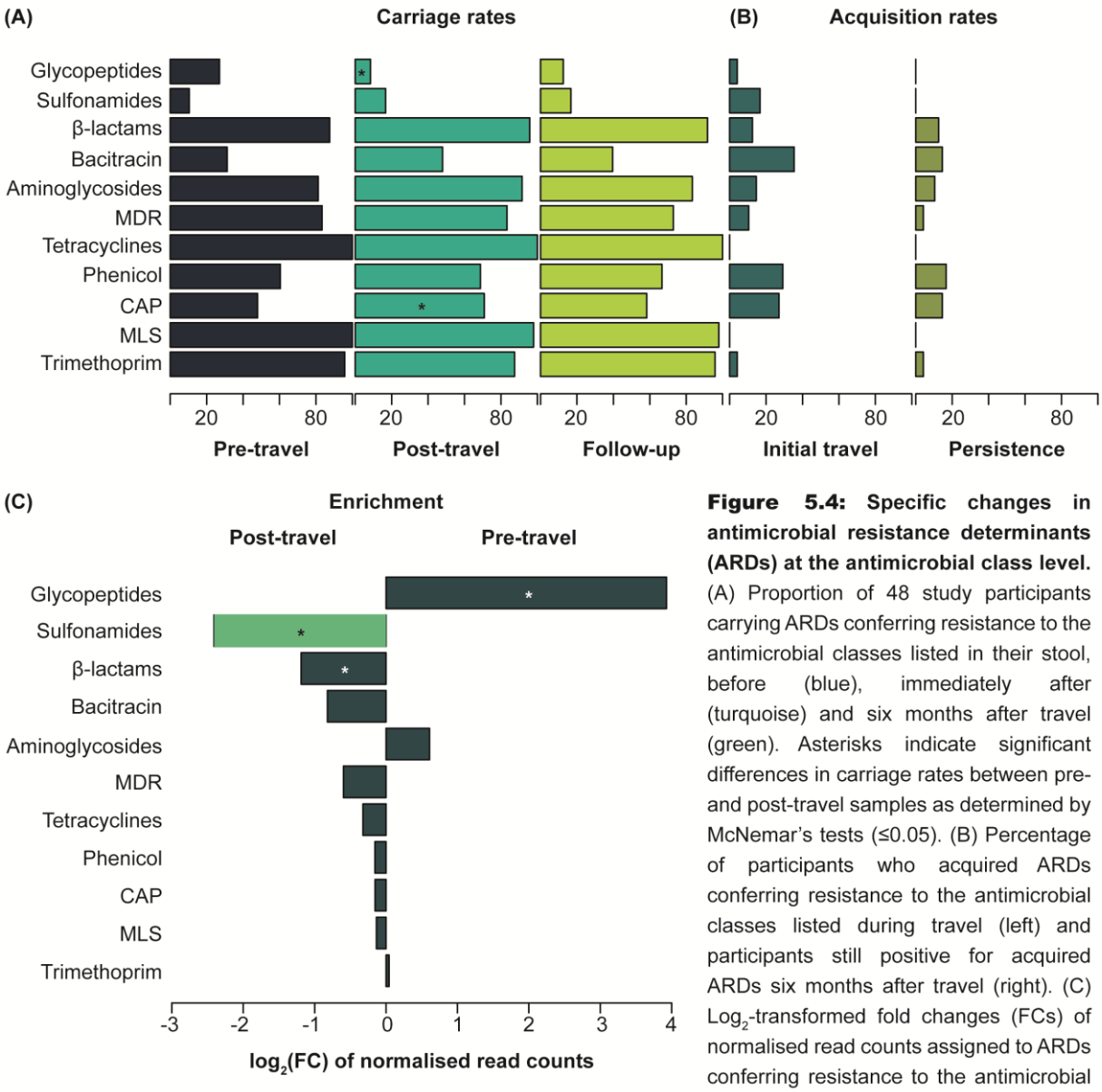

Figure 5.4: Specific changes in antimicrobial resistance determinants (ARDs) at the antimicrobial class level. (A) Proportion of 48 study participants carrying ARDs conferring resistance to the antimicrobial classes listed in their stool, before (blue), immediately after (turquoise) and six months after travel (green). Asterisks indicate significant differences in carriage rates between pre- and post-travel samples as determined by McNemar's tests (≤0.05). (B) Percentage of participants who acquired ARDs conferring resistance to the antimicrobial classes listed during travel (left) and participants still positive for acquired ARDs six months after travel (right). (C) Log_2-transformed fold changes (FCs) of normalised read counts assigned to ARDs conferring resistance to the antimicrobial classes listed, between pre- and post-travel samples. Positive FCs represent enrichment before, negative FCs enrichment after travel. Asterisks denote significant differences in ARD abundance between timepoints as determined by zero-inflated Gaussian mixture models (p≤0.05). Traveller's diarrhoea (TD), antibiotic and antimalarial use during travel were added as covariates. The lighter bar indicates a change which remains significant six months after travel. Antibiotic use before provision of the follow-up sample, re-travel and TD during re-travel were added as further covariates for this analysis. MDR, multidrug resistance; CAP, cationic antimicrobial peptides; MLS, macrolide-lincosamide-streptogramin.

reduction in carriage to 87.50% after travel. Significant differences in carriage rates were seen for glycopeptide and cationic antimicrobial peptide (CAP) ARDs, changing from 27.08% to 8.33% (p=0.03) and 47.92% to 70.83% (p=0.01), respectively. The Bonferroni-corrected p-value for these comparisons was $5x10^{-3}$. Non-significant

increases in carriage were observed for aminoglycoside (81.25% to 91.67%), bacitracin (31.25% to 47.92%), β-lactam (87.50% to 95.83%), phenicol (60.42% to 68.75%) and sulphonamide ARDs (10.42% to 16.67%). For most antimicrobial classes, ARD carriage had not returned to pre-travel rates at follow-up. Since fluoroquinolone ARDs were only detected in nine samples in total, two pre-, five post-travel and two follow-up samples, they were excluded from statistical analyses. None of the participants carried fluoroquinolone ARDs at more than one timepoint.

Acquisition rates were highest for bacitracin ARDs: seventeen participants (35.42%), who tested negative for these ARDs before travel, tested positive afterwards (**Figure 5.4B**). This was followed by acquisition rates for phenicol (29.17%), CAP (27.08%) and sulphonamide ARDs (16.67%). The proportion of participants still carrying acquired ARDs six months after initial travel was highest for phenicol (16.67%). Acquired sulphonamide ARDs did not persist in any of the participants. Conversely, only six participants (12.5%) acquired β-lactam ARDs but all six still tested positive at follow-up.

The abundance of sulphonamide and β-lactam ARDs was significantly higher after travel (log_2(FC)=2.41 and log_2(FC)=1.19, respectively) while glycopeptide ARDs were less abundant (log_2(FC)=3.93) (**Figure 5.4C**). Only for sulphonamide ARDs did this change persist up to six months after travel (log_2(FC)=4.60).

At the mechanism level, the changes in glycopeptide and β-lactam ARD abundance were driven by a persistent decrease in vanD-type regulator (log_2(FC)=5.87) and a persistent enrichment of class A β-lactamase genes (log_2(FC)=2.61). Although there were no significant differences in the sum of aminoglycoside ARDs, a significant decrease in the abundance of aminoglycoside O-nucleotidyltransferase genes was observed after travel (log_2(FC)=2.07).

Investigations at the group level revealed that the increase in CAP ARD carriage was partly mediated by *pmrC* genes, involved in lipid A modification, with their carriage rates increasing from 27.08% before to 54.17% after travel (p=0.01) (**Figure 5.5A**). Further ARDs present in a significantly higher number of post-travel samples were *cpxA* genes, which encode regulatory machinery for multidrug efflux systems (8.33% pre-travel vs. 31.25% post-travel, p=0.01), chloramphenicol acetyltransferase genes (18.75% vs. 41.67%, p=0.04) and *aph(3")*-type aminoglycoside O-

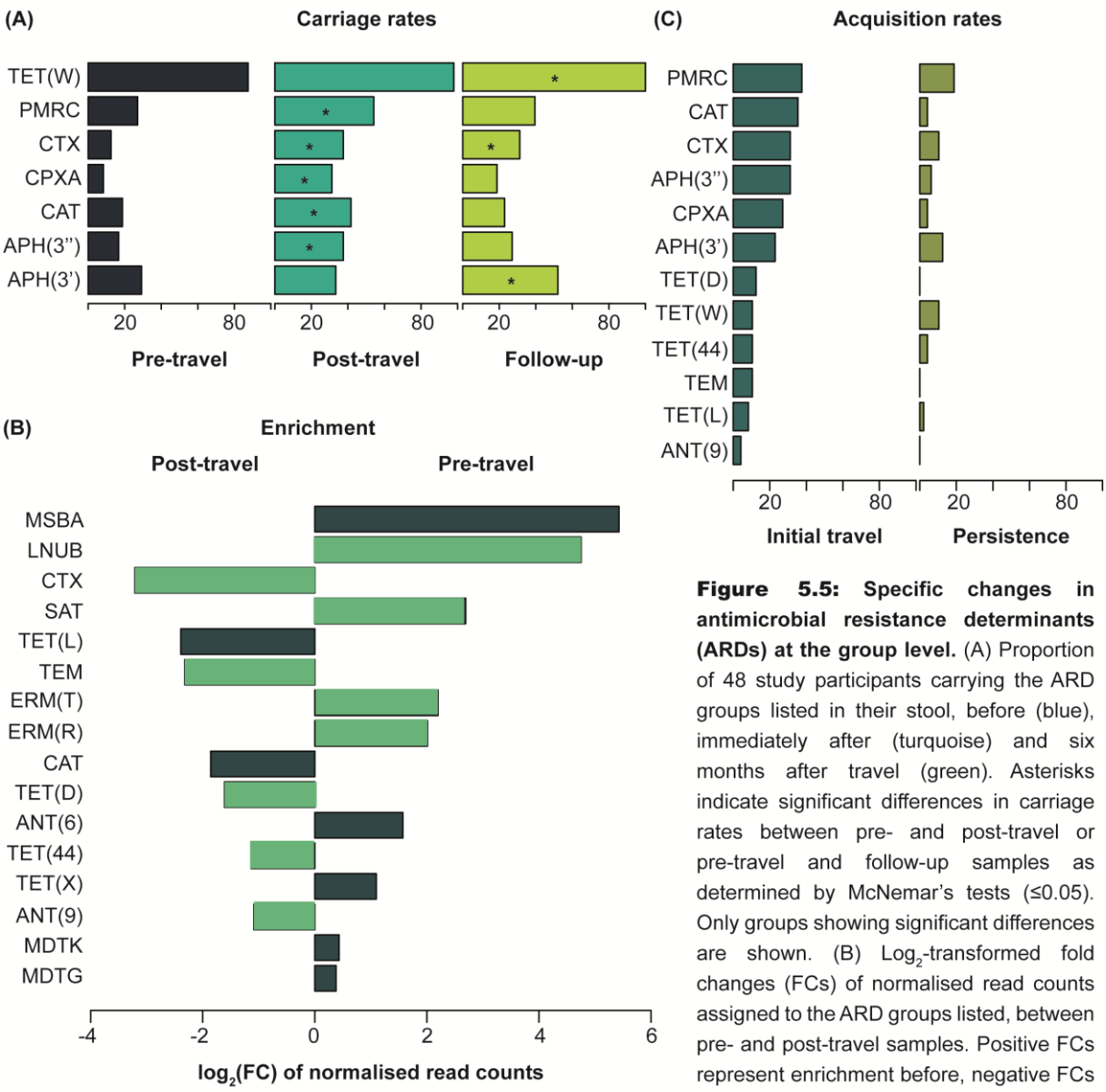

Figure 5.5: Specific changes in antimicrobial resistance determinants (ARDs) at the group level. (A) Proportion of 48 study participants carrying the ARD groups listed in their stool, before (blue), immediately after (turquoise) and six months after travel (green). Asterisks indicate significant differences in carriage rates between pre- and post-travel or pre-travel and follow-up samples as determined by McNemar's tests (≤0.05). Only groups showing significant differences are shown. (B) Log_2-transformed fold changes (FCs) of normalised read counts assigned to the ARD groups listed, between pre- and post-travel samples. Positive FCs represent enrichment before, negative FCs enrichment after travel. Only groups showing significant differences in abundance between timepoints, as determined by zero-inflated Gaussian mixture models, are shown (p≤0.05). Traveller's diarrhoea (TD), antibiotic and antimalarial use during travel were added as covariates. Lighter bars indicate changes which remain significant six months after travel. Antibiotic use before provision of the follow-up sample, re-travel and TD during re-travel were added as further covariates for this analysis. (C) Percentage of participants who acquired the ARD groups listed during travel (left) and participants still positive for acquired ARDs six months after travel (right). Only groups, for which significant differences in carriage rates or read counts were observed, are shown.

phosphotransferase genes (16.67% vs. 37.50%, p=0.04). Increased carriage did not persist until follow-up for any of these ARDs. However, carriage rates were significantly higher for another type of aminoglycoside O-phosphotransferase genes, namely those of the *aph(3')* type, six months after travel, increasing from 29.17% in

pre-travel to 52.08% in follow-up samples (p=0.04). Carriage rates of *tet*(W) tetracycline ribosomal protection protein genes increased non-significantly from 87.5% before to 97.92% after travel. In their follow-up samples, all participants carried *tet*(W) genes (p=0.04). The Bonferroni-corrected p-value for these comparisons was 5.26×10^{-4}.

Carriage of CTX group ARDs significantly increased from 12.5% before to 37.5% after travel (p=0.01) and remained elevated at follow-up (31.25%, p=0.04). Three different CTX group genes were detected in pre-travel samples: three participants carried $bla_{\text{CTX-M-54}}$ only, two $bla_{\text{CTX-M-159}}$ only and one participant carried both $bla_{\text{CTX-M-17}}$ and $bla_{\text{CTX-M-159}}$. Carriage of $bla_{\text{CTX-M-54}}$ persisted in post-travel samples of all three participants and was acquired by an additional nine during travel. Those initially testing positive for $bla_{\text{CTX-M-17}}$ and $bla_{\text{CTX-M-159}}$, on the other hand, tested negative for the genes immediately after travel. However, $bla_{\text{CTX-M-17}}$ was acquired by six and $bla_{\text{CTX-M-159}}$ by nine participants. Seven participants carried more than one of the previously mentioned genes after travel. While acquired $bla_{\text{CTX-M-159}}$ and $bla_{\text{CTX-M-54}}$ genes persisted up to six months after travel in one and four participants, respectively, $bla_{\text{CTX-M-17}}$ carriage did not persist. No participant persistently carried the same gene at all three timepoints. Another 54 CTX group genes were identified at follow-up. However, all of these were detected in a single sample from a 79-year old male who had travelled to Brazil and had not experienced TD, used antibiotics or re-travelled.

For some of the ARDs mentioned above, increased post-travel carriage rates were also reflected by an increase in abundance (**Figure 5.5B**): There was an enrichment of reads assigned to *cat* ($\log_2$(FC)=1.86) and bla_{CTX} genes ($\log_2$(FC)=3.21). The latter was still observed six months after travel ($\log_2$(FC)=2.86). Another group contributing to the enrichment in class A β-lactamase ARDs were bla_{TEM} genes. While *tet*(L) ($\log_2$(FC)=2.39) and *tet*(D) ($\log_2$(FC)=1.61) tetracycline efflux pump as well as *tet*(44) ($\log_2$(FC)=1.14) ribosomal protection protein genes were more abundant after travel, there was a decrease in abundance of *tet*(X) tetracycline inactivation enzyme genes ($\log_2$(FC)=1.10). Similarly, there was a persistent increase in *ant(9)* aminoglycoside O-nucleotidyltransferase gene abundance ($\log_2$(FC)=1.09) but a lower abundance of another O-nucleotidyltransferase gene group, *ant(6)* ($\log_2$(FC)=1.57). Additional

groups less abundant after travel were *msbA* (log_2(FC)=5.43), *mdtK* (log_2(FC)=0.43) and *mdtG* (log_2(FC)=0.38) multidrug efflux pump genes. A persistent decrease was detected for *sat* aminoglycoside N-acetyltransferase genes (log_2(FC)=2.69) as well as *lnuB* lincosamide nucleotidyltransferase (log_2(FC)=4.75) and *erm*(T) (log_2(FC)=2.20) and *erm*(R) (log_2(FC)=2.01) rRNA methyltransferase genes involved in MLS resistance.

Among the ARD groups differing significantly either in carriage rate or abundance, acquisition rates were highest for *pmrC* genes (37.5%), followed by *cat* (35.42%), bla_{CTX} (31.25%), *aph(3'')* (31.25%), *cpxA* (27.08%) and *aph(3')* (22.92%) genes (**Figure 5.5C**). The number of participants, who still tested positive for acquired ARDs six months after travel, was also highest for *pmrC* genes (18.75%), followed by *aph(3')* (12.5%), *tet*(W) and bla_{CTX} genes (both 10.42%).

5.5.3.4 Effect of traveller's diarrhoea and antimicrobial use on the post-travel resistome

Given that TD and consumption of antimicrobials have been identified as risk factors for acquisition of ARDs in previous studies, their effect on diversity and composition of the study participants' resistomes after travel was evaluated. Participants will be referred to as TD-positive or –negative and antimicrobial-positive or –negative in the following paragraphs, depending on whether or not they experienced TD and used antimicrobials.

Immediately after travel, the 22 TD-positive participants carried a mean of 62.95±32.65 and the 26 TD-negative ones 70.46±43.21 individual ARDs. The difference between the two groups was not significant. Similarly, there was no separation of TD-positive and TD-negative post-travel resistomes at either the class, mechanism or group level of the AmrPlusPlus output after CCA.

Chi-square tests were carried out to investigate whether post-travel carriage or acquisition rates of specific ARDs were associated with TD. There was an association between TD and carriage of polymyxin B resistance regulator genes, which were identified in 68.18% of TD-positive and 30.77% of TD-negative participants after travel (p=0.02). Furthermore, acquisition rates for aminoglycoside efflux regulator genes were higher after TD. The genes were acquired by 54.55% of

TD-positive and only 19.23% of TD-negative participants (p=0.02). The Bonferroni-corrected p-value for these tests was $2x10^{-3}$.

In line with carriage rates, polymyxin B resistance regulator genes were more abundant in TD-positive participants after travel (log_2(FC)=2.84). This difference was still observed at follow-up (log_2(FC)=2.27). Another mechanism persistently enriched after TD were streptogramin resistance ATP-binding cassette ABC efflux pumps (log_2(FC)=3.04). At the group level, the enrichment of these two mechanisms in TD-positive participants was partly due to a persistently higher abundance of *pmrA* (log_2(FC)=3.24) and *lsa* genes (log_2(FC)= 2.54) (**Figure 5.6**). Additional groups found to be non-persistently enriched after TD included several involved in MDR, such as *cpxA* (log_2(FC)=2.74), *baeR* (log_2(FC)=2.65), *emrE* (log_2(FC)=2.64) and *mdfA* genes (log_2(FC)=2.11), *cat* genes (log_2(FC)=3.40), penicillin-binding protein genes of the *pbp4b* (log_2(FC)=2.15) and *ampH* type (log_2(FC)=1.57) as well as CFX group class A β-lactamase genes (log_2(FC)=1.18). ARD groups more abundant in TD-negative participants included several involved in resistance to MLS, namely *erm*(G) (log_2(FC)=7.19), *erm*(T) (log_2(FC)=1.79), *erm*(F) (log_2(FC)=3.74), *lnuB* (log_2(FC)=5.52) and *mphA* genes (log_2(FC)=0.79), in resistance to aminoglycosides, specifically *acrS* (log_2(FC)=1.38), *ant(9)* (log_2(FC)=4.94) and *sat* genes (log_2(FC)=3.49), and in resistance to tetracyclines, namely *tet*(X) (log_2(FC)=7.08) and *tet*(D) genes (log_2(FC)=2.94).

