Genetic Diversity and Antimicrobial Resistance in a Global Collection of the Emerging Human Pathogen *Salmonella* Infantis

Jennifer Mattock

Thesis submitted in fulfilment of the requirement for the degree of
Doctor of Philosophy

University of East Anglia
Norwich Medical School
February 2020

Abstract

High levels of *Salmonella* Infantis have been identified in poultry globally; for several years it has been the serovar most frequently found in broilers in Europe. It is also one of the most common serovars causing infection in humans, although responsible for lower numbers of cases than other *Salmonella* serovars.

Very little is known about the genetic diversity of *S*. Infantis, with previous research only comparing up to 264 genomes. In this project a collection of 4,670 *S*. Infantis genomes was amassed. The aims of this thesis were to determine: the global population structure of the serovar; the levels of antimicrobial resistance (AMR) and plasmids and whether genetic differences between human and poultry *S*. Infantis could explain the difference in incidence seen between these sources.

S. Infantis splits into two eBurstGroups (eBG), eBG31 and eBG297, the former comprising 96% of the global collection. However, the proportion of isolates belonging to either eBG varied geographically, with eBG297 strongly associated with isolation from Africa.

High levels of AMR were present in the eBG31 population; 39% of the isolates were multidrug resistant. This was associated with the presence of plasmid of emerging *S*. Infantis, which was identified in 34% of the eBG31 genomes, in particular from 69% of the poultry isolates.

Upon comparison of the eBG31 human and poultry genomes, a greater genetic diversity was observed amongst the human isolates. Furthermore, several thousand genes and intergenic regions were significantly associated with isolation source. This thesis concluded that the differences in the pathogenicity of *S*. Infantis between humans and poultry is due to either only a subgroup of poultry *S*. Infantis being capable of infecting humans; or that other sources are the cause of human infections. Public health teams worldwide will benefit from the increased understanding this work provides on this emerging pathogen.

Table of Contents

List of Figures

List of Tables

List of Accompanying Material

The following files are available in the CD accompanying this thesis:

shorten_node_names.pl
get_eBG31_representitives.py
Infantis_Metadata_and_Results.xlsx
Annotated_eBG31_Phylogeny.pdf
Associated_Virulence_Factors.xlsx
Associated_Unitigs_Region1.gff
Associated_Unitigs_Region2.gff
Associated_Unitigs_Region3.gff

Information about the scripts and legends for each table and figure can be found in Appendix VI. Electronic Appendices.

Declaration

The work presented in this thesis was my own work, except when expressly mentioned. Two journal articles are currently in preparation, listed in Appendix VII. Publications Arising from this Thesis, VII.III Journal Articles. None of the results to be included in these articles has been presented in this thesis, with the exception of Figure IV.2 pESI presence in eBG31 from England & Wales and South Africa. This figure, which I created, is planned to be included in the following article: Distinct genetic phylogeny in human *Salmonella* Infantis from South Africa and the United Kingdom: implications for management.

Acknowledgements

I would like to thank Paul Hunter and the Health Protection Research Unit in Gastrointestinal Infections for giving me the opportunity to do this project. I would also like to thank Roberto La Ragione for suggesting I focus on *Salmonella* and volunteering the use of his chicken caecum model, even if that couldn't happen due to time constraints.

I am particularly grateful to my collaborators in this project, Liljana Petrovska for generously sharing *S*. Infantis sequence data and DNA; Karen Keddy and Anthony Smith for welcoming me into their lab and Shannon Smouse, Tina Duze and Nomsa Tau for their tireless efforts in helping me find and extract the DNA of the South African isolates. Also, all of the people at PHE who helped me including: Marie Anne Chattaway for her guidance and for letting me come to PHE during the peak season; Anaïs Painset for her help in installing SnapperDB; Tim Dallman and Hassan Hartman for sharing sequence data and their guidance with database and phylogeny creation; David Greig for his tips with long-read sequence data assembly; Martin Day, Claire Maguire and Tracey Dealey for helping me find the historical PHE isolates and Amina Ismail for extracting the DNA.

I would also like to thank Alison Mather for the guidance she gave me with identifying pESI presence in my genomes; Dave Baker for the sequencing that he did for my project and Emma Manners for the training she gave me in DNA quantification and library preparation. I am particularly grateful for everyone in the MMRL for being so welcoming, supportive and letting me moan about *Salmonella* whenever I needed to.

My greatest appreciation goes to John Wain and Gemma Langridge for welcoming in me into their team and providing all the supervision for the majority of my project. This project would have been very different and nowhere near as enjoyable to do if not for your help and I will be eternally grateful to you both for "adopting" me, for your support and for all your proof reading.

Finally, I would like to thank my parents for their encouragement, my grandparents for always being on the other end of the phone to talk about my problems, my friend Heather for always giving me opinions on the aesthetics of my graphs and most of all, my boyfriend Chris for his tireless support.

1. Chapter 1. Introduction

Salmonella enterica subspecies *enterica* serovar Infantis (*S*. Infantis) is a bacterial pathogen which causes disease in humans. Although one of the serovars most frequently isolated in public health reference laboratories, often amongst the top six, little is known about transmission routes and reservoirs (National Center for Emerging and Zoonotic Infectious Diseases (NCEZID), 2018; European Food Safety Authority (EFSA) and European Centre for Disease Prevention and Control (ECDC), 2019a). In chickens, *S.* Infantis infection is common, with the serovar being the most frequently identified from broilers in European Union (EU) member states (EFSA and ECDC, 2019a). The genetic basis for the difference in pathogenicity between isolates causing infection in humans and poultry, which could explain why *S.* Infantis is not seen in higher numbers in humans, is not clear.

1.1 Infectious Intestinal Disease

Diarrhoea, or infectious intestinal disease (IID), is a very variable disease state ranging from chronic intestinal disorders to dramatic projectile vomiting after ingestion of a toxin; in order to study causation, definitions are needed. IID is defined by the Food Standards Agency (FSA) of the United Kingdom (UK) as individuals suffering from:

> "loose stools or clinically significant vomiting lasting less than 2 weeks, in the absence of a known non-infectious cause, preceded by a symptom-free period of 3 weeks" (FSA, 2016).

It can be caused by a diverse array of microorganisms including bacteria, viruses and protozoa and is the second most prevalent infectious disease occurring in humans after respiratory tract infection (Vos *et al.*, 2016).

1.1.1 Occurrence and Causes of IID

IID can have serious consequences, with pathogens that cause diarrhoea accounting for 10.5% of deaths of children under the age of 5 worldwide (Liu *et al.*, 2012). In 2016 there was an estimated 4.5 billion episodes of diarrhoeal disease and 1.7 million deaths, with diarrhoea being the second greatest cause of years of life lost after cardiovascular diseases (Global Health Metrics, 2017; Troeger *et al.*, 2018). The mortality associated with diarrhoeal disease varies greatly by location, with 89% of deaths due to diarrhoea occurring in sub-Saharan Africa and South Asia (Troeger *et al.*, 2018).

The incidence of IID in the UK community is high, with 25% of the population having an episode each year (FSA, 2016). There were approximately 16.9 million cases in 2009, resulting in 1 million visits to a general practitioner for IID (Tam *et al.*, 2012). This has an impact on the economy; it has been calculated that each year in the UK, 50% of those with IID take time off work/school, equating to 19 million work days lost, 11 million of those at working age (FSA, 2016).

In England and Wales bacterial pathogens were the most common cause of gastrointestinal infection reported to Public Health England (PHE) in 2014, with *Campylobacter* spp., *Salmonella enterica*, *Shigella sonnei* and *Escherichia coli* (*E. coli*) being the most prevalent bacterial pathogens (PHE, 2015). Norovirus and Rotavirus were the most common viral enteropathogens and the protozoa *Cryptosporidium* and *Giardia* were also frequently reported. It is estimated that in the community viral enteropathogens are the most prevalent cause of IID (Tam *et al.*, 2012).

1.1.2 Transmission Routes

IID can be spread via multiple routes including: from person-to-person via the faecal-oral route; by animal contact or contact with environmental contamination; or by contact with contaminated water or food (Fletcher, Stark and Ellis, 2011; Centers for Disease Control and Prevention (CDC), 2018).

Person-to-person transmission of IID usually occurs due to contamination of the hands with faeces (Rao, 1995). The risk of spreading IID via this route can be reduced by hand washing after possible contamination or before preparing or eating food (Ejemot-Nwadiaro RI and Critchley, 2015). Decontamination of the environment may also be helpful, for example, in hospitals door handles and other surfaces should be disinfected; the transfer of symptomatic patients or staff between wards in hospitals can also increase the spread of IID (Rao, 1995). Another source of person-to-person spread can be sexual contact, specifically oro-anal sex (Farthing and Kelly, 2007). Pathogens that require a low infectious dose for transmission are more likely to be spread from person to person, such as *Shigella*. In the case of some pathogens, person-to-person contact can have an important role in transmission; for example, children with haemolytic uraemic syndrome (HUS), due to verotoxin-producing *E. coli*, are significantly more likely than controls to have come into contact with someone with diarrhoea up to 2 weeks prior to the onset of infection (Rowe *et al.*, 1993).

Animals are a reservoir for human enteropathogens, including *Salmonella;* direct contact with animals, or contamination of the environment due to animals, can lead to human IID (Zambrano *et al.*, 2014). Domestic livestock or poultry being kept in or near the home increase the risk of the spread of faecal contamination and therefore transmission of pathogens. Household pets can also be a source of human IID as *C. jejuni* and *S. enterica* carriage has been observed in cats and dogs (Shimi and Barin, 1977; Svedhem and Kaijser, 1981; Leonard *et al.*, 2011). For example, in a study of 138 dogs in Canada, 23% of the dogs had a stool sample positive for *Salmonella*, with 87.5% of those dogs not having experienced diarrhoea in a month (Leonard *et al.*, 2011). Human *Salmonella* infections have also been found to be associated with dry cat and dog food (Behravesh *et al.*, 2010). Petting farms have been connected with outbreaks of human IID, for example, in North Carolina 108 people fell ill after attending a state fair, 78% of which had visited the petting zoo at the fair (CDC, 2005). Shiga toxin-producing *E. coli* (STEC) was the confirmed cause in 41 of the cases and extensive STEC contamination was found at the petting zoo; 14% of patients developed HUS.

Enteropathogens have been identified in the environment, another route for IID transmission. A study in Tanzania identified enterovirus, pathogenic *E. coli* and rotavirus genes in soil samples, with higher levels of *E. coli* found in soil from house floors than latrine floors (Pickering *et al.*, 2012). *Clostridium difficile* has been identified in the environment in Wales; of 104 soil samples taken, 21% were positive for the pathogen (Al Saif and Brazier, 1996).

Enteropathogens can also be transmitted in water; between 1974 and 2001 in Canada there were 288 outbreaks associated with drinking water, with the top three causative agents being *Giardia*, *Campylobacter* and *Cryptosporidium* (Schuster *et al.*, 2005). In England and Wales between 1992 and 2003 there were 89 waterborne IID outbreaks; 55% were associated with drinking water and 39% with swimming pools; the top three causative agents were *Cryptosporidium*, *Campylobacter* and *Giardia* (Smith *et al.*, 2006). *Cryptosporidium* and *Giardia* are both protozoa which are capable of surviving for long periods in water and resisting disinfection, they also both cause low numbers of fatalities and hospitalisations (Medema and Schijven, 2001; Snel *et al.*, 2009).

It is estimated that globally, 600 million people become ill after eating contaminated food each year; 420,000 of these die (World Health Organization (WHO), 2015). In the United States of America (USA) in 2011 there were approximately 48 million cases of foodborne illness, resulting in 128,000 hospitalisations and 3,000 fatalities (CDC, 2018). Also, in Japan over 1,000 outbreaks of foodborne disease occur each year (Hara-Kudo *et al.*, 2013).

1.1.3 Foodborne Infection

A foodborne infection is an infection that results from ingesting food contaminated with a pathogen (Plaut, 2000). The WHO estimated that in 2010, approximately 92% of cases of foodborne illnesses were caused by diarrhoea causing infectious agents; bacterial pathogens causing 64% of these cases (Havelaar *et al.*, 2015). The top three bacterial pathogens were *Campylobacter spp.*, Enterotoxigenic *E. coli* and *Salmonella enterica*, causing in total 75% of the cases. *Salmonella enterica* was predicted to be the pathogen to cause the most fatalities, accounting for 32% of deaths caused by bacterial diarrhoeal foodborne infections. Other important bacterial foodborne pathogens include *Shigella spp.* and STEC which caused 51 million and 1 million cases respectively.