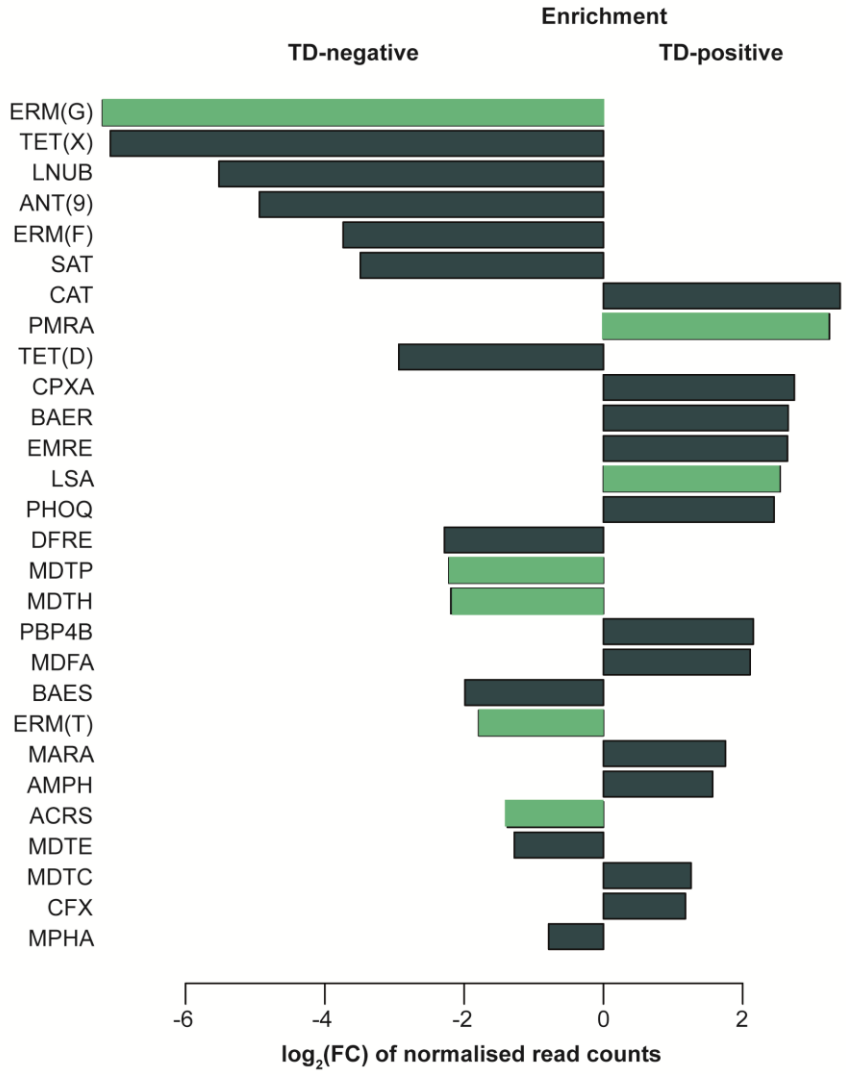

Figure 5.6: Post-travel changes in abundance of specific antimicrobial resistance determinant (ARD) groups after traveller's diarrhoea (TD). Log_2-transformed fold changes (FCs) of normalised read counts assigned to the ARD groups listed, between 22 participants who did and 26 who did not experience TD. Positive FCs represent enrichment in TD-positive, negative FCs enrichment in TD-negative travellers. Only groups showing significant differences in abundance, as determined by zero-inflated Gaussian mixture models, are shown ($p \leq 0.05$). Antibiotic and antimalarial use during travel were added as covariates. Lighter bars indicate changes which remain significant six months after travel. Antibiotic use before provision of the follow-up sample, re-travel and TD during re-travel were added as further covariates for this analysis.

Like TD, antimicrobial use did not have a significant effect on resistome diversity, with the 11 antibiotic-positive participants carrying a mean of 71.91±35.63 and the 37 antibiotic-negative ones 65.57±39.69 individual ARDs immediately after travel. However, a significant separation of antimicrobial-positive from antimicrobial-negative post-travel resistomes was observed at the class ($p=0.01$), mechanism ($p=7x10^{-3}$)

and group level ($p=5x10^{-3}$) of the AmrPlusPlus output after CCA (**Figure 5.7A**). This separation was not apparent anymore at follow-up.

Comparison of post-travel carriage rates between the two groups using chi-squared tests provided little insight into which ARDs were responsible for the

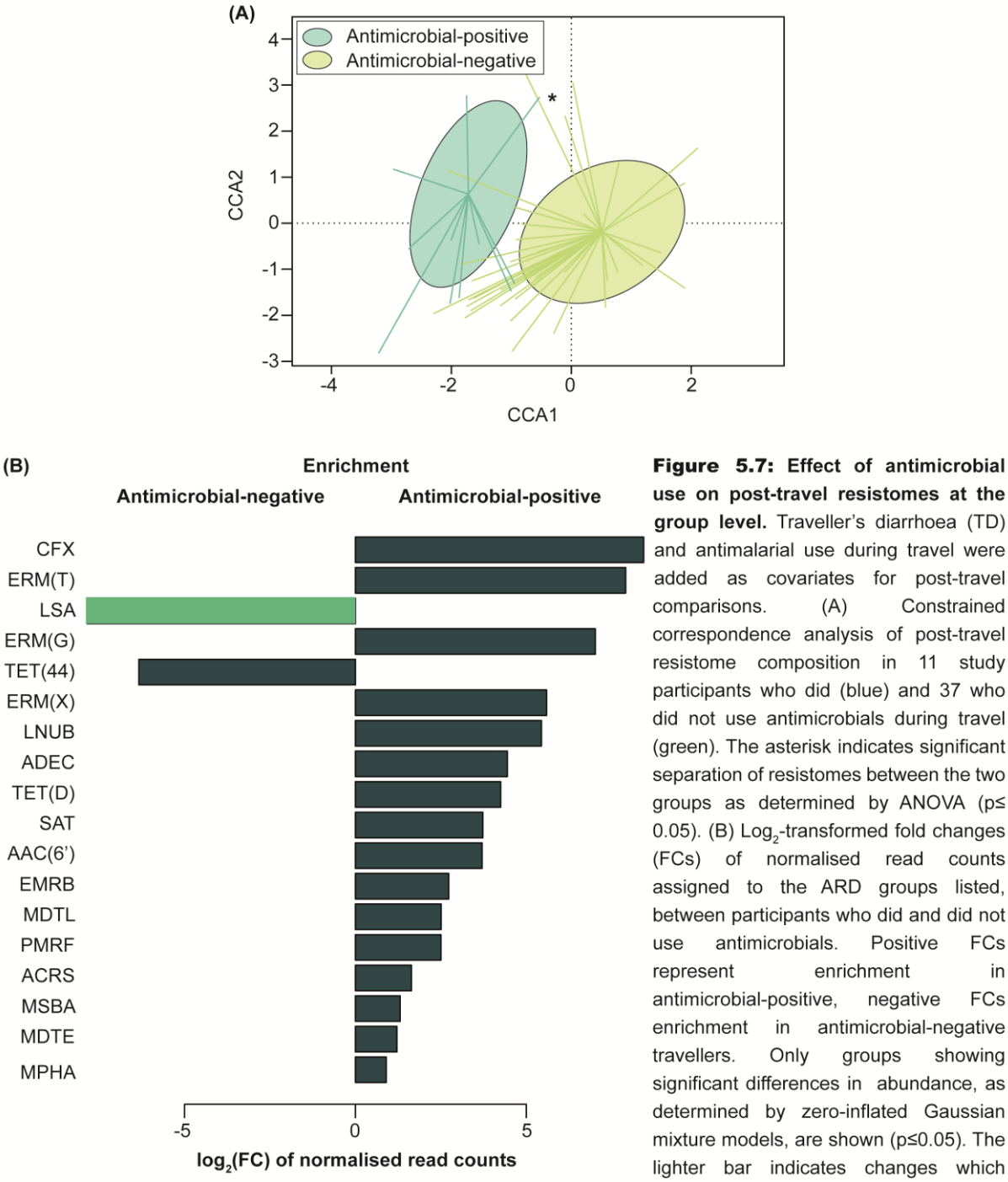

Figure 5.7: Effect of antimicrobial use on post-travel resistomes at the group level. Traveller's diarrhoea (TD) and antimalarial use during travel were added as covariates for post-travel comparisons. (A) Constrained correspondence analysis of post-travel resistome composition in 11 study participants who did (blue) and 37 who did not use antimicrobials during travel (green). The asterisk indicates significant separation of resistomes between the two groups as determined by ANOVA ($p \leq 0.05$). (B) Log_2-transformed fold changes (FCs) of normalised read counts assigned to the ARD groups listed, between participants who did and did not use antimicrobials. Positive FCs represent enrichment in antimicrobial-positive, negative FCs enrichment in antimicrobial-negative travellers. Only groups showing significant differences in abundance, as determined by zero-inflated Gaussian mixture models, are shown ($p \leq 0.05$). The lighter bar indicates changes which remain significant six months after travel. Antimicrobial use before provision of the follow-up sample, re-travel and TD during re-travel were added as further covariates for this analysis.

separation. Only *lnuC* genes were found to be more prevalent after antimicrobial use, detected in 81.82% of antimicrobial-positive and 40.54% of antimicrobial-negative participants (p=0.04). Five antimicrobial users (45.45%) acquired *mecB* genes during travel, compared to eight antimicrobial-negative participants (21.62%). While the acquired genes were still detected in only three of the antimicrobial-negative travellers at follow-up, they persisted in all antimicrobial users. The association between antimicrobial use and persistent *mecB* gene acquisition was significant (p=0.01). The Bonferroni-corrected p-value for these tests was 5.26×10^{-4}.

Examining ARD abundance, glycopeptide ARDs were persistently enriched after antimicrobial use ($\log_2$(FC)=5.26). At the mechanism level, this change was driven by vanD-type regulator genes ($\log_2$(FC)=4.32). An enrichment in reads assigned to *cfx* class A beta-lactamase ($\log_2$(FC)=8.43) and *tet*(D) tetracycline efflux pump genes ($\log_2$(FC)=4.25), genes involved in resistance to MLS, such as *erm*(T) ($\log_2$(FC)=7.91), *erm*(G) ($\log_2$(FC)=7.02), *erm*(X) ($\log_2$(FC)=5.59) and *lnuB* ($\log_2$(FC)=5.44), resistance to aminoglycosides, specifically *sat* ($\log_2$(FC)=3.73), *aac(6')* ($\log_2$(FC)=3.70) and *acrS* ($\log_2$(FC)=1.64), in MDR, including *adeC* ($\log_2$(FC)=4.45), *emrB* ($\log_2$(FC)=2.73) and *mdtL* ($\log_2$(FC)=2.51), and in *pmrF* genes ($\log_2$(FC)=2.50), involved in lipid A modification, was observed in antimicrobial users at the group level (**Figure 5.7B**). Streptogramin resistance ATP-binding cassette ABC efflux pump genes of the *lsa* type ($\log_2$(FC)=7.85) and *tet*(44) tetracycline resistance ribosomal protection protein genes ($\log_2$(FC)=6.33) were more abundant in antimicrobial-negative participants. Only the changes in *lsa* genes persisted at follow-up ($\log_2$(FC)=7.36).

5.5.4 Prevalence and persistence of changes in taxonomic composition after travel

5.5.4.1 Taxonomic diversity of the microbiota

The number of individual OTUs detected in the samples was used as an estimate of taxonomic diversity of the microbiota. A total of 2,685 different OTUs were identified by the Kraken taxonomic classifier, including bacteria, archaea and viruses/phages. Twenty-six participants (54.17%) carried a higher number of OTUs immediately after

travel than before. For 18 out of these 26 (69.23%), the number remained increased in the follow-up samples. The mean number of individual OTUs per sample increased slightly from a mean of 1,643.25±96.81 (range: 1,242-1,775) before travel to 1,656.40±80.53 (range: 1,458-1,823) immediately after travel and reached 1,676.21±66.10 (range: 1,541-1,854) at follow-up (**Figure 5.8**). The difference between the mean number of OTUs before and six months after travel was statistically significant as determined by paired t-tests (p=0.02). Again, repeat travel might play a part in this observation.

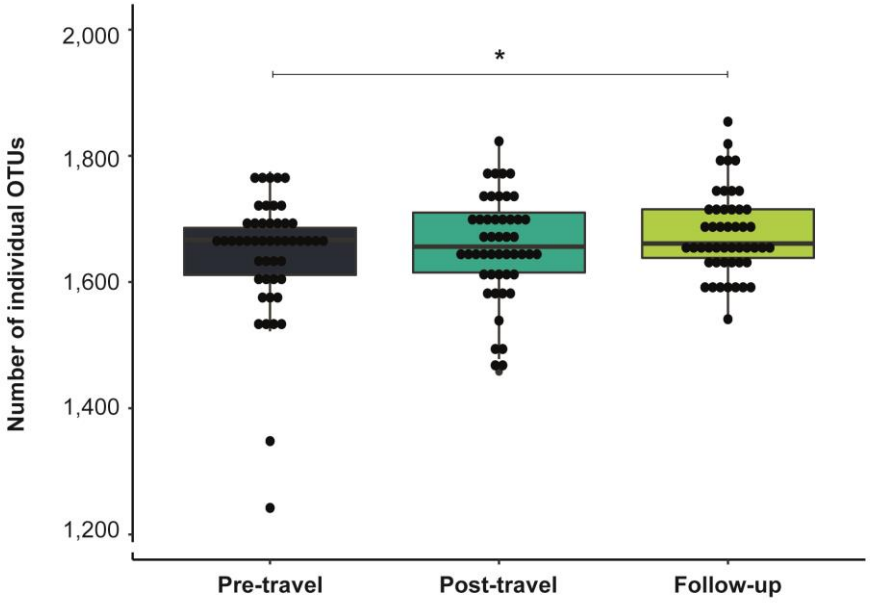

Figure 5.8: Changes in diversity of gut microbiota taxonomic composition after travel. Number of individual operational taxonomic units (OTUs) detected in the stools of 48 study participants before (blue), immediately after (turquoise) and six months after travel (green). Boxes show medians and upper and lower quartiles, dots represent individual measurements. The asterisk denotes a significant difference in OTU numbers between timepoints, as determined by paired t-tests (p≤0.05).

The majority of the 2,685 total OTUs were only detected in a few samples and at a very low abundance. To focus on the most abundant OTUs, those that were not present at a relative abundance of ≥0.01% in at least ten samples were removed before further analysis. The resulting number of remaining individual OTUs was 457.

5.5.4.2 Differences in pre- and post-travel taxonomic compositions

CCA of the 457 most abundant OTUs was carried out, with OTUs aggregated at either the phylum, family, genus or species level. Travel impacted on overall taxonomic composition at the species level only (p=0.03) (**Figure 5.9**). This impact was still apparent at follow-up ($p=1x10^{-3}$).

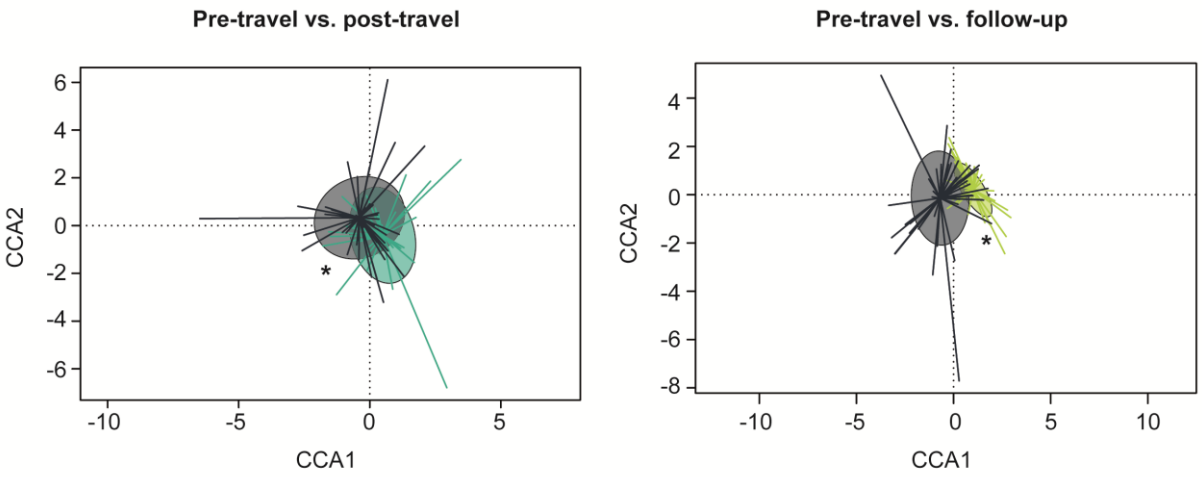

Figure 5.9: Changes in gut microbiota taxonomic composition after travel. Constrained correspondence analysis of taxonomic composition in 48 study participants before and immediately after travel (left) and before and six months after travel (right). The plot shows results at the species level. Only OTUs present at a relative abundance ≥0.01% in at least ten samples were included in the analysis. Traveller's diarrhoea (TD) and the use of antibiotics and antimalarials were added as covariates in comparisons of pre- and post-travel samples. For comparing pre-travel and follow-up samples, antibiotic use after initial travel, repeat travel and TD during repeat travel were added as further covariates. Asterisks indicate significant separation of microbiotas between timepoints as determined by ANOVA (p≤0.05).

5.5.4.3 Differences in specific gut microbiota taxa after travel

In agreement with CCA results, no significant differences in the abundance of specific taxa were observed at the phylum level after travel. At the family level, the largest post-travel enrichment was observed for Enterobacteriaceae ($log_2(FC)=1.93$) but again this change was non-significant.

Carriage rates for potentially pathogenic genera of the Enterobacteriaceae family, with a relative abundance >0.01% being classed as positive carriage, were significantly higher after travel, increasing from 75% to 95.83% for *Escherichia* ($p=9.38x10^{-3}$), 27.08% to 68.75% for *Klebsiella* ($p=5.23x10^{-4}$), 39.58% to 72.92% for *Salmonella* ($p=2.2x10^{-3}$) and 31.25% to 72.92% for *Shigella* ($p=1.05x10^{-4}$) (**Figure**

5.10A). *Campylobacter* carriage rates also increased, albeit non-significantly, from 89.58% to 97.92%. Carriage remained increased for these genera six months after travel at 95.83% (p=0.02), 66.67% ($p=5.32x10^{-4}$), 77.08% ($p=8.56x10^{-4}$), 52.08% (p=0.03) and 100%, respectively. The Bonferroni-corrected p-value for these tests was 0.01.

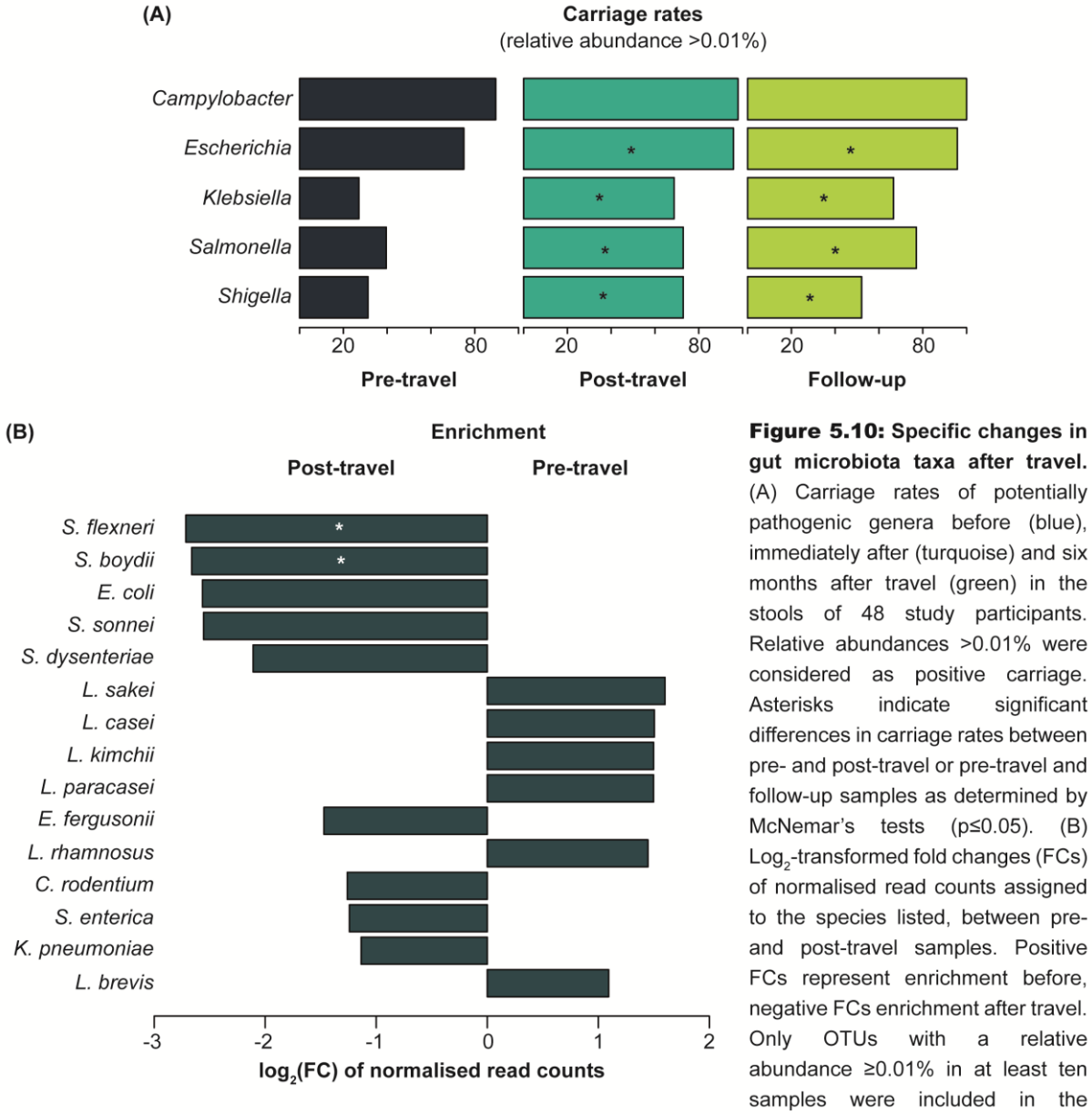

Figure 5.10: Specific changes in gut microbiota taxa after travel. (A) Carriage rates of potentially pathogenic genera before (blue), immediately after (turquoise) and six months after travel (green) in the stools of 48 study participants. Relative abundances >0.01% were considered as positive carriage. Asterisks indicate significant differences in carriage rates between pre- and post-travel or pre-travel and follow-up samples as determined by McNemar's tests (p≤0.05). (B) Log_2-transformed fold changes (FCs) of normalised read counts assigned to the species listed, between pre- and post-travel samples. Positive FCs represent enrichment before, negative FCs enrichment after travel. Only OTUs with a relative abundance ≥0.01% in at least ten samples were included in the analysis and only the species with the 15 highest FCs are shown. Asterisks denote significant differences in abundance between timepoints as determined by zero-inflated Gaussian mixture models (p≤0.05). Traveller's diarrhoea (TD), antibiotic and antimalarial use during travel were added as covariates.