Viral pathogens also have a significant foodborne infection burden; in 2010 an estimated 125 million cases of food-associated norovirus infection occurred (Havelaar *et al.*, 2015). Protozoa such as *Giardia* spp. and *Cryptosporidium* spp. were also predicted to be responsible for 67 million cases of diarrhoea associated with foodborne infection.

1.1.4 Sources of Bacterial Foodborne Infection

Multiple different food groups have been identified as causes of bacterial foodborne infection. In 2005 and 2006, 135,014 *Salmonella* outbreak-associated human cases, with a known source attributed to them, were identified in EU member states; higher than seen with *Campylobacter* (129,603) (Pires *et al.*, 2010). Land animals were the source associated with the highest proportion of *Salmonella* and *Campylobacter* cases, causing 94% and 92% of cases respectively. Of the *Salmonella* cases attributed to land animals, 65% were associated with eggs and 24% with meat and poultry.

In the USA between 1998 and 2008, products from land animals caused the majority, 64%, of bacterial foodborne infections. Other sources included plants (32.1%) and aquatic animals (3.9%) (Painter *et al.*, 2013). Of all land animal produce, meat and poultry caused the highest number of bacterial foodborne illnesses (41.1%); 17.9% were attributed to poultry, 13.2% to beef, 9.8% to pork and 4.9% to eggs. Dairy products were also associated with a large number of cases (18%). The largest contributor to death from foodborne illnesses was poultry, associated with 19% of all deaths, which were in most cases caused by *Listeria* or *Salmonella*. This demonstrates the need for research in reducing the levels of pathogenic bacteria in chickens.

Using a combination of public health data and estimates of IID incidence in the community in the UK from 2009, *Campylobacter* was predicted to be the most frequently identified foodborne pathogen, causing 280,400 cases (O'Brien *et al.*, 2016). However, although *Salmonella* was seen less frequently, it was predicted to cause the highest number of hospitalisations at 2,490, compared to 562 admissions due to *Campylobacter* infection. Also, in Canada between 2000-2010, *Salmonella* was the bacterial pathogen that caused the highest number of domestic foodborne hospitalisations (Thomas *et al.*, 2015). This suggests that *Salmonella* causes large numbers of foodborne infections but also causes more severe infections than other foodborne pathogens.

1.2 Salmonella

Salmonella is part of the *Enterobacteriaceae* family and is comprised of gram negative rod shaped bacteria; the majority of the genus is motile with several peritrichous flagella present on the cell surface (Sanderson and Nair, 2013; Jajere, 2019).

1.2.1 Classification

The *Salmonella* genus is split into two species, *Salmonella enterica* and *Salmonella bongori;* the former is further split into six subspecies (subsp.): *enterica* (subsp. I), *salamae* (subsp. II), *arizonae* (subsp. IIIa), *diarizonae* (subsp. IIIb), *houtenae* (subsp. IV) and *indica* (subsp. VI) (Fookes *et al.*, 2011; Crump and Wain, 2017). Figure 1.1 shows the phylogenetic relationship between the *Salmonella* species and subspecies.

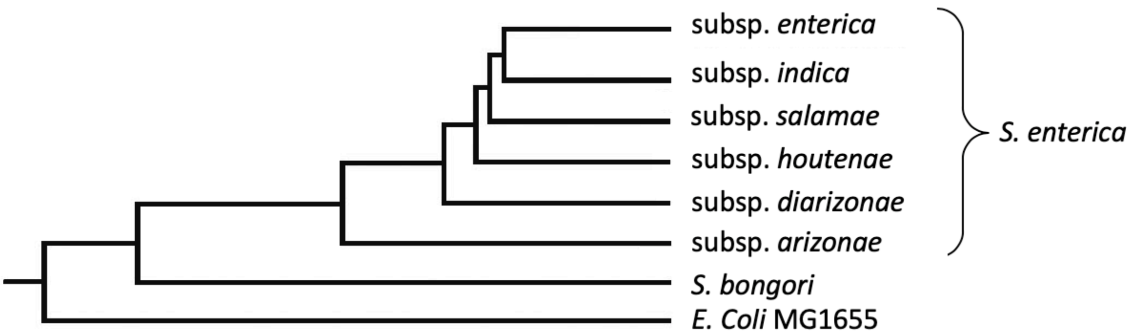


Figure 1.1 *Salmonella* subspecies evolution

Phylogram of the *Salmonella enterica* subspecies, *Salmonella bongori* with an *E. coli* strain as an outgroup. Reconstructed from Desai *et al.*, 2013.

1.2.1.1 *Salmonella enterica*

Of the six subspecies within *Salmonella enterica*, the host each is associated with varies (Crump and Wain, 2017). They have varying biochemical characteristics which can be used to differentiate between them; for example all but *S. enterica* subsp. *enterica* are gelatinase positive and only *S. enterica* subsp. *houtenae* is salicin positive (Lamas *et al.*, 2018). It is estimated that *S. enterica* subsp. *salamae* and *indica* are the closest related to *S. enterica* subsp. *enterica*; diverging 20 million years ago.

S. enterica subsp. *enterica* contains 99% of the serovars that cause animal and human disease and is comprised of 2637 serovars; the majority of which can cause foodborne infection (Issenhuth-Jeanjean *et al.*, 2014; Crump and Wain, 2017). Henceforth, all *S. enterica* subsp. *enterica* serovars will be referred to with the genus and serovar name. The serovars are clustered based on their host specificity and the type of infection they cause; which can range from invasive systemic infection to asymptomatic carriage (Gal-Mor, Boyle and Grassl, 2014). The serovars in this subspecies are often split into two groups dependent on their ability to cause typhoid fever; *S.* Typhi and *S.* Paratyphi A, B and C cause either typhoid fever or paratyphoid fever and are thus referred to as Typhoidal *Salmonella*. The other *S. enterica* subsp. *enterica* serovars, which cause gastroenteritis or extra-intestinal infection are called Non-Typhoidal *Salmonella* (NTS).

Although *S. enterica* subspecies *arizonae*, *diarizonae* and *salamae* have been identified in warm-blooded animals, and all of the subspecies have caused human cases of salmonellosis, their presence in these hosts is rare (Lamas *et al.*, 2018). A study in the Netherlands, between 1984 and 2014, determined that each of the *S. enterica* subsp. II-VI caused between 0.02% and 0.06% of human *Salmonella* infections (Mughini-Gras, Heck and van Pelt, 2016). They are more commonly associated with reptiles, in the same study 59% of reptile *Salmonella* samples were from *S. enterica* subsp. II-IV. Reptiles, whilst frequently carrying these subspecies, also carry *S. enterica* subsp. *enterica* as part of their intestinal microbiota, often asymptomatically (Silva, Calva and Maloy, 2014). Interestingly, very high levels of *S. enterica* subsp. *salamae* serovar Sofia have been observed in the Australian poultry industry, between 2005 and 2009 it was the dominant serovar isolated from broiler meat (Duffy, Dykes and Fegan, 2012). This was not associated with high numbers of human cases.

1.2.1.2 *Salmonella bongori*

S. bongori is primarily associated with cold-blooded animals although cases have been seen in humans and other warm blooded animals (Giammanco *et al.*, 2002; Fookes *et al.*, 2011). It is estimated that *S. bongori* diverged from *S. enterica* 40-63 million years ago (Lamas *et al.*, 2018). All *Salmonella* contain *Salmonella* Pathogenicity Island (SPI)-1; however, whilst *S. enterica* isolates contain SPI-2, *S. bongori* does not; the acquisition of this SPI is predicted to be key in the divergence of these species (McQuiston *et al.*, 2008).

1.2.1.3 *Serology*

The *Salmonella* genus is classified by serotyping against specific antibodies, which involves sorting strains by their antigenic structure (Grimont and Weill, 2007; Ryan, Dwyer and Adley, 2017). This is done with three of *Salmonella*'s surface antigens: Vi capsular antigens, flagellar H antigens and somatic O antigens.

The somatic antigen, a component of lipopolysaccharides in the outer cell membrane, is encoded for by the *rfb* genes (Luk *et al.*, 1993). The H antigen may be present in one of two phases, encoded by the *fliC* and *fljB* genes (McQuiston *et al.*, 2008). Isolates that can only express one antigen are monophasic. For diphasic serovars, whilst these two phases can be detected within a culture at the same time, individual bacteria only express one at a time.

Once the antigens present have been identified, the Kauffman-White scheme is used to determine the serovar, 67 O and 117 H antigens have been recognised (Grimont and Weill, 2007; Sanderson and Nair, 2013). There are a total of 2,659 serovars in the *Salmonella* genus (Table 1.1), with the majority of those identified belonging to *S. enterica* subsp*. enterica* (Issenhuth-Jeanjean *et al.*, 2014). Whilst serology is a well-established method for characterising *Salmonella* strains globally, it is expensive, low-throughput and prone to human error (Achtman *et al.*, 2012).

Species and subspecies	Number of serovars
S. enterica	2637
subsp. *enterica*	1586
subsp. *salamae*	522
subsp. *arizonae*	102
subsp. *diarizonae*	338
subsp. *houtenae*	76
subsp. *indica*	13
S. bongori	22

Table 1.1 Number of serovars in each *Salmonella* species and subspecies

Number of serovars from each *Salmonella* species and subspecies in the current Kauffman-White scheme (Issenhuth-Jeanjean *et al.*, 2014).

1.2.1.4 *Other Typing Methods*

To allow for differentiation of isolates within serovars, other methods are being used alongside serotyping. One method used to identify variants within serovars is phage typing; the bacterium's susceptibility to a set of bacteriophages is determined, this information can then be used to determine whether a particular variant is being spread (Tizard, 2004). Pulsed-field electrophoresis is also used to identify variants of *Salmonella*; restriction endonucleases are used to digest the chromosomal deoxyribonucleic acid (DNA) into fragments that are separated using electrophoresis; the results of different strains are then compared to show to relatedness. Other methods of differentiating between variants of *Salmonella* include antimicrobial resistance (AMR) typing, plasmid typing, biochemical tests and single nucleotide polymorphism (SNP) detection in the flagellar antigens (Achtman *et al.*, 2012; Sanderson and Nair, 2013; Crump and Wain, 2017).

Due to the expense, the need for animals to raise antibodies and the low throughput of serotyping, new techniques are being used to classify *Salmonella* (Achtman *et al.*, 2012). Multi-locus sequence typing (MLST), identifies the variant of each of the 7 housekeeping genes present in the genome; genes that are ubiquitous in *Salmonella* and involved in metabolism (Cooper and Feil, 2004; Achtman *et al.*, 2012). The alleles are: *purE*, *aroC*, *dnaN*, *hemD*, *hisD*, *sucA* and *thrA* (Achtman *et al.*, 2012). Isolates with identical alleles are grouped into a Sequence Type (ST). STs are then clustered into an eBurstGroup (eBG) based on allelic similarity; resulting in groups of related organisms. All of the STs within an eBG are a separated from another ST in the eBG by a single-locus variant (SLV), an alternate version of one of the 7 housekeeping alleles used to define that ST. In some cases, when defining eBGs, singleton STs that were a double locus variant from an eBG and a common serovar were also included in that eBG. As *fliC* and *fljB* can be horizontally transferred between strains, this results in isolates being serotyped as a serovar that they are genetically distinct to. MLST avoids this problem and is therefore believed to be a superior typing method.

1.2.2 *Salmonella* Occurrence

In 2010 there was an estimated 9 million cases globally of enteric fever caused by *S.* Typhi and *S.* Paratyphi A and 79 million cases of NTS foodborne infection (Havelaar *et al.*, 2015). The economic cost of NTS is therefore extensive, NTS infections, which caused either diarrhoea or invasive infection, had the biggest economic burden in 2010 out of all foodborne diseases worldwide, resulting in 4.07 million disability adjusted life years, a measurement of overall disease burden and recording of the number of years lost due to ill health (Kirk *et al.*, 2015). In 2013, the estimated cost of human NTS infection in the United States in 2013 was $3.7 billion (Economic Research Service and U.S. Department of Agriculture, 2014).

1.2.2.1 *NTS Occurrence in Humans*

NTS are the most frequently identified causative agents of foodborne outbreaks in Europe and the second most common cause of reported bacterial gastrointestinal infection (Papadopoulos *et al.*, 2017; EFSA and ECDC, 2019a). After several years of declining levels of human cases of salmonellosis in EU member states, over the last five years the numbers of cases have stabilised, with 91,857 cases confirmed in 2018 from 28 European countries (EFSA and ECDC, 2019a). In 2018, 1,229 outbreaks of *Salmonella* in humans occurred, with *S.* Enteritidis causing the majority. Concerningly, in the USA the number of cases of salmonellosis in humans has significantly increased between 2015 and 2018; the authors attribute this partly to new non-culture based methods being used to diagnose cases, but this could represent a true increase (Tack *et al.*, 2019).