Several species from the above-mentioned genera were found to display the largest changes in abundance after travel, including *S. flexneri* (log_2(FC)=2.71, p=0.02), *S. boydii* (log_2(FC)=2.66, p=0.02), *E. coli* (log_2(FC)=2.57), *S. enterica* (log_2(FC)=1.24) and *K. pneumoniae* (log_2(FC)=1.14) (**Figure 5.10B**). Significant enrichment of *S. flexneri* and *S. boydii* did not persist up to the follow-up period. Species less abundant in post- compared to pre-travel samples included members of the Lactobacillales order such as *L. sakei* (log_2(FC)=1.60), *L. casei* (log_2(FC)=1.50) and *L. kimchii* (log_2(FC)=1.50). There were no significant differences in pre- or post-travel taxonomic compositions between the 15 participants who acquired CTX group genes during travel and those who did not.

5.5.4.4 Effect of traveller's diarrhoea and antimicrobial use on taxonomic composition after travel

Neither TD nor antimicrobial use affected post-travel microbiota diversity with TD-positive participants carrying a mean of 1,642.14±79.77 individual OTUs compared to 1,668.46±80.72 in TD-negative ones and antimicrobial-positive participants carrying a mean of 1,626.82±69.55 individual OTUs compared to 1,665.19±82.31 in antimicrobial-negative participants.

An impact on overall post-travel taxonomic composition at the family ($p=4x10^{-3}$), genus ($p=3x10^{-3}$) and species level (p=0.03) was observed after TD only (**Figure 5.11**). Separation did not persist until follow-up. Significant changes in abundances of specific taxa were not detected, nor were differences in pre-travel microbiota composition between TD-positive and –negative participants. Furthermore, chi-square tests did not reveal any differences in carriage rates of potentially pathogenic genera after either TD or antibiotic use.

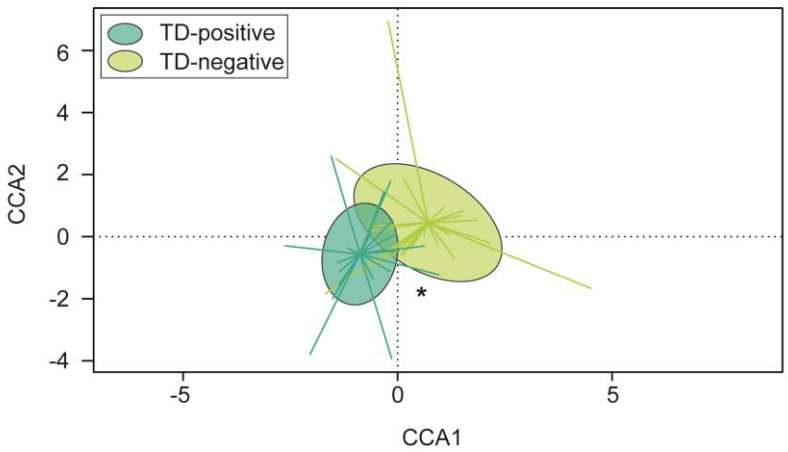

Figure 5.11: Effect of traveller's diarrhoea (TD) on post-travel gut microbiota taxonomic composition at the species level. Constrained correspondence analysis of post-travel resistome composition in 22 study participants who did (blue) and 26 who did not use antibiotics during travel (green). The asterisk indicates significant separation of microbiotas between the two groups as determined by ANOVA ($p \leq 0.05$). Antibiotic and antimalarial use during travel were added as covariates for this analysis.

5.6 Discussion

5.6.1 Effect of travel on gut microbiota resistome and taxonomic composition

In line with previous findings[374,378], travel induced changes in the gut microbiota ARD pool in the present study cohort, including an overall enrichment in the number of ARDs detected. This enrichment could be due to ingestion of bacteria carrying the genes though contaminated food[123] or water[124] in regions where resistance rates are high. Exposure to a foreign environment and its microorganisms might also be expected to increase bacterial diversity[383] but the number of OTUs was found to be relatively stable over time. It should be noted, however, that this metric is only a rough estimate of true diversity and measures that take into account abundance data and phylogenetic relationships between organisms are generally favourable. The range of 7-127 individual ARDs identified in the pre-travel samples indicates that some ARDs are already widespread in the UK community. Almost all participants carried ARDs conferring resistance to tetracyclines, MLS and trimethoprim before travel. Tetracyclines and MLS are amongst the antibiotics with the highest consumption rates in the UK[384]. Although all participants were UK residents at the time of recruitment only 54% are UK natives. More lenient prescription practices in other countries might mean that some participants are more likely to have been exposed to antimicrobials at a higher rate at some point in their lives, which could contribute to the varying number of ARDs observed before travel[385]. Nonetheless, the clear, and persistent, separation observed between pre- and post-travel resistomes at all levels indicates a large-scale shift in resistance potential after exposure to a foreign environment. Taxonomic changes, on the other hand, only became apparent at species level resolution. This suggests that while travel might open up a niche for specific organisms and lead to displacement of others, it does not cause large-scale imbalances in the gut microbiota.

Enterobacteriaceae, the focus of many previous studies[368,369,372], were the taxon displaying the largest differences in this cohort. Carriage rates and abundances of Escherichia, Klebsiella, Salmonella and Shigella, all members of the family capable

of causing disease, were significantly higher immediately after travel. Along with changes in Enterobacteriaceae, carriage and abundance of CTX group ESBL genes also increased. Genes were detected in 12.5% of participants before travel and acquired by 31.25% of individuals whilst abroad. The rates are comparable to those observed by von Wintersdorff et al.[374] in their study of 122 Dutch travellers. Acquired blaCTX genes persisted in 10.42% of participants, thus creating a pool for onward transmission in the community up to six months after return. Isolation of strains to prove ESBL production was not carried out, thus the genes detected might not necessarily be expressed. A previous study by Bengtsson-Palme et al.[378] combining both metagenomics and culture found that in some cases ESBL-E could be isolated even when blaCTX-M genes were not detected in the metagenome, possibly because they were present below the detection limit. However, the same study also reported cases in which ESBL genes were detected without successful isolation of ESBL-E indicating that the genes might be carried by other organisms.

This highlights the importance of recognising the entire gut microbiota community as an ARD reservoir. Use of an unbiased metagenomics approach allowed detection of an enrichment in sulphonamide and additional β-lactam ARDs after travel as well as increased carriage and abundance and high acquisition rates for cat genes involved in chloramphenicol resistance. Carriage rates of ARDs conferring resistance to CAPs were also elevated after travel. Antimicrobial peptides are part of the innate immune defence in many organisms and some are being investigated as promising alternatives to classical antibiotics for a more targeted treatment of infections in the era of AMR[62]. Mechanisms of CAP resistance, such as lipid A modification, have been described for many of the Enterobacteriaceae enriched after travel[386]. While potential pathogens were more abundant after travel, the abundance of several Lactobacillus and Leuconostoc species, generally thought to be beneficial to the host, was decreased. Many lactic acid bacteria are intrinsically resistant to the glycopeptide antibiotic vancomycin[387,388], which likely contributed to the observed reduction in carriage and abundance of glycopeptide ARDs.

5.6.1.1 Risk factors for resistome and taxonomic changes

TD and antimicrobial use, the most consistently identified risk factors for acquisition of resistant organisms, only had limited effects on the gut microbiota of participants in this study. Just under half of the participants experienced TD, which is in agreement with the proportion of 20-50% expected for travellers to developing regions[383]. TD and antimicrobial use have both been shown to decrease microbiota diversity[383] but this could not be reproduced in the present cohort. Again, this might be due to the limitations of the diversity metric used. The effect of TD on overall taxonomic composition was more pronounced than that of antimicrobial use while the opposite was the case for resistome composition. This makes sense in the light of findings that antimicrobial exposure selects for resistant organisms[125,126], especially those not susceptible to the particular drug used[372]. Only two of the eight participants who used ciprofloxacin acquired fluoroquinolone ARDs. MLS ARDs were already highly prevalent before travel so that an effect of azithromycin use in three participants could not be detected. The association between persistent acquisition of *mecB* genes and antimicrobial consumption is worrying as *mecB* is a homologue of *mecA*[389], which is part of the resistance machinery of methicillin-resistant *Staphylococcus aureus* (MRSA), and has recently been described in a clinical MRSA isolate[390]. Duration of TD and antibiotic use was short in most participants and might have occurred early on in their travels so that microbiota balance had already been restored by the time they provided their post-travel sample. The majority of changes observed did not persist until follow-up, which is in line with findings that single, short insults to the system are not sufficient to cause long-term dysbiosis[391].

Apart from TD and antibiotic use, the exact travel destination, has been identified as a risk factor for AMR acquisition during travel, with the highest rates reported for South Asia, in particular India[369,380]. Differences between travel destinations were not evaluated in this study as stratification of results by geographic region was not feasible due to the low sample number. Inconsistently identified risk factors include old age[375,377,378], chronic disease[392] and consumption of certain foods such as ice cream, streetfood, raw vegetables and fruits[368,393]. While these aspects were captured by the study questionnaires, statistical analysis of their influence was again hindered by the low sample number.

5.6.2 Clinical relevance of study results

A survey of travellers recently found that, while the majority were aware of the AMR problem, only few realised that travel was associated with a high risk of acquiring resistant organisms and the potential of transferring them to others[394]. The growing body of evidence this study feeds into provides a useful resource for healthcare practitioners providing pre-travel consultations. They should promote measures proven to be effective in preventing TD, such as regular hand washing and avoidance of specific drink and food items. Depending on the extent of distress the illness causes and the degree to which it interferes with planned activities, TD is categorised as either mild, moderate or severe[354]. Bismuth subsalicylate and loperamide should be advertised as potent alternatives to antimicrobials for treatment of mild and moderate cases of TD[395]. Antibiotic prophylaxis, although highly efficient at preventing TD[396], is not recommended anymore for healthy travellers. Even provision of stand-by antibiotics might be an issue: their availability did not prevent travellers from seeking healthcare abroad and indeed increased antibiotic use[397]. Interestingly, most participants in this study who consumed antibiotics used ciprofloxacin despite reported fluoroquinolone resistance rates of 93.1% in China and 70-80% in Nepal and Thailand for *Campylobacter*[398] and of 65% for ETEC and EAEC isolated in India[399]. Patients presenting with travel-related diarrhoea were found to be less likely to respond to fluoroquinolone treatment than those who acquired the infection at home[400]. Furthermore, there is evidence that ciprofloxacin use favours acquisition of fluoroquinolone-resistant strains[372], which further exacerbates the problem.

The results also highlight the importance of gathering information on travel history, ideally along with other travel-related factors, when a patient presents with gastrointestinal disease that requires antimicrobial treatment. Depending on this travel history, clinicians can then assess which antimicrobials are more likely to be ineffective due to acquired resistance. The present study suggests that patients might not respond to sulphonamides and β-lactams after recent travel.

5.7 Conclusions

Although limited by its sample size, this proof-of-concept study highlights the usefulness of a metagenomics approach to assess the resistance potential in travellers outside a healthcare setting, taking into account not just pathogenic species but the entire gut microbiota community as a reservoir for AMR. A surveillance system based on healthy travellers could contribute to advancing our understanding of AMR dissemination in the community in countries like the UK where resistance is on the rise. Information gathered from this surveillance could then lead to updated policies for medical professionals providing pre-travel consultations.

CHAPTER SIX

6. General discussion

6.1 Summary of findings

6.1.1 Diagnosis of gastrointestinal infections via '–omics' approaches

The use of culture-free metataxonomic, metabolomic and metagenomic approaches for identification of aetiological agents of gastroenteritis was explored in chapters 2 and 3. Direct detection of bacterial enteropathogens by 16S rRNA gene sequencing seemed to depend heavily on sample quality, pathogen load and the type of pathogen present. Only a subset of samples testing positive for *Campylobacter* during conventional diagnostics contained the pathogen at a relative abundance detectable by the taxonomic classifier used and *Salmonella* could not be identified at genus level in any of the *Salmonella enterica*-positive samples. Similarly, *S. enterica* was not identified in all samples from patients diagnosed with salmonellosis using a shotgun sequencing approach and relative abundances of the pathogen in positive samples varied widely. Some of these results could be due to the sequence clustering algorithms and the databases that the taxonomic classifiers, which were chosen for analysis, rely on. While QIIME 2[173], Kraken[260] and MetaPhlAn2[258] are all widely used in the field, none of them will achieve 100% classification accuracy. Almeida et al. (2018)[401] evaluated the performance of four of the most commonly used tools for 16S rRNA gene sequencing data analysis with the help of mock communities of known composition. While the use of QIIME 2 resulted in the lowest number of unclassified sequences and the most accurate representation of taxon abundances, it also yielded a higher proportion of misclassified reads. It is therefore possible that additional *Campylobacter*- or *S. enterica*-derived sequences were present in the study samples but were falsely assigned to other taxa. An equivalent study comparing tools for the analysis of metagenomic sequencing data found that the accuracy of taxonomic assignments by both Kraken and MetaPhlAn2 decreased beyond the family level[402].

Colonisation of the gastrointestinal (GI) tract with foreign enteropathogens is expected to cause a disturbance of the usual metabolic activity of microorganisms in the gut, which should be reflected in measurable differences in the concentrations of

gut microbiota-derived metabolites. When direct detection of the pathogen is not possible, metabolic biomarkers of infection, established using a metabolomics approach, could therefore support differential diagnoses. In this study, identification of individual pathogen-specific biomarkers was not entirely successful. The same applied for taxonomic biomarkers. This was due to the fact that, within sample groups, relative bacterial abundances and concentrations of potential metabolic and taxonomic biomarkers ranged widely, possibly due to pre-infection interindividual variations in microbiota composition. Different communities might vary in their response to the same insult caused by an invading pathogen. Zaneveld et al. (2017)[403] proposed that, rather than adhering to a predictable pattern, changes induced by gut microbiota perturbations are stochastic, following the so-called 'Anna Karenina principle'. Based on this suggestion, dysbiotic microbiotas will be more dissimilar to one another than healthy, balanced ones. The tight clustering of control samples compared to the dispersal of infection samples observed when analysing overall taxonomic and metabolic profiles supports this hypothesis. Interestingly, protozoal infections appeared to result in less severe disturbances than bacterial ones. Whether this is because they are eukaryotes from a different domain of life that only makes up a small proportion of the normal microbiota, and commensal activity is affected more strongly by competition from organisms that share a closer phylogenetic relationship, remains to be elucidated.

While direct detection of aetiological agents of gastroenteritis and pathogen-specific biomarkers was of limited success, the use of a supervised learning algorithm, which takes into account changes in a multitude of features from metataxonomic and metabolomics datasets, yielded more promising results. Prediction accuracy could probably be increased even further by analysing the same number of samples per group. It has to be kept in mind, however, that correct classification of samples into groups using a supervised learning approach strongly depends on the quality of the training data set. Samples were labelled according to the results of conventional culture- or PCR-based diagnostics, which in this case was deemed the gold standard. Appropriate training of the classifier can therefore only be assumed in the absence of misdiagnoses but, as discussed in chapter 2, current diagnostic methods for gastroenteritis are imperfect. An undetected co-infection with

a second enteropathogen could also distort the results. Whilst evaluating a multiplex PCR panel targeting thirteen GI bacteria, Martín et al. (2018)[144] detected co-infections in 15% of samples that were missed by culture. Similar studies reported co-infections in 14.1%-33.5% of samples[404,405], with *Campylobacter* and EPEC being the enteropathogens most frequently encountered together. Training datasets would therefore have to be curated carefully to exclude or account for such samples before implementing a Random Forest classifier as a diagnostic tool.

6.1.2 Use of '–omics' technologies for surveillance of gastrointestinal pathogens

The advent of high throughput sequencing technologies has revolutionised how infectious diseases are tracked throughout the population and vastly increased the resolution of genome-based surveillance. Since the first complete bacterial genome of a *Haemophilus influenza* strain was published in 1995 based on traditional Sanger sequencing[406], turnaround times and costs of pathogen genome sequencing have decreased steadily. Nowadays, a complete genome can be generated for approximately £50 in less than a day. Along with the sequencing technologies, computational tools needed to evolve to deal with the large amount of data produced by the new approaches. Public Health England (PHE) was among the pioneer public health laboratories that validated whole genome sequencing (WGS) for subtyping and further characterisation of pathogen isolates to replace laborious, more time-consuming traditional methods based on, for example, restriction fragment length polymorphisms[407], pulsed-field gel electrophoresis[408], phage[409] or multilocus sequence typing[410]. *Salmonella* was the first enteropathogen, for which surveillance was moved to a WGS basis in April 2015[254], followed by *Escherichia coli* [411], *Shigella*[412], *Listeria*[413], *Campylobacter*[414] and *Yersinia*[415]. By now, a multitude of other countries, including the USA[416] and many EU member states[417] have surveillance systems based on WGS. The technology is not limited to enteric pathogens but has also proven successful for, amongst others, *Mycobacterium tuberculosis*[418], *Neisseria meningitidis*[419] and *N. gonorrhoeae*[420].

Initial studies using WGS for outbreak detection were retrospective, analysing archived isolates, for example from a Shiga toxin-producing *E. coli* (STEC) O157

outbreak in 2013 associated with the consumption of watercress[421] or an outbreak of enterohaemorrhagic *E. coli* on an open farm in 2009[422]. Today, the established surveillance system at PHE, along with the bioinformatic analysis pipelines developed in-house[254,315], allows detection in real time: A STEC O157 outbreak in South West England was successfully traced back to consumption of raw cow's milk and WGS-based single nucleotide polymorphism typing was able to link a further four cases, who did not recall drinking the contaminated milk when questioned initially, to the outbreak[423]. WGS also aided the discovery of an *S*. Enteritidis outbreak caused by reptile feeder mice that would not have been picked up by conventional surveillance methods due to the geographical and temporal spread of cases[97]. Sharing of national surveillance data, coordinated by the European Centre for Disease Prevention and Control (ECDC), even allows outbreak investigation across borders, as illustrated by an *S*. Enteritidis outbreak traced from cases in England and other European countries to an egg producer in Germany[309,424]. As discussed in Chapter 3, culture-free metagenomic sequencing of *Salmonella*-positive stool samples did not result in the phylogenetic resolution provided by culture-dependent WGS that is required for outbreak investigations of this scale. This was mainly due to the fact that an insufficient number of *Salmonella*-specific reads could be identified over the background noise from the remaining gut microbiota community, leading to recovery of incomplete genomes and low genome coverage. Sufficient genome coverage can often only be obtained from the strains dominating the microbiota and the results presented in this study suggest that the pathogen does not necessarily become one of the major taxa, although its load is clearly sufficient to cause gastroenteritis symptoms, which complicates culture-independent methods of diagnosis.

Apart from tracking the exact identity of enteropathogens, some of their functional properties can also be inferred from WGS-based surveillance. Given the threat posed by antimicrobial resistance (AMR)[120], identification of antimicrobial resistance determinants (ARDs) in the sequenced genomes is one of the major pieces of additional information public health reference laboratories make use of. All EU member states as well as Iceland and Norway submit national AMR surveillance data to the European Antimicrobial Resistance Surveillance Network coordinated by the

ECDC. Based on this data, an annual report is issued detailing trends in AMR for different organisms, which helps countries to adapt treatment recommendations in order to minimise treatment failures, which is of particular importance in the case of invasive disease. It also allows a rapid response to the emergence of new resistance mechanisms. As discussed in Chapter 4, prediction of AMR profiles in non-typhoidal *S. enterica* (NTS) from WGS data is suitable as a replacement for traditional antimicrobial susceptibility testing, with predicted profiles being in very good agreement with isolate phenotypes. Similar results were obtained for a variety of other enteropathogens, including *S. sonnei*[274], *E. coli*[337] and typhoidal *Salmonellae*[273]. If turnaround times from initial diagnosis to availability of surveillance were decreased further, it would be plausible for WGS-based AMR profiling to guide treatment decisions. Again, for the reasons outlined above, metagenomic sequencing of *S. enterica*-positive stool samples did not provide sufficient resolution for the generation of AMR profiles matching those derived from the corresponding WGS.

While surveillance of AMR is generally based on data from infectious disease cases, the metagenomic study presented in Chapter 5, with a cohort of healthy travellers, illustrates that the gut microbiome can be a reservoir for ARDs even in the absence of disease. Since the colon is densely populated by microorganisms, horizontal gene transfer is facilitated by spatial proximity so that ARDs carried by commensals could be transferred to invading pathogens. Import of newly acquired ARDs by travellers could lead to their dissemination in the community. With international travel being commonplace, containing pathogens and their pathogenicity factors in their endemic environment is almost impossible in this day and age, as is sadly currently illustrated by the Coronavirus pandemic. Investigating transmission networks before they become an obvious public health concern, however, could lead to appropriate interventions being put into place in time and unbiased metagenomics is a useful tool for such studies.

6.1.3 Completion of thesis objectives

A summary of the five thesis objectives designed to achieve the thesis aim and the extent to which they were completed can be found below:

1. Use of 16S rRNA gene sequencing of the gut microbiota of patients with symptoms of acute gastroenteritis for culture-independent detection of aetiological agents of disease.

Direct detection of aetiological agents of gastroenteritis by 16S rRNA gene sequencing was partially successful for *Campylobacter* only.

2. Identification of pathogen-specific taxonomic and metabolic biomarkers of GI disease with the help of 16S rRNA gene sequencing and nuclear magnetic resonance spectroscopy.

Individual taxonomic and metabolic biomarkers associated with a specific pathogen could not be identified. A supervised learning approach, however, making use of various taxonomic and metabolic features, was able to distinguish controls, samples from patients diagnosed with a bacterial and those from patients diagnosed with a protozoal infection with an accuracy of 81.61%. Prediction of the exact aetiological agent was less accurate.

3. Detection and characterisation of *S. enterica* strains in stool samples derived from patients diagnosed with salmonellosis using metagenomic sequencing.

S. enterica was detected in 70% of samples by metagenomic sequencing. Sequencing resolution was neither high enough to establish phylogenetic relationships with corresponding isolate WGS or reference genomes nor for the generation of AMR profiles.

4. Evaluation of the sensitivity and specificity of WGS-derived AMR profile predictions compared to phenotypic AST for NTS isolates.

Predicted and phenotypic AMR profiles were in complete agreement for 97.82% of NTS isolates. Sensitivity of prediction exceeded 95% and specificity 98% for fourteen antimicrobials tested.

5. Examination of the prevalence and persistence of changes in gut microbiota composition and resistome induced by exposure to a foreign environment by conducting a longitudinal metagenomic study involving international travellers.

A 1.27-fold increase in the number of ARDs as well as increased carriage of potentially pathogenic species was detected after travel. Some of the changes persisted up to six months.

6.1.4 Study limitations and future work

Limitations of the study and suggestions for future work were discussed in detail in the previous chapters and are summarised again below:

The technology used to assess culture-free detection of enteropathogens in chapter 2, 16S rRNA gene sequencing, specifically targets a region in the prokaryotic genome. The study samples, however, were comprised of faeces from patients diagnosed with both bacterial and protozoal gastroenteritis, which excluded the possibility of direct pathogen detection for almost half of the samples. The results could therefore be complemented by sequencing the eukaryotic equivalent of the 16S rRNA marker gene, the 18S rRNA region. In addition to this, storage and processing of stool samples after collection was not identical, which could both mask or falsely attribute importance of certain taxonomic and metabolic features during biomarker discovery and Random Forest classification, only four of the many aetiological agents capable of causing gastroenteritis were studied in detail and sample numbers for these four aetiological agents were not equal. Furthermore, no information about the composition of the patient's pre- and post-infection microbiota was available and no information about patient demographics or lifestyle was recorded. Repeating the study with longitudinal sampling and further enteropathogens, ensuring identical sample storage and processing, matching sample numbers and accounting for factors other than infection that could influence microbiota composition would make the results more robust.

The limited success of using metagenomic sequencing for detection and characterisation of *S. enterica*, detailed in Chapter 3, is most likely due to the inability of the method to resolve the low-abundance pathogen signal in the presence of other high-abundance constituents of the microbiota. Enrichment of *S. enterica*, for example with the help of RNA baits, could amplify the pathogen-specific signal. In addition, a DNA extraction protocol developed specifically for metagenomic sequencing could improve results.

The validation of WGS-based AMR profiling in NTS described in Chapter 4 did not include azithromycin or colistin resistance, which are of increasing clinical relevance. Routine surveillance of resistance to these antimicrobials should be carried out to ensure that its prevalence is monitored.

Evaluation of ARD acquisition during international travel in Chapter 5 was purely genome-based. It is therefore uncertain whether the genetic potential for AMR is translated into functional proteins that confer resistance to antimicrobials. The results could therefore be corroborated using a transcriptomics approach or phenotypic AST. The study cohort was relatively small so that further stratification of samples to investigate the impact of the exact travel destination and of demographic and lifestyle factors on the results could not be carried out. Furthermore, some of the initial exclusion criteria had to be relaxed in order not to exclude too many volunteers from follow up. Future studies should therefore seek to recruit a larger number of volunteers.

In general, all parts of this thesis focussed on bacterial members of the gut microbiota community. However, the importance of the mycobiome[425,426] and the virome[427] is increasingly recognised and further work should include an analysis of the contribution of these microorganisms to GI health.

6.2 The role of next-generation sequencing technologies in healthcare

Although the results presented in this thesis suggest further work is needed before next-generation sequencing (NGS) technologies can replace culture- and PCR-based methods for diagnosis of gastrointestinal infections, they are already in routine use in other areas of the healthcare sector. While the Human Genome Project took almost fifteen years to complete, providing a first draft of the human genome at a price of approximately $2.7 billion[428], now a complete genome sequence can be generated within a day. Instead of targeting individual mutation-prone exons, the UK's National Health Service (NHS) offers whole gene screens, currently sequenced with Illumina® technology, to look for pathogenic variants in any region of a gene. The most recent National Genomic Test Directory for both rare and inherited disease and cancer specifies whole exome or even whole genome sequencing as the method of

choice for a variety of conditions including, for example, congenital muscular dystrophy and leukaemia. In 2019, Health Secretary Matt Hancock even suggested that WGS would routinely be offered to all children at birth in the near future to predict susceptibility to inherited diseases. The estimated cost of sequencing a human genome in an NHS laboratory ranges from £2,350-£3,420[429].

Metagenomic sequencing is already in routine use for diagnosis of meningitis and encephalitis from cerebrospinal fluid at a clinical laboratory in the United States[239,242] and similar assays are currently being developed for diagnosis of sepsis and pneumonia as well as Lyme disease and other tickborne infections. The first genome sequence of SARS-CoV-2, the virus causing the current global COVID-19 pandemic, was made publicly available on 10th January 2020[430], less than two weeks after an outbreak of pneumonia of unknown aetiology had first been reported in China. The genome forms the basis of research efforts aimed at developing a vaccine and appropriate treatments and it could not have been made available at such unprecedented speed without the use of NGS. Recently, the COVID-19 Genomics UK Consortium was announced, a collaboration between the NHS, public health agencies, the Wellcome Sanger Institute and academic institutions in 13 UK cities, which promises rapid, large-scale sequencing of the virus from confirmed national cases to monitor its spread and evolution.

6.2.1 Challenges associated with the use of next-generation sequencing in routine healthcare

Although they can be of great use, the biggest hurdle for the use of NGS approaches in all diagnostic laboratories is the infrastructure required for their implementation[431]. Not only the sequencing platforms themselves but also the computational infrastructure for data analysis requires space. Quite often, efficient data analysis can only be guaranteed when a high-performance computing cluster is available. A shared resource, similar to the Medical Research Council's Cloud Infrastructure for Microbial Bioinformatics used by researchers to perform microbial bioinformatics[34], for computing and storage of clinical data might be a solution. On top of this, although sequencing costs have decreased vastly compared to initial use, they are still prohibitive to widespread use in many healthcare settings[432]. This holds true

especially for developing regions, where the incidence of infectious disease is higher than in industrialised countries and often associated with increased mortality rates due to insufficient provision of healthcare. Thus, areas that would benefit the most from good surveillance systems are the ones most likely to not have access to them. Some of these caveats could be overcome by using the Oxford Nanopore instead of Illumina® technology: The Nanopore MinION is a portable device that can be plugged into a laptop for real-time display of sequencing results, making it suitable for processing samples in the field, and can be purchased for a fraction of the price of an Illumina® machine. The technology has successfully been used for rapid on-site sequencing during the Ebola outbreak in Africa[433] and the Zika virus outbreak in Brazil[434]. In the UK, Nanopore sequencing of the gut microbiota of preterm infants allowed detection of enteropathogens and associated AMR profiles within 1 hour of sequencing[435]. However, the Nanopore approach is prone to sequencing errors and has a lower throughput, which increases its per-read cost compared to other platforms[436].

The amount of data generated by next generation sequencing poses another problem: Unlike a simple PCR, where a signal indicates a positive sample, patterns in sequencing data cannot be visualised without complex processing. This requires skilled, bioinformatically trained staff. Only a few clinical laboratories have such staff at their disposal. Implementation of NGS-based assays therefore needs to be accompanied by either building a bioinformatics workforce or training existing staff in data analysis. Alternatively, analysis workflows could be automated so that laboratory staff only see the final, simplified output, which means that they would not be equipped to troubleshoot any issues with the process. On top of this, while bioinformaticians might have the skills to detect patterns in the data, their clinical relevance can only be determined by staff with knowledge of other areas such as microbiology and physiology. As such, the so-called genomic revolution has probably encouraged interaction between scientific disciplines like few other advances in technology.

Finally, the results of an NGS-based assay might vary depending on the analysis pipeline used and the database this pipeline bases its output on. Common clinically relevant pathogens tend to be over-represented in reference databases so that their

detection and profiling is likely to be more reliable than that of less frequently encountered ones[135]. New algorithms for processing of NGS data are constantly being developed. To ensure that results are comparable between laboratories, both nationally and internationally, regulatory bodies need to provide guidance on appropriate validation of assays, similar to the report issued on the role of WGS in AST of bacteria by a EUCAST subcommittee[349]. Again, a shared computational resource for clinical laboratories could ensure that the same analysis tools are available to all.

6.3 Conclusion

The research presented in this thesis highlights that, while '–omics' technologies hold great potential for the future of diagnosis and surveillance of gastrointestinal infections, some approaches need further development work before they can be put into routine clinical use. Genome-based resistance profiling of enteropathogen isolates has already successfully replaced phenotypic testing for monitoring trends in AMR and metagenomic sequencing and can be utilised to study AMR transmission networks in healthy populations. When it comes to clinical decision making, however, culture-independent testing currently still fails to achieve the performance of the gold standard methods. Since technologies and data analysis tools are constantly under development to address their shortfalls, it is highly plausible that '–omics' technologies will soon find their way into the frontline clinical diagnostic setting, even for conditions affecting such complex microbial communities as the gut microbiota.

Bibliography

1. Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature.* 2010;464(7285):59-65.
2. Luckey TD. Introduction to intestinal microecology. *Am J Clin Nutr.* 1972;25(12):1292-1294.
3. Savage DC. Microbial ecology of the gastrointestinal tract. *Annu Rev Microbiol.* 1977;31:107-133.
4. Sender R, Fuchs S, Milo R. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell.* 2016;164(3):337-340.
5. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol.* 2016;14(8):e1002533.
6. Lederberg J, McCray A. 'Ome Sweet 'Omics - A Genealogical Treasury of Words. In. Vol 15. Scientist2001:8.
7. Grice EA, Kong HH, Conlan S, et al. Topographical and temporal diversity of the human skin microbiome. *Science.* 2009;324(5931):1190-1192.
8. Aas JA, Paster BJ, Stokes LN, Olsen I, Dewhirst FE. Defining the normal bacterial flora of the oral cavity. *J Clin Microbiol.* 2005;43(11):5721-5732.
9. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature.* 2012;486(7402):207-214.
10. Jiménez E, Marín ML, Martín R, et al. Is meconium from healthy newborns actually sterile? *Res Microbiol.* 2008;159(3):187-193.
11. Moles L, Gómez M, Heilig H, et al. Bacterial diversity in meconium of preterm neonates and evolution of their fecal microbiota during the first month of life. *PLoS One.* 2013;8(6):e66986.
12. Yatsunenko T, Rey FE, Manary MJ, et al. Human gut microbiome viewed across age and geography. *Nature.* 2012;486(7402):222-227.
13. Palmer C, Bik EM, DiGiulio DB, Relman DA, Brown PO. Development of the human infant intestinal microbiota. *PLoS Biol.* 2007;5(7):e177.
14. Dominguez-Bello MG, Costello EK, Contreras M, et al. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc Natl Acad Sci U S A.* 2010;107(26):11971-11975.

15. Aagaard K, Riehle K, Ma J, et al. A metagenomic approach to characterization of the vaginal microbiome signature in pregnancy. *PLoS One.* 2012;7(6):e36466.
16. Bennet R, Nord CE. Development of the faecal anaerobic microflora after caesarean section and treatment with antibiotics in newborn infants. *Infection.* 1987;15(5):332-336.
17. Jakobsson HE, Abrahamsson TR, Jenmalm MC, et al. Decreased gut microbiota diversity, delayed Bacteroidetes colonisation and reduced Th1 responses in infants delivered by caesarean section. *Gut.* 2014;63(4):559-566.
18. Barrett E, Kerr C, Murphy K, et al. The individual-specific and diverse nature of the preterm infant microbiota. *Arch Dis Child Fetal Neonatal Ed.* 2013;98(4):F334-340.
19. Arboleya S, Binetti A, Salazar N, et al. Establishment and development of intestinal microbiota in preterm neonates. *FEMS Microbiol Ecol.* 2012;79(3):763-772.
20. Hunt KM, Foster JA, Forney LJ, et al. Characterization of the diversity and temporal stability of bacterial communities in human milk. *PLoS One.* 2011;6(6):e21313.
21. Cabrera-Rubio R, Collado MC, Laitinen K, Salminen S, Isolauri E, Mira A. The human milk microbiome changes over lactation and is shaped by maternal weight and mode of delivery. *Am J Clin Nutr.* 2012;96(3):544-551.
22. Zivkovic AM, German JB, Lebrilla CB, Mills DA. Human milk glycobiome and its impact on the infant gastrointestinal microbiota. *Proc Natl Acad Sci U S A.* 2011;108 Suppl 1:4653-4658.
23. Koenig JE, Spor A, Scalfone N, et al. Succession of microbial consortia in the developing infant gut microbiome. *Proc Natl Acad Sci U S A.* 2011;108 Suppl 1:4578-4585.
24. Faith JJ, Guruge JL, Charbonneau M, et al. The long-term stability of the human gut microbiota. *Science.* 2013;341(6141):1237439.
25. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. Bacterial community variation in human body habitats across space and time. *Science.* 2009;326(5960):1694-1697.

26. Eckburg PB, Bik EM, Bernstein CN, et al. Diversity of the human intestinal microbial flora. *Science.* 2005;308(5728):1635-1638.
27. Erwin G. Zoetendal ADLAWMA-vVJAGMdVWMdV. The Host Genotype Affects the Bacterial Community in the Human Gastronintestinal Tract. *Microbial Ecology in Health and Disease.* 2001;13(3):129-134.
28. Stewart JA, Chadwick VS, Murray A. Investigations into the influence of host genetics on the predominant eubacteria in the faecal microflora of children. *J Med Microbiol.* 2005;54(Pt 12):1239-1242.
29. Turnbaugh PJ, Hamady M, Yatsunenko T, et al. A core gut microbiome in obese and lean twins. *Nature.* 2009;457(7228):480-484.
30. De Filippo C, Cavalieri D, Di Paola M, et al. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc Natl Acad Sci U S A.* 2010;107(33):14691-14696.
31. De Filippis F, Pellegrini N, Vannini L, et al. High-level adherence to a Mediterranean diet beneficially impacts the gut microbiota and associated metabolome. *Gut.* 2016;65(11):1812-1821.
32. Schnorr SL, Candela M, Rampelli S, et al. Gut microbiome of the Hadza hunter-gatherers. *Nat Commun.* 2014;5:3654.
33. Barton W, Penney NC, Cronin O, et al. The microbiome of professional athletes differs from that of more sedentary subjects in composition and particularly at the functional metabolic level. *Gut.* 2018;67(4):625-633.
34. Claesson MJ, Jeffery IB, Conde S, et al. Gut microbiota composition correlates with diet and health in the elderly. *Nature.* 2012;488(7410):178-184.
35. Biagi E, Nylund L, Candela M, et al. Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. *PLoS One.* 2010;5(5):e10667.
36. Drago L, Toscano M, Rodighiero V, De Vecchi E, Mogna G. Cultivable and pyrosequenced fecal microflora in centenarians and young subjects. *J Clin Gastroenterol.* 2012;46 Suppl:S81-84.
37. Arumugam M, Raes J, Pelletier E, et al. Enterotypes of the human gut microbiome. *Nature.* 2011;473(7346):174-180.

38. Knights D, Ward TL, McKinlay CE, et al. Rethinking "enterotypes". *Cell Host Microbe.* 2014;16(4):433-437.
39. Segain JP, Raingeard de la Blétière D, Bourreille A, et al. Butyrate inhibits inflammatory responses through NFkappaB inhibition: implications for Crohn's disease. *Gut.* 2000;47(3):397-403.
40. Säemann MD, Böhmig GA, Osterreicher CH, et al. Anti-inflammatory effects of sodium butyrate on human monocytes: potent inhibition of IL-12 and up-regulation of IL-10 production. *FASEB J.* 2000;14(15):2380-2382.
41. Waldecker M, Kautenburger T, Daumann H, Busch C, Schrenk D. Inhibition of histone-deacetylase activity by short-chain fatty acids and some polyphenol metabolites formed in the colon. *J Nutr Biochem.* 2008;19(9):587-593.
42. Donohoe DR, Holley D, Collins LB, et al. A gnotobiotic mouse model demonstrates that dietary fiber protects against colorectal tumorigenesis in a microbiota- and butyrate-dependent manner. *Cancer Discov.* 2014;4(12):1387-1397.
43. Fleming SE, Fitch MD, DeVries S, Liu ML, Kight C. Nutrient utilization by cells isolated from rat jejunum, cecum and colon. *J Nutr.* 1991;121(6):869-878.
44. Roediger WE. Utilization of nutrients by isolated epithelial cells of the rat colon. *Gastroenterology.* 1982;83(2):424-429.
45. Donohoe DR, Collins LB, Wali A, Bigler R, Sun W, Bultman SJ. The Warburg effect dictates the mechanism of butyrate-mediated histone acetylation and cell proliferation. *Mol Cell.* 2012;48(4):612-626.
46. Peng L, Li ZR, Green RS, Holzman IR, Lin J. Butyrate enhances the intestinal barrier by facilitating tight junction assembly via activation of AMP-activated protein kinase in Caco-2 cell monolayers. *J Nutr.* 2009;139(9):1619-1625.
47. den Besten G, Lange K, Havinga R, et al. Gut-derived short-chain fatty acids are vividly assimilated into host carbohydrates and lipids. *Am J Physiol Gastrointest Liver Physiol.* 2013;305(12):G900-910.
48. den Besten G, Bleeker A, Gerding A, et al. Short-Chain Fatty Acids Protect Against High-Fat Diet-Induced Obesity via a PPARγ-Dependent Switch From Lipogenesis to Fat Oxidation. *Diabetes.* 2015;64(7):2398-2408.

49. Frost G, Sleeth ML, Sahuri-Arisoylu M, et al. The short-chain fatty acid acetate reduces appetite via a central homeostatic mechanism. *Nat Commun.* 2014;5:3611.
50. Hüfner E, Britton RA, Roos S, Jonsson H, Hertel C. Global transcriptional response of Lactobacillus reuteri to the sourdough environment. *Syst Appl Microbiol.* 2008;31(5):323-338.
51. Pompei A, Cordisco L, Amaretti A, Zanoni S, Matteuzzi D, Rossi M. Folate production by bifidobacteria as a potential probiotic property. *Appl Environ Microbiol.* 2007;73(1):179-185.
52. Fukiya S, Arata M, Kawashima H, et al. Conversion of cholic acid and chenodeoxycholic acid into their 7-oxo derivatives by Bacteroides intestinalis AM-1 isolated from human feces. *FEMS Microbiol Lett.* 2009;293(2):263-270.
53. Hirokane H, Nakahara M, Tachibana S, Shimizu M, Sato R. Bile acid reduces the secretion of very low density lipoprotein by repressing microsomal triglyceride transfer protein gene expression mediated by hepatocyte nuclear factor-4. *J Biol Chem.* 2004;279(44):45685-45692.
54. Watanabe M, Houten SM, Mataki C, et al. Bile acids induce energy expenditure by promoting intracellular thyroid hormone activation. *Nature.* 2006;439(7075):484-489.
55. Fang Y, Studer E, Mitchell C, et al. Conjugated bile acids regulate hepatocyte glycogen synthase activity in vitro and in vivo via Galphai signaling. *Mol Pharmacol.* 2007;71(4):1122-1128.
56. De Angelis M, Ferrocino I, Calabrese FM, et al. Diet influences the functions of the human intestinal microbiome. *Sci Rep.* 2020;10(1):4247.
57. van der Waaij D, Berghuis-de Vries JM, Lekkerkerk Lekkerkerk-v. Colonization resistance of the digestive tract in conventional and antibiotic-treated mice. *J Hyg (Lond).* 1971;69(3):405-411.
58. Ng KM, Ferreyra JA, Higginbottom SK, et al. Microbiota-liberated host sugars facilitate post-antibiotic expansion of enteric pathogens. *Nature.* 2013;502(7469):96-99.