Children are more likely to get *Salmonella* infections than adults; in 2015 in the USA there were 7,719 laboratory confirmed cases of *Salmonella*, 38% of which were in people aged 0-19 years (CDC, 2017). The majority of these were under the age of 5. In Shanghai, China, between July 2010 and December 2011, 1,833 cases of children with suspected bacterial diarrhoea were recognised; 17.2% of the cases were due to NTS, 7.1% to *Campylobacter* and 5.7% to *Shigella* (Li *et al.*, 2014). As children excrete *Salmonella* in their faeces for longer than adults post-infection, this increases this risk of transmission to others (Buchwald and Blaser, 1984).

With large outbreaks it is often possible to attribute cases to a common cause, however, 60-80% of salmonellosis cases are not identified as part of an outbreak and are either classed as sporadic cases or not classed at all (WHO, 2016). In 2015 in the USA, only 6.9% of diagnosed human *Salmonella* cases were outbreak associated (CDC, 2017). It is thought that for every identified *Salmonella* case, 28.3 cases occur in the community which are never reported, meaning the numbers of cases reported to public health teams are not an accurate representation of actual case numbers. (Scallan *et al.*, 2011).

1.2.2.2 *NTS Occurrence in Poultry*

Salmonella is found at high levels in food animals globally. In EU member states in 2018, 2% of the 14,000 *Gallus gallus* breeding flocks that were tested were positive for *Salmonella*; 4% of the 40,000 laying flocks and 3.5% of the 360,000 broiler flocks tested were also positive (EFSA and ECDC, 2019a). *Salmonella* in eggs also represents the highest risk of foodborne disease. Additionally, *Salmonella* was present in other food animals; of the 35,524 cattle units tested, 4% were positive, furthermore, 41% of the 92,089 pig units that were tested were also positive. In 2017, of the 36,079 samples of fresh broiler meat tested across the European Member States, *Salmonella* was detected in 4.85% (EFSA and ECDC, 2018b).

In Ghana, *Salmonella* was found in 47% of laying hen and broiler samples (Andoh *et al.*, 2016). Between 1998-2008 in Japan, *Salmonella* was isolated in 33.5% of ground chicken samples; it was also found in chicken sold for raw consumption at 12.7% (Hara-Kudo *et al.*, 2013). This shows how *Salmonella* contamination of poultry produce remains a problem worldwide.

As *Salmonella* is still a large cause of severe human infection there is clearly a need for further research in this area. With poultry products causing the highest number of hospitalisation and deaths from *Salmonella* cases; reducing the level of *Salmonella* colonisation in chickens could drastically reduce the numbers of human infection; making research in this area worthwhile (Painter *et al.*, 2013).

1.2.2.3 *Serovar Distribution*

The serovars that are the most prevalent in humans vary depending on the country of isolation (Hendriksen *et al.*, 2011). The most common serovars in humans from 37 countries across the world between 2001 and 2007 were compared; in North America, Australia and New Zealand, *S.* Typhimurium was the most common serovar followed by *S.* Enteritidis. In all other regions, *S.* Enteritidis was the dominant serovar. Figure 1.2 shows more recent data of the top 10 *Salmonella* serovars in the USA, EU and New Zealand.

The most common serovar also varies by isolation source; in 2016 in the USA the serovars causing the most clinical infections in chickens, turkeys, pigs, and cattle were *S.* Enteritidis, *S.* Senftenberg, monophasic *S.* Typhimurium and *S.* Dublin respectively (Morningstar-Shaw *et al.*, 2016). In routine flock screening and environmental sampling, the most common serovars differed from those causing infection; for chickens, turkeys, pigs and cattle they were *S.* Senftenberg, *S.* Hadar, monophasic *S.* Typhimurium and *S.* Cerro respectively.