59. Battaglioli EJ, Hale VL, Chen J, et al. Clostridioides difficile uses amino acids associated with gut microbial dysbiosis in a subset of patients with diarrhea. *Sci Transl Med.* 2018;10(464).
60. Hryckowian AJ, Van Treuren W, Smits SA, et al. Microbiota-accessible carbohydrates suppress Clostridium difficile infection in a murine model. *Nat Microbiol.* 2018;3(6):662-669.
61. Wlodarska M, Willing BP, Bravo DM, Finlay BB. Phytonutrient diet supplementation promotes beneficial Clostridia species and intestinal mucus secretion resulting in protection against enteric infection. *Sci Rep.* 2015;5:9253.
62. Garcia-Gutierrez E, Mayer MJ, Cotter PD, Narbad A. Gut microbiota as a source of novel antimicrobials. *Gut Microbes.* 2019;10(1):1-21.
63. Aoki SK, Pamma R, Hernday AD, Bickham JE, Braaten BA, Low DA. Contact-dependent inhibition of growth in Escherichia coli. *Science.* 2005;309(5738):1245-1248.
64. Russell AB, Hood RD, Bui NK, LeRoux M, Vollmer W, Mougous JD. Type VI secretion delivers bacteriolytic effectors to target cells. *Nature.* 2011;475(7356):343-347.
65. Jacobson A, Lam L, Rajendram M, et al. A Gut Commensal-Produced Metabolite Mediates Colonization Resistance to Salmonella Infection. *Cell Host Microbe.* 2018;24(2):296-307.e297.
66. Buffie CG, Bucci V, Stein RR, et al. Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile. *Nature.* 2015;517(7533):205-208.
67. Ridge JP, Fuchs EJ, Matzinger P. Neonatal tolerance revisited: turning on newborn T cells with dendritic cells. *Science.* 1996;271(5256):1723-1726.
68. Lathrop SK, Bloom SM, Rao SM, et al. Peripheral education of the immune system by colonic commensal microbiota. *Nature.* 2011;478(7368):250-254.
69. Al Nabhani Z, Dulauroy S, Marques R, et al. A Weaning Reaction to Microbiota Is Required for Resistance to Immunopathologies in the Adult. *Immunity.* 2019;50(5):1276-1288.e1275.

70. Dethlefsen L, Huse S, Sogin ML, Relman DA. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLoS Biol.* 2008;6(11):e280.
71. Jernberg C, Löfmark S, Edlund C, Jansson JK. Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. *ISME J.* 2007;1(1):56-66.
72. Britton RA, Young VB. Role of the intestinal microbiota in resistance to colonization by Clostridium difficile. *Gastroenterology.* 2014;146(6):1547-1553.
73. Li YT, Cai HF, Wang ZH, Xu J, Fang JY. Systematic review with meta-analysis: long-term outcomes of faecal microbiota transplantation for Clostridium difficile infection. *Aliment Pharmacol Ther.* 2016;43(4):445-457.
74. Fouhy F, Guinane CM, Hussey S, et al. High-throughput sequencing reveals the incomplete, short-term recovery of infant gut microbiota following parenteral antibiotic treatment with ampicillin and gentamicin. *Antimicrob Agents Chemother.* 2012;56(11):5811-5820.
75. Russell SL, Gold MJ, Hartmann M, et al. Early life antibiotic-driven changes in microbiota enhance susceptibility to allergic asthma. *EMBO Rep.* 2012;13(5):440-447.
76. Turnbaugh PJ, Gordon JI. The core gut microbiome, energy balance and obesity. *J Physiol.* 2009;587(Pt 17):4153-4158.
77. Cani PD, Possemiers S, Van de Wiele T, et al. Changes in gut microbiota control inflammation in obese mice through a mechanism involving GLP-2-driven improvement of gut permeability. *Gut.* 2009;58(8):1091-1103.
78. Mouzaki M, Comelli EM, Arendt BM, et al. Intestinal microbiota in patients with nonalcoholic fatty liver disease. *Hepatology.* 2013;58(1):120-127.
79. Cadwell K, Patel KK, Maloney NS, et al. Virus-plus-susceptibility gene interaction determines Crohn's disease gene Atg16L1 phenotypes in intestine. *Cell.* 2010;141(7):1135-1145.
80. Noor SO, Ridgway K, Scovell L, et al. Ulcerative colitis and irritable bowel patients exhibit distinct abnormalities of the gut microbiota. *BMC Gastroenterol.* 2010;10:134.

81. Jiang H, Ling Z, Zhang Y, et al. Altered fecal microbiota composition in patients with major depressive disorder. *Brain Behav Immun.* 2015;48:186-194.
82. Jiang HY, Zhang X, Yu ZH, et al. Altered gut microbiota profile in patients with generalized anxiety disorder. *J Psychiatr Res.* 2018;104:130-136.
83. Adams JB, Johansen LJ, Powell LD, Quig D, Rubin RA. Gastrointestinal flora and gastrointestinal status in children with autism--comparisons to typical children and correlation with autism severity. *BMC Gastroenterol.* 2011;11:22.
84. Carabotti M, Scirocco A, Maselli MA, Severi C. The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems. *Ann Gastroenterol.* 2015;28(2):203-209.
85. Fhogartaigh C, Dance D. Bacterial gastroenteritis. In. Vol 41. Medicine2013:693-699.
86. Public Health England. UK Standards for Microbiology Investigations - Investigation of Faecal Specimens for Enteric Pathogens. 2014.
87. Joensen KG, Engsbro ALØ, Lukjancenko O, et al. Evaluating next-generation sequencing for direct clinical diagnostics in diarrhoeal disease. *Eur J Clin Microbiol Infect Dis.* 2017;36(7):1325-1338.
88. Kosek M, Bern C, Guerrant RL. The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000. *Bull World Health Organ.* 2003;81(3):197-204.
89. Checkley W, Buckley G, Gilman RH, et al. Multi-country analysis of the effects of diarrhoea on childhood stunting. *Int J Epidemiol.* 2008;37(4):816-830.
90. Tam CC, Rodrigues LC, Viviani L, et al. Longitudinal study of infectious intestinal disease in the UK (IID2 study): incidence in the community and presenting to general practice. *Gut.* 2012;61(1):69-77.
91. Roberts JA, Cumberland P, Sockett PN, et al. The study of infectious intestinal disease in England: socio-economic impact. *Epidemiol Infect.* 2003;130(1):1-11.
92. Desselberger U. Rotaviruses. *Virus Res.* 2014;190:75-96.

93. Sandmann FG, Shallcross L, Adams N, et al. Estimating the Hospital Burden of Norovirus-Associated Gastroenteritis in England and Its Opportunity Costs for Nonadmitted Patients. *Clin Infect Dis.* 2018;67(5):693-700.
94. Goddard EA, Lastovica AJ, Argent AC. Campylobacter 0:41 isolation in Guillain-Barré syndrome. *Arch Dis Child.* 1997;76(6):526-528.
95. Ali M, Nelson AR, Lopez AL, Sack DA. Updated global burden of cholera in endemic countries. *PLoS Negl Trop Dis.* 2015;9(6):e0003832.
96. Samore MH. Epidemiology of nosocomial clostridium difficile diarrhoea. *J Hosp Infect.* 1999;43 Suppl:S183-190.
97. Kanagarajah S, Waldram A, Dolan G, et al. Whole genome sequencing reveals an outbreak of Salmonella Enteritidis associated with reptile feeder mice in the United Kingdom, 2012-2015. *Food Microbiology.* 2017;In Press.
98. Behravesh CB, Ferraro A, Deasy M, et al. Human Salmonella infections linked to contaminated dry dog and cat food, 2006-2008. *Pediatrics.* 2010;126(3):477-483.
99. Byrne L, Adams N, Glen K, et al. Epidemiological and Microbiological Investigation of an Outbreak of Severe Disease from Shiga Toxin-Producing Escherichia coli O157 Infection Associated with Consumption of a Slaw Garnish. *J Food Prot.* 2016;79(7):1161-1168.
100. Clark CG, Price L, Ahmed R, et al. Characterization of waterborne outbreak-associated Campylobacter jejuni, Walkerton, Ontario. *Emerg Infect Dis.* 2003;9(10):1232-1241.
101. Gormley FJ, Little CL, Rawal N, Gillespie IA, Lebaigue S, Adak GK. A 17-year review of foodborne outbreaks: describing the continuing decline in England and Wales (1992-2008). *Epidemiol Infect.* 2011;139(5):688-699.
102. O'Brien SJ. The "decline and fall" of nontyphoidal salmonella in the United kingdom. *Clin Infect Dis.* 2013;56(5):705-710.
103. Scanlan PD, Stensvold CR, Rajilic-Stojanovic M, et al. The microbial eukaryote Blastocystis is a prevalent and diverse member of the healthy human gut microbiota. *FEMS Microbiol Ecol.* 2014;90(1):326-330.

104. Parfrey LW, Walters WA, Lauber CL, et al. Communities of microbial eukaryotes in the mammalian gut within the context of environmental eukaryotic diversity. *Front Microbiol.* 2014;5:298.
105. Lukes J, Stensvold CR, Jirku-Pomajbikova K, Wegener Parfrey L. Are Human Intestinal Eukaryotes Beneficial or Commensals? *PLoS Pathog.* 2015;11(8):e1005039.
106. Fletcher SM, Stark D, Harkness J, Ellis J. Enteric protozoa in the developed world: a public health perspective. *Clin Microbiol Rev.* 2012;25(3):420-449.
107. Tan KS, Mirza H, Teo JD, Wu B, Macary PA. Current Views on the Clinical Relevance of Blastocystis spp. *Curr Infect Dis Rep.* 2010;12(1):28-35.
108. Barratt JL, Harkness J, Marriott D, Ellis JT, Stark D. A review of Dientamoeba fragilis carriage in humans: several reasons why this organism should be considered in the diagnosis of gastrointestinal illness. *Gut Microbes.* 2011;2(1):3-12.
109. Döller PCD, Karl; Filipp, Norbert; Brockmann, Stefan; Dreweck, Caroline: Vonthein, Reinhard; Wagner-Wiening, Christiane; Wiedenmann, Albrecht. Cyclosporiasis outbreak in Germany associated with the consumption of salad. *Emerg Infect Dis.* 2002;8(9):992-994.
110. Ho AYL, Adriana S.; Eberhart, Michael G.; Levenson, Robert; Finkel, Bernard S.; da Silva, Alexandre J.; Roberts, Jacquelin M.; Orlandi, Palmer A.; Johnson, Caroline C.; Herwald, Barbara L. . Outbreak of cyclosporiasis associated with imported raspberries, Philadelphia, Pennsylvania, 2000. *Emerg Infect Dis.* 2002;8(8):783-788.
111. Karanis P, Kourenti C, Smith H. Waterborne transmission of protozoan parasites: a worldwide review of outbreaks and lessons learnt. *J Water Health.* 2007;5(1):1-38.
112. Stuart JM, Orr HJ, Warburton FG, et al. Risk factors for sporadic giardiasis: a case-control study in southwestern England. *Emerg Infect Dis.* 2003;9(2):229-233.
113. Feng Y, Xiao L. Zoonotic potential and molecular epidemiology of Giardia species and giardiasis. *Clin Microbiol Rev.* 2011;24(1):110-140.

114. Rogawski ET, Bartelt LA, Platts-Mills JA, et al. Determinants and Impact of Giardia Infection in the First 2 Years of Life in the MAL-ED Birth Cohort. *J Pediatric Infect Dis Soc.* 2017;6(2):153-160.
115. McLauchlin JA, Corinne; Pedraza-Diaz, S.; Nichols, G. S. Molecular epidemiological analysis of Cryptosporidium spp. in the United Kingdom: Results of genotyping Cryptosporidium spp. in 1,705 fecal samples from humans and 105 fecal samples from livestock animals. *Journal of Clinical Microbiology.* 2000;38(11):3984-3990.
116. Hunter PRH, S.; Woodhouse, S.; Syed, Q.; Verlander, N. Q.; Chalmers, R. M.; Morgan, K.; Nichols, G.; Beeching, N.; Osborn, K. Sporadic cryptosporidiosis case-control study with genotyping. *Emerg Infect Dis.* 2004;10(7):1241-1249.
117. Zylberberg HM, Green PH, Turner KO, Genta RM, Lebwohl B. Prevalence and Predictors of Giardia in the United States. *Dig Dis Sci.* 2017;62(2):432-440.
118. Laupland KB, Church DL. Population-based laboratory surveillance for Giardia sp. and Cryptosporidium sp. infections in a large Canadian health region. *BMC Infect Dis.* 2005;5:72.
119. Rosenblatt-Farrell N. The landscape of antibiotic resistance. *Environ Health Perspect.* 2009;117(6):A244-250.
120. The Review on Antimicrobial Resistance. Tackling drug-resistant infections globally: Final report and recommendations. 2016.
121. Baym M, Lieberman TD, Kelsic ED, et al. Spatiotemporal microbial evolution on antibiotic landscapes. *Science.* 2016;353(6304):1147-1151.
122. World Health Organisation. Prioritization of pathogens to guide discovery, research and development of new antibiotics for drug-resistant bacterial infections, including tuberculosis. 2017.
123. Ruimy R, Brisabois A, Bernede C, et al. Organic and conventional fruits and vegetables contain equivalent counts of Gram-negative bacteria expressing resistance to antibacterial agents. *Environ Microbiol.* 2010;12(3):608-615.
124. De Boeck H, Miwanda B, Lunguya-Metila O, et al. ESBL-positive Enterobacteria isolates in drinking water. *Emerg Infect Dis.* 2012;18(6):1019-1020.

125. Pehrsson EC, Tsukayama P, Patel S, et al. Interconnected microbiomes and resistomes in low-income human habitats. *Nature.* 2016;533(7602):212-216.
126. Gullberg E, Albrecht LM, Karlsson C, Sandegren L, Andersson DI. Selection of a multidrug resistance plasmid by sublethal levels of antibiotics and heavy metals. *mBio.* 2014;5(5):e01918-01914.
127. Guenther S, Wuttke J, Bethe A, et al. Is fecal carriage of extended-spectrum-β-lactamase-producing Escherichia coli in urban rats a risk for public health? *Antimicrob Agents Chemother.* 2013;57(5):2424-2425.
128. Hansen TA, Joshi T, Larsen AR, et al. Vancomycin gene selection in the microbiome of urban Rattus norvegicus from hospital environment. *Evol Med Public Health.* 2016;2016(1):219-226.
129. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature.* 2008;456(7218):53-59.
130. Greninger AL, Naccache SN, Federman S, et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med.* 2015;7:99.
131. Neefs JM, Van de Peer Y, De Rijk P, Chapelle S, De Wachter R. Compilation of small ribosomal subunit RNA structures. *Nucleic Acids Res.* 1993;21(13):3025-3049.
132. Nicholson JK, Lindon JC, Holmes E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica.* 1999;29(11):1181-1189.
133. Bezabeh T, Somorjai RL, Smith IC. MR metabolomics of fecal extracts: applications in the study of bowel diseases. *Magn Reson Chem.* 2009;47 Suppl 1:S54-61.
134. Guerrant RL, Van Gilder T, Steiner TS, et al. Practice guidelines for the management of infectious diarrhea. *Clin Infect Dis.* 2001;32(3):331-351.
135. Simner PJ, Miller S, Carroll KC. Understanding the Promises and Hurdles of Metagenomic Next-Generation Sequencing as a Diagnostic Tool for Infectious Diseases. *Clin Infect Dis.* 2018;66(5):778-788.

136. Humphries RM, Linscott AJ. Laboratory diagnosis of bacterial gastroenteritis. *Clin Microbiol Rev.* 2015;28(1):3-31.
137. Lewis SJ, Heaton KW. Stool form scale as a useful guide to intestinal transit time. *Scand J Gastroenterol.* 1997;32(9):920-924.
138. Nye KJ, Fallon D, Frodsham D, et al. Comparison of the performance of lactose and mannitol selenite enriched broths, subcultured to DCA and XLD agars, in the isolation of Salmonella spp. from faeces. *Commun Dis Public Health.* 2002;5(4):285-288.
139. Public Health England. UK Standards for Microbiology Investigations - Identification of Shigella species. 2015.
140. He Y, Li H, Lu X, Stratton CW, Tang YW. Mass spectrometry biotyper system identifies enteric bacterial pathogens directly from colonies grown on selective stool culture media. *J Clin Microbiol.* 2010;48(11):3888-3892.
141. Public Health England. UK Standards for Microbiology Investigations - Staining procedures. 2019.
142. Public Health England. UK Standards for Microbiology Investigations - Processing of Faeces for Clostridium difficile. 2018.
143. Wohlwend N, Tiermann S, Risch L, Risch M, Bodmer T. Evaluation of a Multiplex Real-Time PCR Assay for Detecting Major Bacterial Enteric Pathogens in Fecal Specimens: Intestinal Inflammation and Bacterial Load Are Correlated in Campylobacter Infections. *J Clin Microbiol.* 2016;54(9):2262-2266.
144. Martín A, Pérez-Ayala A, Chaves F, Lora D, Orellana M. Evaluation of the multiplex PCR Allplex-GI assay in the detection of bacterial pathogens in diarrheic stool samples. *J Microbiol Methods.* 2018;144:33-36.
145. Stokes W, Simner PJ, Mortensen J, et al. Multicenter Clinical Validation of the Molecular BD Max Enteric Viral Panel for Detection of Enteric Pathogens. *J Clin Microbiol.* 2019;57(9).
146. Madison-Antenucci S, Relich RF, Doyle L, et al. Multicenter Evaluation of BD Max Enteric Parasite Real-Time PCR Assay for Detection of Giardia duodenalis, Cryptosporidium hominis, Cryptosporidium parvum, and Entamoeba histolytica. *J Clin Microbiol.* 2016;54(11):2681-2688.

147. Won EJ, Kim SH, Kee SJ, et al. Multiplex Real-Time PCR Assay Targeting Eight Parasites Customized to the Korean Population: Potential Use for Detection in Diarrheal Stool Samples from Gastroenteritis Patients. *PLoS One.* 2016;11(11):e0166957.
148. Claas EC, Burnham CA, Mazzulli T, Templeton K, Topin F. Performance of the xTAG® gastrointestinal pathogen panel, a multiplex molecular assay for simultaneous detection of bacterial, viral, and parasitic causes of infectious gastroenteritis. *J Microbiol Biotechnol.* 2013;23(7):1041-1045.
149. Spina A, Kerr KG, Cormican M, et al. Spectrum of enteropathogens detected by the FilmArray GI Panel in a multicentre study of community-acquired gastroenteritis. *Clin Microbiol Infect.* 2015;21(8):719-728.
150. McAuliffe G, Bissessor L, Williamson D, et al. Use of the EntericBio Gastro Panel II in a diagnostic microbiology laboratory: challenges and opportunities. *Pathology.* 2017;49(4):419-422.
151. Ake JA, Jelacic S, Ciol MA, et al. Relative nephroprotection during Escherichia coli O157:H7 infections: association with intravenous volume expansion. *Pediatrics.* 2005;115(6):e673-680.
152. Liu J, Almeida M, Kabir F, et al. Direct Detection of Shigella in Stool Specimens by Use of a Metagenomic Approach. *J Clin Microbiol.* 2018;56(2).
153. Abubakar I, Irvine L, Aldus CF, et al. A systematic review of the clinical, public health and cost-effectiveness of rapid diagnostic tests for the detection and identification of bacterial intestinal pathogens in faeces and food. *Health Technol Assess.* 2007;11(36):1-216.
154. Harrington SM, Buchan BW, Doern C, et al. Multicenter evaluation of the BD max enteric bacterial panel PCR assay for rapid detection of Salmonella spp., Shigella spp., Campylobacter spp. (C. jejuni and C. coli), and Shiga toxin 1 and 2 genes. *J Clin Microbiol.* 2015;53(5):1639-1647.
155. Persson S, Petersen HM, Jespersgaard C, Olsen KE. Real-time TaqMan polymerase chain reaction-based genus-identification and pyrosequencing-based species identification of Campylobacter jejuni, C. coli, C. lari, C. upsaliensis, and C. fetus directly on stool samples. *Diagn Microbiol Infect Dis.* 2012;74(1):6-10.

156. Porte L, García P, Braun S, et al. Head-to-head comparison of Microflex LT and Vitek MS systems for routine identification of microorganisms by MALDI-TOF mass spectrometry in Chile. *PLoS One.* 2017;12(5):e0177929.
157. McHardy IH, Wu M, Shimizu-Cohen R, Couturier MR, Humphries RM. Detection of intestinal protozoa in the clinical laboratory. *J Clin Microbiol.* 2014;52(3):712-720.
158. Planche T, Aghaizu A, Holliman R, et al. Diagnosis of Clostridium difficile infection by toxin detection kits: a systematic review. *Lancet Infect Dis.* 2008;8(12):777-784.
159. Myers AL, Jackson MA, Selvarangan R. False-positive results of Campylobacter rapid antigen testing. *Pediatr Infect Dis J.* 2011;30(6):542.
160. Goldenberg SD, Bacelar M, Brazier P, Bisnauthsing K, Edgeworth JD. A cost benefit analysis of the Luminex xTAG Gastrointestinal Pathogen Panel for detection of infectious gastroenteritis in hospitalised patients. *J Infect.* 2015;70(5):504-511.
161. Abbott S, Janda J. Revisiting Bacterial Gastroenteritis: Issues, Possible Approaches, and an Ever-Expanding List of Etiologic Agents, Part II. *Clinical Microbiology Newsletter.* 2011;33:79-86.
162. Tam CC, O'Brien SJ, Tompkins DS, et al. Changes in causes of acute gastroenteritis in the United Kingdom over 15 years: microbiologic findings from 2 prospective, population-based studies of infectious intestinal disease. *Clin Infect Dis.* 2012;54(9):1275-1286.
163. Barman M, Unold D, Shifley K, et al. Enteric salmonellosis disrupts the microbial ecology of the murine gastrointestinal tract. *Infect Immun.* 2008;76(3):907-915.
164. David LA, Weil A, Ryan ET, et al. Gut microbial succession follows acute secretory diarrhea in humans. *mBio.* 2015;6(3):e00381-00315.
165. Gigliucci F, von Meijenfeldt FAB, Knijn A, et al. Metagenomic Characterization of the Human Intestinal Microbiota in Fecal Samples from STEC-Infected Patients. *Front Cell Infect Microbiol.* 2018;8:25.
166. Barash NR, Maloney JG, Singer SM, Dawson SC. Giardia Alters Commensal Microbial Diversity throughout the Murine Gut. *Infect Immun.* 2017;85(6).

167. Chen SY, Tsai CN, Lee YS, et al. Intestinal microbiome in children with severe and complicated acute viral gastroenteritis. *Sci Rep.* 2017;7:46130.
168. Dinleyici EC, Martínez-Martínez D, Kara A, et al. Time Series Analysis of the Microbiota of Children Suffering From Acute Infectious Diarrhea and Their Recovery After Treatment. *Front Microbiol.* 2018;9:1230.
169. Huang AD, Luo C, Pena-Gonzalez A, Weigand MR, Tarr CL, Konstantinidis KT. Metagenomics of Two Severe Foodborne Outbreaks Provides Diagnostic Signatures and Signs of Coinfection Not Attainable by Traditional Methods. *Appl Environ Microbiol.* 2017;83(3).
170. Kellingray L, Tapp HS, Saha S, Doleman JF, Narbad A, Mithen RF. Consumption of a diet rich in Brassica vegetables is associated with a reduced abundance of sulphate-reducing bacteria: A randomised crossover study. *Mol Nutr Food Res.* 2017;61(9).
171. Caporaso JG, Lauber CL, Walters WA, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A.* 2011;108 Suppl 1:4516-4522.
172. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods.* 2010;7(5):335-336.
173. Bolyen E, Rideout JR, Dillon MR, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol.* 2019;37(8):852-857.
174. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods.* 2016;13(7):581-583.
175. Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 2013;41(Database issue):D590-596.
176. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013;30(4):772-780.

177. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol.* 2009;26(7):1641-1650.
178. Faith DP. Conservation evaluation and phylogenetic diversity. *Biological Conservation.* 1992;61(1):1-10.
179. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. *ISME J.* 2011;5(2):169-172.
180. Douglas GM, Maffei VJ, Zaneveld J, et al. PICRUSt2: An improved and extensible approach for metagenome inference. *bioRxiv.* 2019:672295.
181. Langille MG, Zaneveld J, Caporaso JG, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol.* 2013;31(9):814-821.
182. Franzosa EA, McIver LJ, Rahnavard G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods.* 2018;15(11):962-968.
183. Abubucker S, Segata N, Goll J, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol.* 2012;8(6):e1002358.
184. Ogle D, Wheeler P, Dinno A. FSA: Fisheries Stock Analysis. In. R package version 0.8.302020.
185. Anderson MJ. A new method for non-parametric multivariate analysis of variance. *Austral Ecology.* 2001;26(1):32-46.
186. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol.* 2017;13(11):e1005752.
187. Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. *Nat Methods.* 2013;10(12):1200-1202.
188. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43(7):e47.

189. van der Boogaart K, Tolosana-Delgado R. *Compositional data analysis with 'R' and the package 'compositions'.* Vol 264. London: Geological Society; 2016.
190. Liaw A, Wiener M. Classification and Regression by randomForest. *R News.* 2002;2(3):18-22.
191. Singh P, Teal TK, Marsh TL, et al. Intestinal microbial communities associated with acute enteric infections and disease recovery. *Microbiome.* 2015;3:45.
192. Kellingray L, Gall GL, Defernez M, Beales ILP, Franslem-Elumogo N, Narbad A. Microbial taxonomic and metabolic alterations during faecal microbiota transplantation to treat Clostridium difficile infection. *J Infect.* 2018;77(2):107-118.
193. Monira S, Nakamura S, Gotoh K, et al. Metagenomic profile of gut microbiota in children during cholera and recovery. *Gut Pathog.* 2013;5(1):1.
194. Lopetuso LR, Scaldaferri F, Petito V, Gasbarrini A. Commensal Clostridia: leading players in the maintenance of gut homeostasis. *Gut Pathog.* 2013;5(1):23.
195. Duncan SH, Aminov RI, Scott KP, Louis P, Stanton TB, Flint HJ. Proposal of Roseburia faecis sp. nov., Roseburia hominis sp. nov. and Roseburia inulinivorans sp. nov., based on isolates from human faeces. *Int J Syst Evol Microbiol.* 2006;56(Pt 10):2437-2441.
196. Hugon P, Lagier JC, Colson P, Bittar F, Raoult D. Repertoire of human gut microbes. *Microb Pathog.* 2017;106:103-112.
197. Litvak Y, Byndloss MX, Tsolis RM, Bäumler AJ. Dysbiotic Proteobacteria expansion: a microbial signature of epithelial dysfunction. *Curr Opin Microbiol.* 2017;39:1-6.
198. Derrien M, Collado MC, Ben-Amor K, Salminen S, de Vos WM. The Mucin degrader Akkermansia muciniphila is an abundant resident of the human intestinal tract. *Appl Environ Microbiol.* 2008;74(5):1646-1648.
199. van Passel MW, Kant R, Zoetendal EG, et al. The genome of Akkermansia muciniphila, a dedicated intestinal mucin degrader, and its use in exploring intestinal metagenomes. *PLoS One.* 2011;6(3):e16876.
200. García B, Merayo-Lloves J, Rodríguez D, et al. Different Use of Cell Surface Glycosaminoglycans As Adherence Receptors to Corneal Cells by Gram

Positive and Gram Negative Pathogens. *Front Cell Infect Microbiol.* 2016;6:173.

201. Camesano TA, Logan BE. Probing Bacterial Electrosteric Interactions Using Atomic Force Microscopy. *Environmental Science & Technology.* 2000;34(16):3354-3362.
202. Amaretti A, Gozzoli C, Simone M, et al. Profiling of Protein Degraders in Cultures of Human Gut Microbiota. *Frontiers in microbiology.* 2019;10:2614-2614.
203. Kim Y-G, Sakamoto K, Seo S-U, et al. Neonatal acquisition of Clostridia species protects against colonization by bacterial pathogens. *Science (New York, NY).* 2017;356(6335):315-319.
204. Macfarlane GT, Allison C, Gibson SAW, Cummings JH. Contribution of the microflora to proteolysis in the human large intestine. *Journal of Applied Bacteriology.* 1988;64(1):37-46.
205. Schoch CL, Seifert KA, Huhndorf S, et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc Natl Acad Sci U S A.* 2012;109(16):6241-6246.
206. Feau N, Decourcelle T, Husson C, Desprez-Loustau ML, Dutech C. Finding single copy genes out of sequenced genomes for multilocus phylogenetics in non-model fungi. *PLoS One.* 2011;6(4):e18803.
207. Johnson JS, Spakowicz DJ, Hong BY, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun.* 2019;10(1):5029.
208. Kieser S, Sarker SA, Sakwinska O, et al. Bangladeshi children with acute diarrhoea show faecal microbiomes with increased Streptococcus abundance, irrespective of diarrhoea aetiology. *Environ Microbiol.* 2018;20(6):2256-2269.
209. Flint HJ, Bayer EA, Rincon MT, Lamed R, White BA. Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. *Nat Rev Microbiol.* 2008;6(2):121-131.
210. Furusawa Y, Obata Y, Fukuda S, et al. Commensal microbe-derived butyrate induces the differentiation of colonic regulatory T cells. *Nature.* 2013;504(7480):446-450.

211. Ng Hublin JSY, Ryan U, Trengove R, Maker G. Metabolomic Profiling of Faecal Extracts from Cryptosporidium parvum Infection in Experimental Mouse Models. *PLOS ONE.* 2013;8(10):e77803.
212. Doyle P, Kanaani J, Wang C. Hypoxanthine, guanine, xanthine phosphoribosyltransferase activity in Cryptosporidium parvum. *Exp Parasitol.* 1998;89(1):9-15.
213. Ott SJ, Musfeldt M, Timmis KN, Hampe J, Wenderoth DF, Schreiber S. In vitro alterations of intestinal bacterial microbiota in fecal samples during storage. *Diagn Microbiol Infect Dis.* 2004;50(4):237-245.
214. Fouhy F, Deane J, Rea MC, et al. The effects of freezing on faecal microbiota as determined using MiSeq sequencing and culture-based investigations. *PLoS One.* 2015;10(3):e0119355.
215. Knudsen BE, Bergmark L, Munk P, et al. Impact of Sample Type and DNA Isolation Procedure on Genomic Inference of Microbiome Composition. *mSystems.* 2016;1(5).
216. Velásquez-Mejía EP, de la Cuesta-Zuluaga J, Escobar JS. Impact of DNA extraction, sample dilution, and reagent contamination on 16S rRNA gene sequencing of human feces. *Appl Microbiol Biotechnol.* 2018;102(1):403-411.
217. Fouhy F, Clooney AG, Stanton C, Claesson MJ, Cotter PD. 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. *BMC Microbiol.* 2016;16(1):123.
218. Rintala A, Pietilä S, Munukka E, et al. Gut Microbiota Analysis Results Are Highly Dependent on the 16S rRNA Gene Target Region, Whereas the Impact of DNA Extraction Is Minor. *J Biomol Tech.* 2017;28(1):19-30.
219. Carroll IM, Ringel-Kulka T, Keku TO, et al. Molecular analysis of the luminal- and mucosal-associated intestinal microbiota in diarrhea-predominant irritable bowel syndrome. *Am J Physiol Gastrointest Liver Physiol.* 2011;301(5):G799-807.
220. David LA, Maurice CF, Carmody RN, et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature.* 2014;505(7484):559-563.

221. Haro C, Rangel-Zúñiga OA, Alcalá-Díaz JF, et al. Intestinal Microbiota Is Influenced by Gender and Body Mass Index. *PLoS One.* 2016;11(5):e0154090.
222. Foodborne Diseases Burden Epidemiology Reference Group 2007-2015. WHO Estimates of the Global Burden of Foodborne Disease. 2015.
223. Threlfall J, Ward L, Old D. Changing the nomenclature of Salmonella. *Commun Dis Public Health.* 1999;2(3):156-157.
224. Grimont P, Weill F. *Antigenic formulae of the Salmonella serovars.* 9th Edition ed. Paris: WHO Collaborating Centre for Reference and Research on Salmonella; 2007.
225. Wain J, Hendriksen RS, Mikoleit ML, Keddy KH, Ochiai RL. Typhoid fever. *Lancet.* 2015;385(9973):1136-1145.
226. Langridge GC, Fookes M, Connor TR, et al. Patterns of genome evolution that have accompanied host adaptation in Salmonella. *Proc Natl Acad Sci U S A.* 2015;112(3):863-868.
227. Roumagnac P, Weill FX, Dolecek C, et al. Evolutionary history of Salmonella typhi. *Science.* 2006;314(5803):1301-1304.
228. Reddy EA, Shaw AV, Crump JA. Community-acquired bloodstream infections in Africa: a systematic review and meta-analysis. *Lancet Infect Dis.* 2010;10(6):417-432.
229. Kingsley RA, Msefula CL, Thomson NR, et al. Epidemic multiple drug resistant Salmonella Typhimurium causing invasive disease in sub-Saharan Africa have a distinct genotype. *Genome Res.* 2009;19(12):2279-2287.
230. Murray CJ, Rosenfeld LC, Lim SS, et al. Global malaria mortality between 1980 and 2010: a systematic analysis. *Lancet.* 2012;379(9814):413-431.
231. Ao TT, Feasey NA, Gordon MA, Keddy KH, Angulo FJ, Crump JA. Global burden of invasive nontyphoidal Salmonella disease, 2010(1). *Emerg Infect Dis.* 2015;21(6).
232. Buckle GC, Walker CL, Black RE. Typhoid fever and paratyphoid fever: Systematic review to estimate global morbidity and mortality for 2010. *J Glob Health.* 2012;2(1):010401.

233. Public Health England. Travel-associated non-typhoidal Salmonella infection in England, Wales and Northern Ireland: 2014. 2017.
234. Public Health England. Enteric fever (typhoid and paratyphoid) England, Wales and Northern Ireland: 2017. 2018.
235. Wilson MR, Suan D, Duggins A, et al. A novel cause of chronic viral meningoencephalitis: Cache Valley virus. *Ann Neurol.* 2017;82(1):105-114.
236. Wilson MR, Zimmermann LL, Crawford ED, et al. Acute West Nile Virus Meningoencephalitis Diagnosed Via Metagenomic Deep Sequencing of Cerebrospinal Fluid in a Renal Transplant Patient. *Am J Transplant.* 2017;17(3):803-808.
237. Chiu CY, Coffey LL, Murkey J, et al. Diagnosis of Fatal Human Case of St. Louis Encephalitis Virus Infection by Metagenomic Sequencing, California, 2016. *Emerg Infect Dis.* 2017;23(10):1964-1968.
238. Murkey JA, Chew KW, Carlson M, et al. Hepatitis E Virus-Associated Meningoencephalitis in a Lung Transplant Recipient Diagnosed by Clinical Metagenomic Sequencing. *Open Forum Infect Dis.* 2017;4(3):ofx121.
239. Wilson MR, O'Donovan BD, Gelfand JM, et al. Chronic Meningitis Investigated via Metagenomic Next-Generation Sequencing. *JAMA Neurol.* 2018;75(8):947-955.
240. Wilson MR, Naccache SN, Samayoa E, et al. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *N Engl J Med.* 2014;370(25):2408-2417.
241. Wilson MR, Shanbhag NM, Reid MJ, et al. Diagnosing Balamuthia mandrillaris Encephalitis With Metagenomic Deep Sequencing. *Ann Neurol.* 2015;78(5):722-730.
242. Miller S, Naccache SN, Samayoa E, et al. Laboratory validation of a clinical metagenomic sequencing assay for pathogen detection in cerebrospinal fluid. *Genome Res.* 2019;29(5):831-842.
243. Wilson MR, Sample HA, Zorn KC, et al. Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis. *N Engl J Med.* 2019;380(24):2327-2340.

244. Graf EH, Simmon KE, Tardif KD, et al. Unbiased Detection of Respiratory Viruses by Use of RNA Sequencing-Based Metagenomics: a Systematic Comparison to a Commercial PCR Panel. *J Clin Microbiol.* 2016;54(4):1000-1007.
245. Doan T, Wilson MR, Crawford ED, et al. Illuminating uveitis: metagenomic deep sequencing identifies common and rare pathogens. *Genome Med.* 2016;8(1):90.
246. Doan T, Acharya NR, Pinsky BA, et al. Metagenomic DNA Sequencing for the Diagnosis of Intraocular Infections. *Ophthalmology.* 2017;124(8):1247-1248.
247. Ferrer R, Martin-Loeches I, Phillips G, et al. Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program. *Crit Care Med.* 2014;42(8):1749-1755.
248. Grumaz S, Stevens P, Grumaz C, et al. Next-generation sequencing diagnostics of bacteremia in septic patients. *Genome Med.* 2016;8(1):73.
249. Abril MK, Barnett AS, Wegermann K, et al. Diagnosis of Capnocytophaga canimorsus sepsis by whole genome next-generation sequencing. *Open Forum Infect Dis.* 2016;3(3):ofw144.
250. Hilty M, Burke C, Pedro H, et al. Disordered microbial communities in asthmatic airways. *PLoS One.* 2010;5(1):e8578.
251. Zhou Y, Wylie KM, El Feghaly RE, et al. Metagenomic Approach for Identification of the Pathogens Associated with Diarrhea in Stool Specimens. *J Clin Microbiol.* 2016;54(2):368-375.
252. Andersen SC, Kiil K, Harder CB, et al. Towards diagnostic metagenomics of Campylobacter in fecal samples. *BMC Microbiol.* 2017;17(1):133.
253. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014;30(15):2114-2120.
254. Ashton PM, Nair S, Peters TM, et al. Identification of Salmonella for public health surveillance using whole genome sequencing. *PeerJ.* 2016;4:e1752.
255. Achtman M, Wain J, Weill FX, et al. Multilocus sequence typing as a replacement for serotyping in Salmonella enterica. *PLoS Pathog.* 2012;8(6):e1002776.

256. Tewolde R, Dallman T, Schaefer U, et al. MOST: a modified MLST typing tool based on short read sequencing. *PeerJ.* 2016;4:e2308.
257. Bushnell B. BBMap: A fast, accurate splice-aware aligner. In:2014.
258. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods.* 2012;9(8):811-814.
259. Truong DT, Franzosa EA, Tickle TL, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods.* 2015;12(10):902-903.
260. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 2014;15(3):R46.
261. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25(16):2078-2079.
262. Kang DD, Li F, Kirton E, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ.* 2019;7:e7359.
263. Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front Genet.* 2013;4:237.
264. Seah BK, Gruber-Vodicka HR. gbtools: Interactive Visualization of Metagenome Bins in R. *Front Microbiol.* 2015;6:1451.
265. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* 2017;27(5):824-834.
266. Tamburini FB, Andermann TM, Tkachenko E, Senchyna F, Banaei N, Bhatt AS. Precision identification of diverse bloodstream pathogens in the gut microbiome. *Nat Med.* 2018;24(12):1809-1814.
267. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 2019;47(W1):W256-W259.
268. Wattam AR, Davis JJ, Assaf R, et al. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res.* 2017;45(D1):D535-D542.

269. Brettin T, Davis JJ, Disz T, et al. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci Rep.* 2015;5:8365.
270. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* 2014;30(9):1312-1313.
271. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012;9(4):357-359.
272. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics.* 2011;27(21):2987-2993.
273. Day MR, Doumith M, do Nascimento V, et al. Comparison of phenotypic and whole genome sequencing-derived antimicrobial resistance profiles of Salmonella enterica serovars Typhi and Paratyphi. *The Journal of Antimicrobial Chemotherapy.* 2017a.
274. Sadouki Z, Day MR, Doumith M, et al. Comparison of phenotypic and WGS-derived antimicrobial resistance profiles of Shigella sonnei isolated from cases of diarrhoeal disease in England and Wales, 2015. *The Journal of antimicrobial chemotherapy.* 2017;72(9):2496–2502.
275. Rowe WPM, Winn MD. Indexed variation graphs for efficient and accurate resistome profiling. *Bioinformatics.* 2018;34(21):3601-3608.
276. Magnet S, Courvalin P, Lambert T. Activation of the cryptic aac(6')-Iy aminoglycoside resistance gene of Salmonella by a chromosomal deletion generating a transcriptional fusion. *J Bacteriol.* 1999;181(21):6650-6655.
277. Stuempfig ND, Seroy J. Viral Gastroenteritis. *StatPearls*. 2020.
278. Vandeputte D, Falony G, Vieira-Silva S, Tito RY, Joossens M, Raes J. Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut.* 2016;65(1):57-62.
279. Rodriguez-R LM, Konstantinidis KT. Estimating coverage in metagenomic data sets and why it matters. *ISME J.* 2014;8(11):2349-2351.
280. Albertsen M, Karst SM, Ziegler AS, Kirkegaard RH, Nielsen PH. Back to Basics--The Influence of DNA Extraction and Primer Choice on Phylogenetic Analysis of Activated Sludge Communities. *PLoS One.* 2015;10(7):e0132783.

281. Stephen AM, Cummings JH. The microbial contribution to human faecal mass. *J Med Microbiol.* 1980;13(1):45-56.
282. Briese T, Kapoor A, Mishra N, et al. Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis. *mBio.* 2015;6(5):e01491-01415.
283. Wylie TN, Wylie KM, Herter BN, Storch GA. Enhanced virome sequencing using targeted sequence capture. *Genome Res.* 2015;25(12):1910-1920.
284. Singh N, Lapierre P, Quinlan TM, et al. Whole-Genome Single-Nucleotide Polymorphism (SNP) Analysis Applied Directly to Stool for Genotyping Shiga Toxin-Producing Escherichia coli: an Advanced Molecular Detection Method for Foodborne Disease Surveillance and Outbreak Tracking. *J Clin Microbiol.* 2019;57(7).
285. Neuert S, Nair S, Day MR, et al. Prediction of Phenotypic Antimicrobial Resistance Profiles From Whole Genome Sequences of Non-typhoidal. *Front Microbiol.* 2018;9:592.
286. Mirza SH, Beeching NJ, Hart CA. Multi-drug resistant typhoid: a global problem. *J Med Microbiol.* 1996;44(5):317-319.
287. Wain J, Hoa NT, Chinh NT, et al. Quinolone-resistant Salmonella typhi in Viet Nam: molecular basis of resistance and clinical response to treatment. *Clin Infect Dis.* 1997;25(6):1404-1410.
288. Threlfall EJ, Ward LR, Skinner JA, Smith HR, Lacey S. Ciprofloxacin-resistant Salmonella typhi and treatment failure. *Lancet.* 1999;353(9164):1590-1591.
289. Hassing RJ, Goessens WH, van Pelt W, et al. Salmonella subtypes with increased MICs for azithromycin in travelers returned to The Netherlands. *Emerg Infect Dis.* 2014;20(4):705-708.
290. Threlfall EJ, Frost JA, Ward LR, Rowe B. Increasing spectrum of resistance in multiresistant Salmonella typhimurium. *Lancet.* 1996;347(9007):1053-1054.
291. Boyd D, Cloeckaert A, Chaslus-Dancla E, Mulvey MR. Characterization of variant Salmonella genomic island 1 multidrug resistance regions from serovars Typhimurium DT104 and Agona. *Antimicrob Agents Chemother.* 2002;46(6):1714-1722.