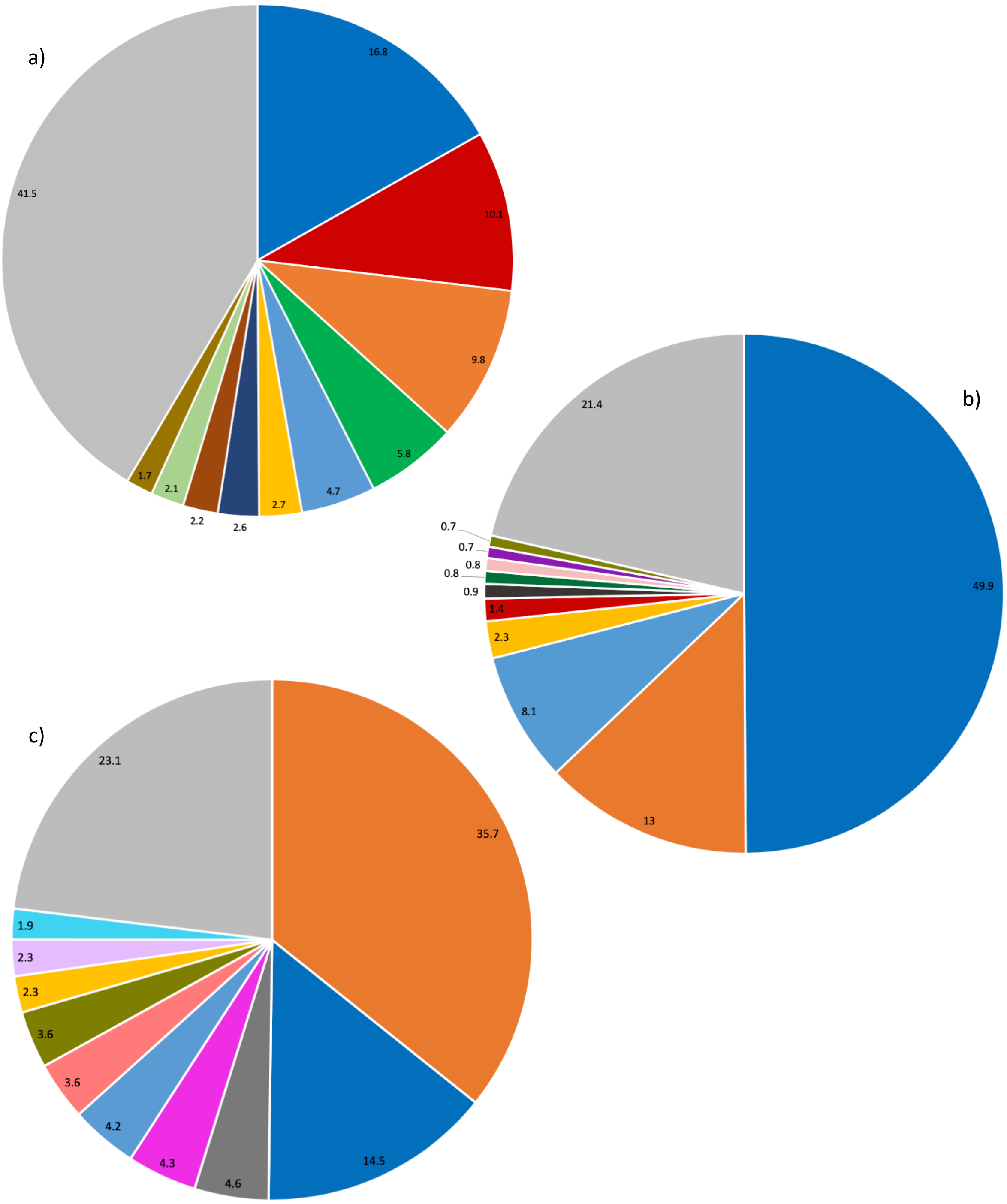


Figure 1.2 Percentage of human salmonellosis cases caused by the top 10 serovars in the USA, EU and New Zealand

a) USA in 2016, n = 46,623
b) EU in 2018, n = 79,698
c) New Zealand in 2019, n = 1153

Agona ■ Bovismorbificans ■ Braenderup ■ Brandenburg ■ Derby ■ Enteritidis ■

Infantis ■ Javiana ■ Kentucky ■ Monophasic Typhimurium 1.4.[5].12:i: ■

Montevideo ■ Muenchen ■ Newport ■ Paratyphi B var. Java ■ Saintpaul ■

Stanley ■ Thompson ■ Typhi ■ Typhimurium ■ Virchow ■ Other serovars ■

Generated using data from: NCEZID, 2018; EFSA and ECDC, 2019; Institute of Environmental Science and Research, 2020

1.2.3 How *Salmonella* Causes Infection in Humans

The infection *Salmonella* causes in humans varies depending on whether it is caused by typhoidal or NTS. Whilst NTS cases are seen more frequently, the consequences of typhoidal infection for the individual are more severe; 0.08% of NTS foodborne associated infections in 2010 resulted in death, compared to 0.7% of typhoidal infections (Havelaar *et al.*, 2015).

1.2.3.1 *Non-Typhoidal Salmonellosis*

NTS typically causes localised gut infections, giving symptoms such as fever, diarrhoea, vomiting and gastroenteritis (Cherubin *et al.*, 1974). The incubation period of the pathogen is usually 12 to 96 hours, although cases with longer periods have been identified, for example, an outbreak of *S.* Nienstedten was identified in a nursery with an incubation period of 7 to 18 days (Seals *et al.*, 1983; Eikmeier, Medus and Smith, 2018). Figure 1.3 shows human host factors that are increased with susceptibility to NTS infection.

- High gastric pH (low acidity) [a]
- Gastric and gastrointestinal surgery [b]
- Antibiotic administration [c]
- Haemoglobin abnormalities (e.g., sickle cell anaemia)
- Cancers
- Leukaemia and lymphoma
- Diabetes mellitus
- Immunosuppressive drugs
- Acquired Immunodeficiency Syndrome (AIDS)

Figure 1.3 Host factors that increase susceptibility to Salmonellosis

a) Normal gastric acidity (pH<3.5) is lethal to *Salmonella*
b) Surgery can inhibit normal gastric emptying and intestinal motility and can be undertaken to reduce gastric acidity
c) Antibiotics alter normal intestinal microflora

Reproduced from (CDC, 2013)

NTS can also cause invasive infection, hereafter referred to as iNTS (Gordon *et al.*, 2008). It was estimated that between 1969 and 1983, 8% of patients admitted to hospital in Manchester (UK), with diarrhoea due to NTS, also had bacteraemia (Mandal and

Brennand, 1988). iNTS infection is particularly common in low to middle income countries where it is linked to immunocompromised individuals, in association with human immunodeficiency virus (HIV), malnutrition or malaria (Ao *et al.*, 2015). Genetic defects, such as those affecting Interleukin (IL)-12 and Interferon gamma, also increase host susceptibility to iNTS infection (Fierer and Guiney, 2001). Globally, in 2010, there was an estimated 3.4 million cases of invasive salmonellosis each year and 681,000 fatalities, with NTS causing up to 39% of community-acquired bacteraemia cases in sub-Saharan Africa (Ao *et al.*, 2015; Uche, MacLennan and Saul, 2017). More recent estimates have predicted that globally, in 2017, there were 535,000 cases of iNTS and 77,500 fatalities, of which 24% were associated with HIV (Stanaway, Parisi, *et al.*, 2019). Other manifestations of NTS infection include *Salmonella* osteomyelitis, systemic lupus erythematosus, septic arthritis, endarteritis and meningitis (Cherubin *et al.*, 1974).

The factors causing salmonellosis to become invasive are not all host driven. Some *Salmonella* serovars are more often associated with invasive disease than others, suggesting some variation in virulence (Kazemi, Gumpert and Marks, 1974; Jones *et al.*, 2008). For example, in humans in the USA between 1996 and 2006, 6% of *S.* Typhimurium cases were invasive whereas 57% of *S.* Choleraesuis and 64% of *S.* Dublin cases were invasive (Jones *et al.*, 2008). Invasiveness can be measured by the invasive index, defined as the percentage of all cases that present as bacteraemia (Langridge, Wain and Nair, 2012). *S.* Choleraesuis and *S.* Enteritidis represent the extremes of the NTS invasiveness spectrum; having an invasive index of 55.2% and 1.8% respectively (Langridge, Wain and Nair, 2012). For some serovars, a complex association with invasiveness exists as differences are reported for a single serovar depending upon where the infection occurs - for example the invasive index of *S.* Dublin ranges from 14.3% in Canada to 71.4% in the USA (Langridge, Nair and Wain, 2009).

1.2.3.2 *Typhoid Fever*

A typical host adapted and restricted *Salmonella* is *S.* Typhi, the causal agent of typhoid fever (Crump and Wain, 2017). The incubation period is typically 10 to 20 days and symptoms initially include malaise, headache and loss of appetite (Stuart and Pullen, 1946). Patients then progress to have a combination of the following symptoms: a fever of on average 40°C, rash, aches and pains, cough and abdominal tenderness. Due to its multi-system nature the infection can be life threatening; causing delirium, intestinal perforation, gastrointestinal bleeding, encephalopathy, sepsis, myocarditis, peritonitis and meningitis (Crump and Wain, 2017; Stanaway, Reiner, *et al.*, 2019). Mortality in the absence of antibiotics is low, in Middlesex Hospital (UK) between 1915-1917, of 1,118 cases of typhoid fever identified, the mortality rate was 7.51% (Webb-Johnson, 1917).

1.2.3.3 *Pathogenesis of Salmonella in Humans*

For *Salmonella* to cause infection it must be ingested and reach the small intestine. Initially NTS and typhoidal serovars both adhere to the epithelium of the small intestine and invade (Gal-Mor, Boyle and Grassl, 2014). This is controlled by SPI-1 which codes for a type III secretion system (T3SS) that injects effector proteins into human epithelial cells; triggering host cell membrane rearrangement and allowing *Salmonella* uptake to occur (Francis, Starnbach and Falkow, 1992; Misselwitz *et al.*, 2011). SPI-1 is also essential for regulating the host immune response as it induces the recruitment of neutrophils and it's T3SS suppresses the expression of proinflammatory cytokines in macrophages (Lou *et al.*, 2019).

T3SSs, encoded by SPI-2, modify the phagosome containing the bacterium to form a *Salmonella* containing vacuole which is resistant to lysosome fusion, enabling intracellular survival (Garcia-Gutierrez *et al.*, 2016). Other T3SS effectors present on SPI-2 include *SifA* which triggers lysosomal hydrolase secretion and *SseL* which controls cell death in macrophages (Jennings, Thurston and Holden, 2017). Most *Salmonella* also contain *iroBCDE* and *iroN* which encode a siderophore that supplies iron and is resistant to the host's antimicrobial peptide lipocalin 2 (Fischbach *et al.*, 2006).

NTS gastroenteritis remains confined to the ileum and colon in immunocompetent patients (Gal-Mor, Boyle and Grassl, 2014). The human immune response to NTS is an increase in helper T cell 1 inducing cytokines; patients with a deficiency in IL-12 and IL-13 have been identified as having a higher risk of invasive salmonellosis (MacLennan *et al.*, 2004; Gal-Mor, Boyle and Grassl, 2014). It is this inflammatory response that triggers the diarrhoea seen in cases of NTS (Galàn, 2001). *S.* Typhimurium has been found to be able to utilise tetrathionate, a respiratory electron acceptor produced during inflammation, allowing it to outcompete the host microbiota (Winter *et al*., 2010).

In cases of iNTS the bacteria are phagocytosed into macrophages which then transports them to systemic sites (Haraga, Ohlson and Miller, 2008). In *S.* Enteritidis, several genes were found to play a role in invasion of Caco-2 cells including the flagellar genes *fljB* and *fljH* and the lipopolysaccharide O antigen genes *rfbM* and *rfbN* (Shah *et al.*, 2012).

While NTS initiates an inflammatory response during invasion, the typhoidal serovars avoid this (Gal-Mor, Boyle and Grassl, 2014). Typhoidal *Salmonella* colonises macrophages and avoids being destroyed by them (Gal-Mor, Boyle and Grassl, 2014). This leads to spreading to the liver, bone marrow, spleen and gallbladder. NTS can be excreted in the faeces for on average 7 weeks after infection in children under the age of five and 3 to 4 weeks in adults and children over the age of 5 (Buchwald and Blaser, 1984). Carriage of *S.* Typhi is seen for 3 months after infection in 10% of recovering, untreated patients; between 1 and 4% of patients carry and excrete *S.* Typhi in their faeces for over 12 months (Parry *et al.*, 2002).

1.2.3.4 *Transmission of Salmonella in Humans*

NTS is transmitted to humans through the ingestion of the pathogen via multiple routes (Figure 1.4) (Gal-Mor, Boyle and Grassl, 2014). *Salmonella* can be transmitted person to person through ingestion of infected faeces. It can also be transmitted from pets such as dogs, cats, rodents, amphibians and reptiles.

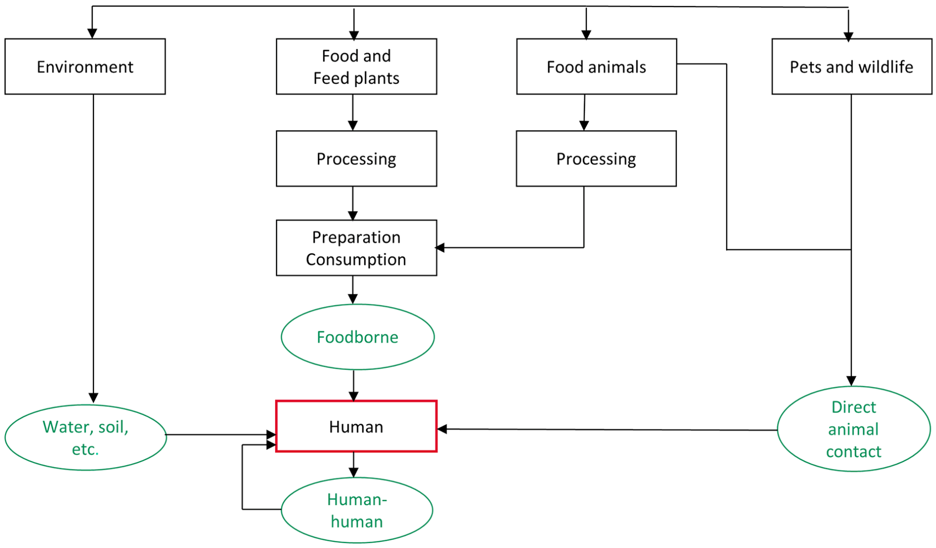


Figure 1.4 Transmission routes of human NTS infection
Reproduced from (Hald *et al.*, 2016)

Foodborne transmission is the most common route of human infection, estimated to cause 95% of NTS cases in the USA (Mead *et al.*, 1999). Poultry products are often the biggest source of human salmonellosis cases; between 1998-2003 in the USA, 48% of reported human salmonellosis cases were attributed to chicken and 6% to egg products (Guo *et al.*, 2011). Between 2007 and 2009 in EU member states the largest source of human salmonellosis was eggs, causing 42.4% of cases (De Knegt, Pires and Hald, 2015). In 2018 eggs were still the largest cause of salmonellosis outbreaks, associated with 45.6% of strong-evidence outbreaks in EU member states; broiler meat was associated with 2.4% (EFSA and ECDC, 2019a).