292. Le Hello S, Bekhit A, Granier SA, et al. The global establishment of a highly-fluoroquinolone resistant Salmonella enterica serotype Kentucky ST198 strain. *Front Microbiol.* 2013;4:395.
293. Nair S, Ashton P, Doumith M, et al. WGS for surveillance of antimicrobial resistance: a pilot study to detect the prevalence and mechanism of resistance to azithromycin in a UK population of non-typhoidal Salmonella. *J Antimicrob Chemother.* 2016;71(12):3400-3408.
294. Villa L, Guerra B, Schmoger S, et al. IncA/C Plasmid Carrying bla(NDM-1), bla(CMY-16), and fosA3 in a Salmonella enterica Serovar Corvallis Strain Isolated from a Migratory Wild Bird in Germany. *Antimicrob Agents Chemother.* 2015;59(10):6597-6600.
295. European Food Safety Authority, European Centre for Disease Prevention and Control. The European Union summary report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2015. *EFSA Journal.* 2017;15(2):e04694.
296. Olaitan AO, Morand S, Rolain J-M. Mechanisms of polymyxin resistance: acquired and intrinsic resistance in bacteria. *Frontiers in microbiology.* 2014;5:643-643.
297. Liu Y-Y, Wang Y, Walsh T, et al. Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: A microbiological and molecular biological study. *The Lancet Infectious Diseases.* 2015;16.
298. Doumith M, Godbole G, Ashton P, et al. Detection of the plasmid-mediated mcr-1 gene conferring colistin resistance in human and food isolates of Salmonella enterica and Escherichia coli in England and Wales. *J Antimicrob Chemother.* 2016;71(8):2300-2305.
299. Varma JK, Molbak K, Barrett TJ, et al. Antimicrobial-resistant nontyphoidal Salmonella is associated with excess bloodstream infections and hospitalizations. *J Infect Dis.* 2005;191(4):554-561.
300. Martin LJ, Fyfe M, Doré K, et al. Increased burden of illness associated with antimicrobial-resistant Salmonella enterica serotype typhimurium infections. *J Infect Dis.* 2004;189(3):377-384.

301. Hopkins KL, Peters TM, Lawson AJ, Owen RJ. Rapid identification of Salmonella enterica subsp. arizonae and S. enterica subsp. diarizonae by real-time polymerase chain reaction. *Diagn Microbiol Infect Dis.* 2009;64(4):452-454.
302. Hopkins KL, Peters TM, de Pinna E, Wain J. Standardisation of multilocus variable-number tandem-repeat analysis (MLVA) for subtyping of Salmonella enterica serovar Enteritidis. *Euro Surveill.* 2011;16(32).
303. Issenhuth-Jeanjean S, Roggentin P, Mikoleit M, et al. Supplement 2008-2010 (no. 48) to the White-Kauffmann-Le Minor scheme. *Research in Microbiology.* 2014;165.
304. Public Health England. Salmonella data 2007 to 2016 - National laboratory data for residents of England and Wales. 2018.
305. Prager R, Rabsch W, Streckel W, Voigt W, Tietze E, Tschäpe H. Molecular properties of Salmonella enterica serotype paratyphi B distinguish between its systemic and its enteric pathovars. *J Clin Microbiol.* 2003;41(9):4270-4278.
306. Maiden MCJ. Multilocus Sequence Typing of Bacteria. *Annual Review of Microbiology.* 2006;60(1):561-588.
307. Köser CU, Ellington MJ, Cartwright EJ, et al. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathog.* 2012;8(8):e1002824.
308. Dallman TJ, Byrne L, Launders N, Glen K, Grant KA, Jenkins C. The utility and public health implications of PCR and whole genome sequencing for the detection and investigation of an outbreak of Shiga toxin-producing Escherichia coli serogroup O26:H11. *Epidemiol Infect.* 2015;143(8):1672-1680.
309. Inns T, Ashton PM, Herrera-Leon S, et al. Prospective use of whole genome sequencing (WGS) detected a multi-country outbreak of Salmonella Enteritidis. *Epidemiol Infect.* 2017;145(2):289-298.
310. Mair-Jenkins J, Borges-Stewart R, Harbour C, et al. Investigation using whole genome sequencing of a prolonged restaurant outbreak of. *Euro Surveill.* 2017;22(49).

311. Kanagarajah S, Waldram A, Dolan G, et al. Whole genome sequencing reveals an outbreak of Salmonella Enteritidis associated with reptile feeder mice in the United Kingdom, 2012-2015. *Food Microbiology.* 2017;71.
312. Pijnacker R, Dallman TJ, Tijsma ASL, et al. An international outbreak of Salmonella enterica serotype Enteritidis linked to eggs from Poland: a microbiological and epidemiological study. *Lancet Infect Dis.* 2019;19(7):778-786.
313. Li H. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinformatics (Oxford, England).* 2009;25:1754-1760.
314. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011;43(5):491-498.
315. Dallman T, Ashton P, Schafer U, et al. SnapperDB: A database solution for routine sequencing analysis of bacterial isolates. *Bioinformatics (Oxford, England).* 2018;34.
316. Ruiz J, Castro D, Goñi P, Santamaria JA, Borrego JJ, Vila J. Analysis of the mechanism of quinolone resistance in nalidixic acid-resistant clinical isolates of Salmonella serotype Typhimurium. *J Med Microbiol.* 1997;46(7):623-628.
317. Yoshida H, Bogaki M, Nakamura M, Nakamura S. Quinolone resistance-determining region in the DNA gyrase gyrA gene of Escherichia coli. *Antimicrob Agents Chemother.* 1990;34(6):1271-1272.
318. Ashton PM, Owen SV, Kaindama L, et al. Public health surveillance in the UK revolutionises our understanding of the invasive Salmonella Typhimurium epidemic in Africa. *Genome Med.* 2017;9(1):92.
319. European Food Safety Authority, European Centre for Disease Prevention and Control. The European Union summary report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2014. *EFSA Journal.* 2016;14(2):4380.
320. Doublet B, Boyd D, Mulvey MR, Cloeckaert A. The Salmonella genomic island 1 is an integrative mobilizable element. *Mol Microbiol.* 2005;55(6):1911-1924.
321. Sood S, Kapil A, Das B, Jain Y, Kabra SK. Re-emergence of chloramphenicol-sensitive Salmonella typhi. *Lancet.* 1999;353(9160):1241-1242.

322. Rahman M, Ahmad A, Shoma S. Decline in epidemic of multidrug resistant Salmonella typhi is not associated with increased incidence of antibiotic-susceptible strain in Bangladesh. *Epidemiol Infect.* 2002;129(1):29-34.

323. Centers for Disease Control and Prevention. National Antimicrobial Resistance Monitoring System for Enteric Bacteria (NARMS): Human Isolates Surveillance Report for 2015
(Final Report). Atlanta, Georgia: U.S. Department of Health and Human Services; 2018.

324. Threlfall EJ, Ward LR, Skinner JA, Graham A. Antimicrobial drug resistance in non-typhoidal salmonellas from humans in England and Wales in 1999: decrease in multiple resistance in Salmonella enterica serotypes Typhimurium, Virchow, and Hadar. *Microb Drug Resist.* 2000;6(4):319-325.

325. García-Fernández A, Fortini D, Veldman K, Mevius D, Carattoli A. Characterization of plasmids harbouring qnrS1, qnrB2 and qnrB19 genes in Salmonella. *J Antimicrob Chemother.* 2009;63(2):274-281.

326. Su LH, Wu TL, Chia JH, Chu C, Kuo AJ, Chiu CH. Increasing ceftriaxone resistance in Salmonella isolates from a university hospital in Taiwan. *J Antimicrob Chemother.* 2005;55(6):846-852.

327. Mataseje LF, Xiao J, Kost S, et al. Characterization of Canadian cefoxitin-resistant non-typhoidal Salmonella isolates, 2005-06. *J Antimicrob Chemother.* 2009;64(4):723-730.

328. Batchelor M, Hopkins KL, Threlfall EJ, et al. Characterization of AmpC-mediated resistance in clinical Salmonella isolates recovered from humans during the period 1992 to 2003 in England and Wales. *J Clin Microbiol.* 2005;43(5):2261-2265.

329. Burke L, Hopkins KL, Meunier D, et al. Resistance to third-generation cephalosporins in human non-typhoidal Salmonella enterica isolates from England and Wales, 2010-12. *J Antimicrob Chemother.* 2014;69(4):977-981.

330. European Food Safety Authority, European Centre for Disease Prevention and Control. The European Union Summary Report on Antimicrobial Resistance in zoonotic and indicator bacteria from humans, animals and food in 2017/2018. *EFSA Journal.* 2020;18(3):e06007.

331. Lee HY, Su LH, Tsai MH, et al. High rate of reduced susceptibility to ciprofloxacin and ceftriaxone among nontyphoid Salmonella clinical isolates in Asia. *Antimicrob Agents Chemother.* 2009;53(6):2696-2699.

332. Hopkins K, Wootton L, Day M, Threlfall E. Plasmid mediated quinolone resistance determinant qnrS1 found in Salmonella enteric strains isolated in the UK. *The Journal of antimicrobial chemotherapy.* 2007;59:1071-1075.

333. Benacer D, Thong KL, Watanabe H, Puthucheary SD. Characterization of drug resistant Salmonella enterica serotype Typhimurium by antibiograms, plasmids, integrons, resistance genes and PFGE. *J Microbiol Biotechnol.* 2010;20(6):1042-1052.

334. Vo AT, van Duijkeren E, Gaastra W, Fluit AC. Antimicrobial resistance, class 1 integrons, and genomic island 1 in Salmonella isolates from Vietnam. *PLoS One.* 2010;5(2):e9440.

335. Stoesser N, Batty EM, Eyre DW, et al. Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data. *J Antimicrob Chemother.* 2013;68(10):2234-2244.

336. Tyson GH, McDermott PF, Li C, et al. WGS accurately predicts antimicrobial resistance in Escherichia coli. *J Antimicrob Chemother.* 2015;70(10):2763-2769.

337. Day M, Doumith M, Jenkins C, et al. Antimicrobial resistance in Shiga toxin-producing Escherichia coli serogroups O157 and O26 isolated from human cases of diarrhoeal disease in England, 2015. *J Antimicrob Chemother.* 2017;72(1):145-152.

338. Zankari E, Hasman H, Kaas RS, et al. Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. *J Antimicrob Chemother.* 2013;68(4):771-777.

339. McDermott PF, Tyson GH, Kabera C, et al. Whole-Genome Sequencing for Detecting Antimicrobial Resistance in Nontyphoidal Salmonella. *Antimicrob Agents Chemother.* 2016;60(9):5515-5520.

340. Garcia-Migura L, Sunde M, Karlsmose S, et al. Establishing streptomycin epidemiological cut-off values for Salmonella and Escherichia coli. *Microb Drug Resist.* 2012;18(1):88-93.
341. Tyson GH, Zhao S, Li C, et al. Establishing Genotypic Cutoff Values To Measure Antimicrobial Resistance in Salmonella. *Antimicrob Agents Chemother.* 2017;61(3).
342. Heider LC, Hoet AE, Wittum TE, et al. Genetic and phenotypic characterization of the bla(CMY) gene from Escherichia coli and Salmonella enterica isolated from food-producing animals, humans, the environment, and retail meat. *Foodborne Pathog Dis.* 2009;6(10):1235-1240.
343. Adesiji YO, Deekshit VK, Karunasagar I. Antimicrobial-resistant genes associated with Salmonella spp. isolated from human, poultry, and seafood sources. *Food Sci Nutr.* 2014;2(4):436-442.
344. Hopkins KL, Davies RH, Threlfall EJ. Mechanisms of quinolone resistance in Escherichia coli and Salmonella: recent developments. *Int J Antimicrob Agents.* 2005;25(5):358-373.
345. Boulund F, Berglund F, Flach CF, et al. Computational discovery and functional validation of novel fluoroquinolone resistance genes in public metagenomic data sets. *BMC Genomics.* 2017;18(1):682.
346. Catry B, Cavaleri M, Baptiste K, et al. Use of colistin-containing products within the European Union and European Economic Area (EU/EEA): development of resistance in animals and possible impact on human and animal health. *Int J Antimicrob Agents.* 2015;46(3):297-306.
347. Carroll L, Gaballa A, Guldimann C, Sullivan G, Henderson L, Wiedmann M. Identification of Novel Mobilized Colistin Resistance Gene mcr-9 in a Multidrug-Resistant, Colistin-Susceptible Salmonella enterica Serotype Typhimurium Isolate. *mBio.* 2019;10.
348. Food and Drug Administration. Antimicrobial Suscepibility Test (AST) Systems - Class II Special Controls Guidance for Industry and FDA. Vol 42. Rockville, MD: US Food and Drug Administration; 2009.

349. Ellington MJ, Ekelund O, Aarestrup FM, et al. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. *Clin Microbiol Infect.* 2017;23(1):2-22.
350. Nordahl Petersen T, Rasmussen S, Hasman H, et al. Meta-genomic analysis of toilet waste from long distance flights; a step towards global surveillance of infectious diseases and antimicrobial resistance. *Scientific Reports.* 2015;5(1):11444.
351. World Trade Organisation. UNWTO Annual Report 2015. Madrid: UNWTO; 2016.
352. Angelo KM, Kozarsky PE, Ryan ET, Chen LH, Sotir MJ. What proportion of international travellers acquire a travel-related illness? A review of the literature. *J Travel Med.* 2017;24(5).
353. Youmans B, Ajami NJ, Jiang Z-D, et al. Characterization of the human gut microbiome during travelers' diarrhea. Gut microbes. 2015;6.
354. Riddle MS, Connor BA. The Traveling Microbiome. Current Infectious Disease Reports. 2016;18(9):29.
355. Harvey K, Esposito DH, Han P, et al. Surveillance for travel-related disease--GeoSentinel Surveillance System, United States, 1997-2011. MMWR Surveill Summ. 2013;62:1-23.
356. Angelo KM, Kozarsky PE, Ryan ET, Chen LH, Sotir MJ. What proportion of international travellers acquire a travel-related illness? A review of the literature. J Travel Med. 2017;24(5).
357. Shah N, DuPont HL, Ramsey DJ. Global etiology of travelers' diarrhea: systematic review from 1973 to the present. Am J Trop Med Hyg. 2009;80(4):609-614.
358. Porter CK, Olson S, Hall A, Riddle MS. Travelers' Diarrhea: An Update on the Incidence, Etiology, and Risk in Military Deployments and Similar Travel Populations. Mil Med. 2017;182(S2):4-10.
359. Van Boeckel TP, Gandra S, Ashok A, et al. Global antibiotic consumption 2000 to 2010: an analysis of national pharmaceutical sales data. Lancet Infect Dis. 2014;14(8):742-750.

360. Van Boeckel TP, Brower C, Gilbert M, et al. Global trends in antimicrobial use in food animals. Proc Natl Acad Sci U S A. 2015;112(18):5649-5654.
361. Woerther PL, Burdet C, Chachaty E, Andremont A. Trends in human fecal carriage of extended-spectrum β-lactamases in the community: toward the globalization of CTX-M. Clin Microbiol Rev. 2013;26(4):744-758.
362. MacVane SH, Tuttle LO, Nicolau DP. Impact of extended-spectrum β-lactamase-producing organisms on clinical and economic outcomes in patients with urinary tract infection. J Hosp Med. 2014;9(4):232-238.
363. Rodríguez-Baño J, Picón E, Gijón P, et al. Community-onset bacteremia due to extended-spectrum beta-lactamase-producing Escherichia coli: risk factors and prognosis. Clin Infect Dis. 2010;50(1):40-48.
364. Mesa RJ, Blanc V, Blanch AR, et al. Extended-spectrum beta-lactamase-producing Enterobacteriaceae in different environments (humans, food, animal farms and sewage). J Antimicrob Chemother. 2006;58(1):211-215.
365. Abdul Rahman EM, El-Sherif RH. High rates of intestinal colonization with extended-spectrum lactamase-producing Enterobacteriaceae among healthy individuals. J Investig Med. 2011;59(8):1284-1286.
366. Rossolini GM, D'Andrea MM, Mugnaioli C. The spread of CTX-M-type extended-spectrum beta-lactamases. Clin Microbiol Infect. 2008;14 Suppl 1:33-41.
367. McNulty CAM, Lecky DM, Xu-McCrae L, et al. CTX-M ESBL-producing Enterobacteriaceae: estimated prevalence in adults in England in 2014. Journal of Antimicrobial Chemotherapy. 2018;73(5):1368-1388.
368. Arcilla MS, van Hattem JM, Haverkate MR, et al. Import and spread of extended-spectrum β-lactamase-producing Enterobacteriaceae by international travellers (COMBAT study): a prospective, multicentre cohort study. Lancet Infect Dis. 2017;17(1):78-85.
369. Tängdén T, Cars O, Melhus A, Löwdin E. Foreign travel is a major risk factor for colonization with Escherichia coli producing CTX-M-type extended-spectrum beta-lactamases: a prospective study with Swedish volunteers. Antimicrob Agents Chemother. 2010;54(9):3564-3568.

370. Kennedy K, Collignon P. Colonisation with Escherichia coli resistant to "critically important" antibiotics: a high risk for international travellers. Eur J Clin Microbiol Infect Dis. 2010;29(12):1501-1506.
371. Ruppé E, Armand-Lefèvre L, Estellat C, et al. Acquisition of carbapenemase-producing Enterobacteriaceae by healthy travellers to India, France, February 2012 to March 2013. Euro Surveill. 2014;19(14).
372. Lübbert C, Straube L, Stein C, et al. Colonization with extended-spectrum beta-lactamase-producing and carbapenemase-producing Enterobacteriaceae in international travelers returning to Germany. Int J Med Microbiol. 2015;305(1):148-156.
373. Paltansing S, Vlot JA, Kraakman ME, et al. Extended-spectrum β-lactamase-producing enterobacteriaceae among travelers from the Netherlands. Emerg Infect Dis. 2013;19(8):1206-1213.
374. von Wintersdorff CJ, Penders J, Stobberingh EE, et al. High rates of antimicrobial drug resistance gene acquisition after international travel, The Netherlands. Emerg Infect Dis. 2014;20(4):649-657.
375. Ostholm-Balkhed A, Tärnberg M, Nilsson M, Nilsson LE, Hanberger H, Hällgren A. Travel-associated faecal colonization with ESBL-producing Enterobacteriaceae: incidence and risk factors. J Antimicrob Chemother. 2013;68(9):2144-2153.
376. Tham J, Odenholt I, Walder M, Brolund A, Ahl J, Melander E. Extended-spectrum beta-lactamase-producing Escherichia coli in patients with travellers' diarrhoea. Scand J Infect Dis. 2010;42(4):275-280.
377. Kantele A, Lääveri T, Mero S, et al. Antimicrobials Increase Travelers' Risk of Colonization by Extended-Spectrum Betalactamase-Producing Enterobacteriaceae. Clinical Infectious Diseases. 2015;60(6):837-846.
378. Bengtsson-Palme J, Angelin M, Huss M, et al. The Human Gut Microbiome as a Transporter of Antibiotic Resistance Genes between Continents. Antimicrob Agents Chemother. 2015;59(10):6551-6560.
379. Angelin M, Forsell J, Granlund M, Evengård B, Palmgren H, Johansson A. Risk factors for colonization with extended-spectrum beta-lactamase

producing Enterobacteriaceae in healthcare students on clinical assignment abroad: A prospective study. Travel Med Infect Dis. 2015;13(3):223-229.

380. Ruppé E, Armand-Lefèvre L, Estellat C, et al. High Rate of Acquisition but Short Duration of Carriage of Multidrug-Resistant Enterobacteriaceae After Travel to the Tropics. Clin Infect Dis. 2015;61(4):593-600.
381. Lakin SM, Dean C, Noyes NR, et al. MEGARes: an antimicrobial resistance database for high throughput sequencing. Nucleic acids research. 2017;45(D1):D574-D580.
382. Oksanen J, Blanchet FG, Kindt R, et al. Vegan: Community Ecology Package. R Package Version 22-1. 2015;2:1-2.
383. Leo S, Lazarevic V, Gaïa N, et al. The intestinal microbiota predisposes to traveler's diarrhea and to the carriage of multidrug-resistant Enterobacteriaceae after traveling to tropical regions. Gut Microbes. 2019;10(5):631-641.
384. World Health Organisation. WHO Report on surveillance of antibiotic consumption: 2016-2018 early implementation. Geneva: World Health Organization; 2018.
385. Forslund K, Sunagawa S, Kultima JR, et al. Country-specific antibiotic use practices impact the human gut resistome. Genome Res. 2013;23(7):1163-1169.
386. Raetz CR, Reynolds CM, Trent MS, Bishop RE. Lipid A modification systems in gram-negative bacteria. Annu Rev Biochem. 2007;76:295-329.
387. Zhang S, Oh JH, Alexander LM, Özçam M, van Pijkeren JP. d-Alanyl-d-Alanine Ligase as a Broad-Host-Range Counterselection Marker in Vancomycin-Resistant Lactic Acid Bacteria. J Bacteriol. 2018;200(13).
388. Gueimonde M, Sánchez B, G de Los Reyes-Gavilán C, Margolles A. Antibiotic resistance in probiotic bacteria. Frontiers in microbiology. 2013;4:202-202.
389. Tsubakishita S, Kuwahara-Arai K, Sasaki T, Hiramatsu K. Origin and molecular evolution of the determinant of methicillin resistance in staphylococci. Antimicrob Agents Chemother. 2010;54(10):4352-4359.

390. Becker K, van Alen S, Idelevich EA, et al. Plasmid-Encoded Transferable mecB-Mediated Methicillin Resistance in Staphylococcus aureus. Emerging infectious diseases. 2018;24(2):242-248.
391. Sullivan A, Edlund C, Nord CE. Effect of antimicrobial agents on the ecological balance of human microflora. Lancet Infect Dis. 2001;1(2):101-114.
392. Rodríguez-Baño J, Picón E, Gijón P, et al. Community-onset bacteremia due to extended-spectrum beta-lactamase-producing Escherichia coli: risk factors and prognosis. Clin Infect Dis. 2010;50(1):40-48.
393. Kuenzli E, Jaeger VK, Frei R, et al. High colonization rates of extended-spectrum β-lactamase (ESBL)-producing Escherichia coli in Swiss travellers to South Asia- a prospective observational multicentre cohort study looking at epidemiology, microbiology and risk factors. BMC Infect Dis. 2014;14:528.
394. Migault C, Kanagaratnam L, Nguyen Y, et al. Poor knowledge among French travellers of the risk of acquiring multidrug-resistant bacteria during travel. J Travel Med. 2016;24(1).
395. Dupont HL, Jiang ZD, Belkind-Gerson J, et al. Treatment of travelers' diarrhea: randomized trial comparing rifaximin, rifaximin plus loperamide, and loperamide alone. Clin Gastroenterol Hepatol. 2007;5(4):451-456.
396. Alajbegovic S, Sanders JW, Atherly DE, Riddle MS. Effectiveness of rifaximin and fluoroquinolones in preventing travelers' diarrhea (TD): a systematic review and meta-analysis. Systematic Reviews. 2012;1(1):39.
397. Vilkman K, Lääveri T, Pakkanen SH, Kantele A. Stand-by antibiotics encourage unwarranted use of antibiotics for travelers' diarrhea: A prospective study. Travel Med Infect Dis. 2019;27:64-71.
398. Post A, Martiny D, van Waterschoot N, et al. Antibiotic susceptibility profiles among Campylobacter isolates obtained from international travelers between 2007 and 2014. Eur J Clin Microbiol Infect Dis. 2017;36(11):2101-2107.
399. Mendez Arancibia E, Pitart C, Ruiz J, Marco F, Gascón J, Vila J. Evolution of antimicrobial resistance in enteroaggregative Escherichia coli and enterotoxigenic Escherichia coli causing traveller's diarrhoea. J Antimicrob Chemother. 2009;64(2):343-347.

400. O'Donnell AT, Vieira AR, Huang JY, Whichard J, Cole D, Karp BE. Quinolone-resistant Salmonella enterica serotype Enteritidis infections associated with international travel. Clin Infect Dis. 2014;59(9):e139-141.
401. Almeida A, Mitchell AL, Tarkowska A, Finn RD. Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. Gigascience. 2018;7(5).
402. Sczyrba A, Hofmann P, Belmann P, et al. Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. Nat Methods. 2017;14(11):1063-1071.
403. Zaneveld JR, McMinds R, Vega Thurber R. Stress and stability: applying the Anna Karenina principle to animal microbiomes. Nat Microbiol. 2017;2:17121.
404. Khare R, Espy MJ, Cebelinski E, et al. Comparative evaluation of two commercial multiplex panels for detection of gastrointestinal pathogens by use of clinical stool specimens. J Clin Microbiol. 2014;52(10):3667-3673.
405. Buss SN, Leber A, Chapin K, et al. Multicenter evaluation of the BioFire FilmArray gastrointestinal panel for etiologic diagnosis of infectious gastroenteritis. J Clin Microbiol. 2015;53(3):915-925.
406. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 1995;269(5223):496-512.
407. Owen RJ. Chromosomal DNA fingerprinting--a new method of species and strain identification applicable to microbial pathogens. J Med Microbiol. 1989;30(2):89-99.
408. Zadoks R, van Leeuwen W, Barkema H, et al. Application of pulsed-field gel electrophoresis and binary typing as tools in veterinary clinical microbiology and molecular epidemiologic analysis of bovine and human Staphylococcus aureus isolates. J Clin Microbiol. 2000;38(5):1931-1939.
409. Sonnenschein C. Bakteriendiagnose mit Bakteriophagen. Beschleunigung der bakteriologischen Typhusdiagnose. Dt Med Wochenschr. 1928:54:1034-1036.
410. Maiden MCJ. Multilocus Sequence Typing of Bacteria. Annual Review of Microbiology. 2006;60(1):561-588.

411. Chattaway MA, Dallman TJ, Gentle A, et al. Whole Genome Sequencing for Public Health Surveillance of Shiga Toxin-Producing Escherichia coli Other than Serogroup O157. Front Microbiol. 2016;7:258.
412. Chattaway MA, Greig DR, Gentle A, Hartman HB, Dallman TJ, Jenkins C. Wole-Genome Sequencing for National Surveillance of Shigella flexneri. Front Microbiol. 2017;8:1700.
413. Elson R, Awofisayo-Okuyelu A, Greener T, et al. Utility of Whole Genome Sequencing To Describe the Persistence and Evolution of Listeria monocytogenes Strains within Crabmeat Processing Environments Linked to Two Outbreaks of Listeriosis. J Food Prot. 2019;82(1):30-38.
414. Painset A, Day M, Doumith M, et al. Comparison of phenotypic and WGS-derived antimicrobial resistance profiles of Campylobacter jejuni and Campylobacter coli isolated from cases of diarrhoeal disease in England and Wales, 2015–16. J Antimicrob Chemother. 2020;75(4):883-889.
415. Hunter E, Greig DR, Schaefer U, et al. Identification and typing of Yersinia enterocolitica and Yersinia pseudotuberculosis isolated from human clinical specimens in England between 2004 and 2018. J Med Microbiol. 2019;68(4):538-548.
416. Brown E, Dessai U, McGarry S, Gerner-Smidt P. Use of Whole-Genome Sequencing for Food Safety and Public Health in the United States. Foodborne pathogens and disease. 2019;16(7):441-450.
417. Revez J, Espinosa L, Albiger B, Leitmeyer KC, Struelens MJ, Group ENMFPaE. Survey on the Use of Whole-Genome Sequencing for Infectious Diseases Surveillance: Rapid Expansion of European National Capacities, 2015-2016. Frontiers in public health. 2017;5:347-347.
418. Kohl TA, Diel R, Harmsen D, et al. Whole-genome-based Mycobacterium tuberculosis surveillance: a standardized, portable, and expandable approach. J Clin Microbiol. 2014;52(7):2479-2486.
419. Jolley KA, Maiden MC. Automated extraction of typing information for bacterial pathogens from whole genome sequence data: Neisseria meningitidis as an exemplar. Euro Surveill. 2013;18(4):20379.

420. Harris SR, Cole MJ, Spiteri G, et al. Public health surveillance of multidrug-resistant clones of Neisseria gonorrhoeae in Europe: a genomic survey. Lancet Infect Dis. 2018;18(7):758-768.
421. Jenkins C, Dallman TJ, Launders N, et al. Public Health Investigation of Two Outbreaks of Shiga Toxin-Producing Escherichia coli O157 Associated with Consumption of Watercress. Appl Environ Microbiol. 2015;81(12):3946-3952
422. Underwood AP, Dallman T, Thomson NR, et al. Public Health Value of Next-Generation DNA Sequencing of Enterohemorrhagic Escherichia coli Isolates from an Outbreak. J Clin Microbiol. 2013;51(1):232-237.
423. Butcher H, Elson R, Chattaway MA, et al. Whole genome sequencing improved case ascertainment in an outbreak of Shiga toxin-producing Escherichia coli O157 associated with raw drinking milk. Epidemiol Infect. 2016;144(13):2812-2823.
424. Dallman T, Inns T, Jombart T, et al. Phylogenetic structure of European Salmonella Enteritidis outbreak correlates with national and international egg distribution network. Microb Genom. 2016;2(8):e000070.
425. Nash AK, Auchtung TA, Wong MC, et al. The gut mycobiome of the Human Microbiome Project healthy cohort. Microbiome. 2017;5(1):153.
426. James S, Phillips S, Baker D, Hall L, Carding S. Characterising the human intestinal mycobiome during healthy ageing. Access Microbiology. 2019;1(1A).
427. Shkoporov AN, Clooney AG, Sutton TDS, et al. The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. Cell Host Microbe. 2019;26(4):527-541.e525.
428. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.
429. Schwarze K, Buchanan J, Fermont JM, et al. The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. Genetics in Medicine. 2020;22(1):85-94.
430. Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265-269.

431. Loman NJ, Constantinidou C, Christner M, et al. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. Jama. 2013;309(14):1502-1510.
432. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think!
433. Hoenen T, Groseth A, Rosenke K, et al. Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool. Emerg Infect Dis. 2016;22(2):331-334.
434. Faria NR, Sabino EC, Nunes MRT, Alcantara LCJ, Loman NJ, Pybus OG. Mobile real-time surveillance of Zika virus in Brazil.
435. Leggett RM, Alcon-Giner C, Heavens D, et al. Rapid MinION profiling of preterm microbiota and antimicrobial-resistant pathogens. Nature Microb. 2020;5:430-442.
436. Gu W, Miller S, Chiu CY. Clinical Metagenomic Next-Generation Sequencing for Pathogen Detection. Annu Rev Pathol. 2019;14:319-338.
437. Pearson K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine. 1901;2:559–572.
438. Hill MO. Correspondence Analysis - neglected multivariate method. Applied Statistics. 1974;23:340–354.
439. Hodges K, Gill R. Infectious diarrhea – Cellular and molecular mechanisms. Gut Microbes. 2010;1(1):4-21.
440. Buret AG. Mechanisms of epithelial dysfunction in giardiasis. Gut. 2007;56(3):316-317.
441. Chin AC, Teoh DA, Scott KGE, et al. Strain-dependent induction of enterocyte apoptosis by Giardia lamblia disrupts epithelial barrier function in a caspase-3-dependent manner. Infect Immun. 2002;70(7):3673-3680.
442. Troeger H, Epple HJ, Schneider T, et al. Effect of chronic Giardia lamblia infection on epithelial transport and barrier function in human duodenum. Gut. 2007;56(3):328-335.

Appendix I: Ethics application form and approval letter for collection of faecal samples

Stool samples for GI tract infection diagnosis Protocol V1.0 9/11/2015

Application for retrieving faecal samples of patients with GI tract infections via UEA biorepository.

Background

Gastrointestinal infections continue to be a considerable cause of morbidity and mortality. The incidence of infectious intestinal disease (IID) or gastroenteritis in the UK is 17 million cases annually, which translates to the high burden of one million general practice consultations a year. The economic impact of IID is considerable, as it accounts for a loss of 11 million working days and 8 million absences from school leading to high levels of healthcare usage.

Current investigation of faecal specimens involves a range of different techniques, including culturing, PCR, enzyme immunoassays and microscopy. However such testing techniques are still unable to find an aetiological agent in 40% of the cases in the UK. Identification of the aetiological agent of the infection is important and will provide information beneficial for both patient and public health. Treatment of the patient could be more specific, especially in severe cases when a specific antibiotic therapy may be recommend. In outbreak cases, identification of the aetiological agent would aid in intervention and control measurements.

Clostridium difficile remains a major cause of nosocomial gastrointestinal infections which is on the increase in incidence. *C. difficile* infections (CDI) follows disruption of the normal GI tract microbiota, often following antibiotic treatment. Pathogenesis involves toxin production, including the two major exotoxins A and B, with severe infections leading to pseudomembranous colitis and death in some cases . It is estimated that in UK there are more than 20000 cased per annum. CDI is currently treated with antibiotics such as vancomycin, but post-treatment relapse is as high as 20% [3]. Some strains of *C. difficile* are becoming resistant to these antibiotics (AMR) and the lack of new classes of antibiotics coming on the market is now a major global issue in tackling the problem of AMR. Alternatives to antibiotics are urgently needed and recently NICE has approved the use of faecal microbiota transplant (FMT) for treatment of recurrent CD and IFR/NNUH have started to use this treatment. However in order to improve the application and efficacy of the FMT we would like to understand the underlying alterations of the microbiome in the patients.

Aims and Objectives

We want to exploit the recent advances in next generation sequencing technologies and advanced metabolomics to developed biomarkers of disease with improved increased sensitivity for detection and diagnosis of infections.

We are going to evaluate the use of these techniques on faecal specimens collected from patients with GI tract infection. We also want to use the same techniques to assess the changes in structure and function of the gut microbiomes in patients treated with FMT

Methodology

We will collect archived faecal samples from the Microbiology lab of the local NHS Trust. We plan to collect and analyse 300 to 400 samples in total over four years. These samples have been sent from either GPs or from the hospital for diagnostic purpose and they are from patients with GI tract infections. A sub-sample of the faecal sample is used for diagnosis and the remainder is stored in the microbiology Dept. for 7 days and at that stage is part of the NNUH's diagnostic archive. The samples we plan to use are part of the NNUH diagnostic archive which falls under the remit of the Biorepository. In accordance with the NRES ethics approval for the Biorepository (Ref: 08/h0304/85+5) samples in the diagnostic archive can be used for research purposes with appropriate ethics approval in place. There is also no requirement for patients to have provided additional consent for their samples to be used in approved research.

9/11/2015

Stool samples for GI tract infection diagnosis Protocol V1.0 9/11/2015

We will collect from the faecal microbiota transplant (FMT) related faecal samples from the UEA biobank. These samples have been deposited (with full consent of the patients and the donors) in the biobank during the NICE approved FMT treatment of the elderly patients with recurrent CDI. The samples deposited are the donor faecal material as well as the pre and post-transplant samples of the patients,

The samples we will collect are completely anonymised and we will not receive any information on the patient other than if they have been diagnosed with any pathogens or in the case of FMT samples if they are of the donor or pre or post FMT patients..

For metagenomic techniques DNA will be extracted from these faecal samples using proprietary kits either at IFR or at BCRE for their DNA extraction facility. The exact methods will be subject to optimization during the project. Any human DNA sequences present will be discarded. Faecal water will also be prepared from the same faecal material for metabolite analysis via proton NMR and/or GS/LS-MS to assess metabolites present.

Metagenomic and metabolic information from healthy individuals and patients samples will be compared to evaluate the use of these techniques in diarrheal infections and also for monitoring the efficacy of infection treatment.
General consent is therefore requested to retrieve from biorepository, diagnostic samples with viral, bacterial or parasite infection and for comparative analysis samples without gastrointestinal infection in order to understand pathophysiological of large bowel intestinal disease and improve accuracy, specificity of diagnostic process and to identify novel biomarkers of GI tract infections. The samples received will be anonymised and will have already been analysed with relevant diagnostics and routine clinical tests.

9/11/2015

Faculty of Medicine and Health Sciences Research Ethics Committee

University of East Anglia

Arjan Narbad
IFR

Research & Enterprise Services
West Office (Science Building)
University of East Anglia
Norwich Research Park
Norwich, NR4 7TJ

Telephone: +44 (0) 1603 591720

Email: fmh.ethics@uea.ac.uk

Web: www.uea.ac.uk/researchandenterprise

20th November 2015

Dear Arjan,

Project title: The microbiome: the role of microbial communities in gastrointestinal infections.
Reference: 20152016 31 HT

The submission of your above proposal for the use of Human Tissue has been considered by a Sub-Committee of the Faculty Research Ethics Committee and we can confirm that your proposal has been approved.

Please could you ensure that any further amendments to either the protocol or documents submitted are notified to us in advance and also that any adverse events which occur during your project are reported to the Committee. Please could you also arrange to send us a report once your project is completed.

Can I remind you that any tissue used should be destroyed at the end of the experiment, or, with prior arrangement returned to the Tissue Bank.

The Committee would like to wish you good luck with your project

Yours sincerely

P.P

Linda Harvey
Chair Human Tissue - FMH Research Ethics Committee

Appendix II: Gut microbiota metabolite concentrations during gastrointestinal infection

	Control (n=41)	Bacteria (n=134)	Protozoa (n=112)	*Campylobacter* (n=107)	*S. enterica* (n=27)	*Cryptosporidium* (n=24)	*G. lamblia* (n=88)
2,3-Butanediol	14.64	92.16	37.66	82.01	132.4	34.47	38.53
2-methylbutyrate	72.39	61.68	113.79	63.48	54.54	107.91	115.39
3-phenylpropionate	27.21	24	39.35	24.28	22.87	22.96	43.82
acetate	3339.79	6778	7993.03	6904.29	6277.51	6306.94	8452.88
alanine	104.83	552.92	368.23	569.44	487.47	386.58	363.22
arabinose	17.66	47.64	33.36	37.6	87.39	16.69	37.91
aspartate	80.57	89.22	152.62	79.11	129.28	139.18	156.28
butyrate	1088.56	1225.28	2154.95	1244.87	1147.66	1732.85	2270.07
cadaverine	24.4	205.68	131.58	220.09	148.54	109.49	137.61
caprate	55.13	39.53	59.38	40.05	37.49	54.43	60.73
caproate	131.3	48.22	98.65	47.65	50.46	94.26	99.84
caprylate	65.41	54.3	73.66	55.47	49.65	68.65	75.03
citrulline	16.22	56.12	33.68	58.35	47.29	32.41	34.02
ethanol	150.98	384.52	256.41	370.21	441.23	278.32	250.44
formate	15.4	28.94	31.44	28.96	28.84	45.2	27.69
fucose	19.25	31.28	31.46	31.22	31.49	22.4	33.93
galactose	28.5	52.46	43.32	41.14	97.31	23.81	48.64
glucose	218.59	247.24	378.62	227.07	327.14	269.6	408.35
glutamate	284.38	377.31	349.3	363.82	430.73	324.84	355.98
glutamine	30.39	87.98	68.86	89.57	81.7	65.84	69.68
glutarate	30.02	66.94	52.27	69.21	57.95	46.95	53.72
glycerol	90.3	81.01	286.17	74.8	105.63	197.19	310.43
glycine	91.08	422.63	274.59	416.98	445	235.39	285.28
hypoxanthine	26.78	19.18	37.73	19.37	18.4	30.69	39.64
isobutyrate	112.87	178.47	236.56	183.59	158.18	196.66	247.44
isoleucine	60.67	236.05	161.77	233.55	245.96	194.4	152.86
isovalerate	129.15	162.16	279.05	171.44	125.38	249.63	287.07
lactate	20.39	572.29	256.02	469.85	978.26	126.66	291.3
leucine	69.79	432.52	216.78	451.94	355.52	282.06	198.98
lysine	44.99	215.31	189.91	217.52	206.56	178.9	192.91

	Control (n=41)	Bacteria (n=134)	Protozoa (n=112)	*Campylobacter* (n=107)	*S. enterica* (n=27)	*Cryptosporidium* (n=24)	*G. lamblia* (n=88)
malonate	35.08	125.84	75.28	116.01	164.78	75.35	75.27
methanol	39.3	65.21	45.12	66.61	59.69	37.13	47.3
methionine	29.94	89.99	61.44	92.08	81.7	71.5	58.7
methylamine	17.34	39.13	55.87	38.73	40.71	39.96	60.21
N6-acetyllysine	9.91	41.49	65.43	42.9	35.9	48.24	70.12
ornithine	21.42	128.23	85.21	131.54	115.11	82.08	86.07
phenylacetate	47.53	64.69	121.19	69.6	45.2	104.7	125.69
phenylalanine	40.29	224.28	117.96	230.36	200.17	154.24	108.07
proline	30.29	192.09	101.64	198.05	168.5	92.47	104.14
propionate	867.35	1945.04	2516.18	2005.54	1705.29	2056.93	2641.43
putrescine	3.89	144.44	64.57	143.95	146.39	54.53	67.31
ribose	117.6	73.14	184.4	66.56	99.2	132.42	198.58
serine	64.47	141.69	145.59	129.06	191.74	128.19	150.34
succinate	64.75	1271.91	424.7	1220.44	1475.9	284.12	463.05
taurine	21.39	139.81	46.5	143.97	123.33	28.51	51.4
threonine	74.19	136.88	144.79	129.45	166.34	128.24	149.3
trimethylamine	8.35	46.37	43.28	47.39	42.34	39.78	44.23
tryptophan	8.11	43.29	23.99	44.07	40.19	28.5	22.76
tyramine	1.59	26.65	10.03	27.57	23.03	5.18	11.35
tyrosine	48.7	199.97	115.28	207.06	171.89	148.65	106.18
uracil	42.59	46.8	89	44.57	55.66	68.4	94.61
valerate	199.1	153.15	403.02	152.64	155.18	288.12	434.35
valine	81.23	398.4	280.06	406.36	366.85	324.85	267.84
xanthine	10.23	26.74	45.94	26.84	26.33	30.47	50.16
xylose	37.86	56.41	46.08	52.81	70.64	34.33	49.28
myo-inositol	9.07	36.76	29.58	36.47	37.92	33.12	28.61