1.2.4 How *Salmonella* Causes Infection in Domestic Fowl

Salmonella infection and carriage in poultry is a major public health issue; although some serovars do cause disease in poultry, those that don't are able to colonise the caeca of poultry and reside there for weeks (Barrow *et al.*, 1987). As these birds don't show any symptoms it is challenging to identify those infected within flocks.

As seen in humans, the introduction of *Salmonella* to the chicken gastrointestinal tract provokes severe inflammation; this reduces invasion but does not provide enough of a response to remove the bacteria near the caecal lumen in the epithelial cells (Withanage *et al.*, 2005; Kogut and Arsenault, 2017). Caecal pro-inflammatory signals are down-regulated 4 days post *Salmonella* infection allowing persistence for several weeks (Kogut *et al.*, 2016).

Differences in the mode of infection are also seen between the serovars that are less invasive to poultry; for example, infection of chicks with *S.* Enteritidis and *S.* Typhimurium resulted in changes of cytokine expression; however *S.* Infantis infection did not change cytokine expression (Setta *et al.*, 2012). In poultry *S.* Pullorum and *S.* Gallinarum cause pullorum disease and fowl typhoid respectively (Shivaprasad, 2000). Both serovars cause systemic disease in poultry and avoid provoking a strong immune response to enable invasion of host tissues (Henderson, Bounous and Lee, 1999; Kaiser *et al.*, 2000; Shivaprasad, 2000). Manifestations of these infections include: anorexia, diarrhoea, droopy wings, depression, laboured breathing and dehydration. Both infections can be fatal in birds.

1.2.4.1 *Transmission of Salmonella in Poultry*

Salmonella colonisation in chickens is affected by the genetic susceptibility of the host, infectious dose and bird stress (Foley *et al.*, 2011). In chickens *Salmonella* can be spread either by horizontal or vertical transmission (Lamont, 2010). If chicks are infected immediately after hatching they can be colonised all the way through to maturity; this may occur because of the low responsiveness of the chicks towards *Salmonella* (Gast and Holt, 1998; Holt *et al.*, 1999). *Salmonella* is shed in the faeces of carrier birds and is released into the environment, enabling horizontal transmission. Vertical transmission occurs when the bacteria invade egg follicles and the ovaries; this is then transmitted into laid eggs (Haider, Chowdhury and Hossain, 2014).

Compared to humans, whose gastrointestinal pH ranges from 1 to 8, the pH of the chicken gastrointestinal tract is lower, ranging between 2.5 and 7.7, a factor that may affect *Salmonella* virulence (Denbow, 2000; Koziolek *et al.*, 2015). Additionally, there is a difference in core body temperature in humans and chickens, 37°C and 42°C respectively (Byrne, Clyne and Bourke, 2007). Consequently, a zoonotic *Salmonella* strain would have to be able to survive in both environments.

While newly hatched chicks are vulnerable to *Salmonella* infection, the mature microbiota of the chicken gastrointestinal system is known to give resistance to *Salmonella* infection using competitive exclusion mechanisms such as nutrient competition and occupation of mucosal attachment sites (Schneitz, 2005; Chambers and Gong, 2011). For example, *Salmonella* increase epithelial oxygenation with virulence factors; the commensal *Enterobacteriaceae* compete with *Salmonella* for this (Litvak *et al.*, 2019). Also, the chicken gut microbiota produces short-chain fatty acids such as acetate and butyrate which reduces the pH of the lumen, inhibiting microorganisms that are acid sensitive (Barua *et al.*, 2002; Czerwiński *et al.*, 2012).

1.2.4.2 *Poultry Industry*

Chickens are the main birds used in the poultry industry, others including ducks, geese and turkeys (Department for Environment, Food & Rural Affairs (DEFRA), 2019a). Chickens are split into three groups: broilers, hens laying eggs and the breeding flock. Broiler chickens are birds that have been reared specifically for the production of meat

(Kruchten, 2002). They have a lifespan of 6 to 7 weeks and are usually transported at least twice, as eggs and then again after hatching; this increases the likelihood of contamination occurring, either due to prior contamination of the transport coops or the stress that transport causes the birds (Mulder, 1995; Northcutt *et al.*, 2003; Mitchell and Kettlewell, 2009; Vinueza-Burgos *et al.*, 2016). Layers are reared for the production of eggs. They begin to lay eggs from the age of 18 weeks and the majority continue to do so until they are around 2 years old (Webster and Fletcher, 1996; Meunier and Latour, 2005).

As broiler chickens have a short lifespan, vaccination of them against *Salmonella* is ineffective due to their immune system not being fully developed (Desin, Köster and Potter, 2013). Vaccination of the breeding flock is therefore required as an alternative. The risks for *Salmonella* infection vary between broiler chickens and layer hens; in broiler chicken farms in Belgium, risks for *Salmonella* infection were found to be infection of other chicks in the flock and *Salmonella* infection in a previous flock (Namata *et al.*, 2009). The main risks of *S.* Enteritidis infection in layer hens from the Netherlands were flock size, the housing system used and a farm containing hens of various ages (Mollenhorst *et al.*, 2005).

Globally, in 2017 it is estimated that approximately 23 billion *Gallus gallus* were bred for agriculture; the largest producer of chickens was China, producing approximately 5 billion (Food and Agriculture Organization of the United Nations, 2019). In total, an estimated 109 million tonnes of chicken meat were produced globally, with the USA producing the most at 19 million tonnes. Furthermore, 1.4 trillion eggs were produced globally in 2017, with China producing 537 billion of these. In 2018 the total number of poultry birds in the UK was 188,442 million; 66% of these birds were broilers, also known as table chickens (DEFRA, 2019a). In 2014, approximately 1 billion broilers were slaughtered in the UK, a figure that has been increasing since the 1990s (DEFRA, 2019b). This illustrates the vastness of the global poultry industry.

1.2.4.3 *Salmonella Testing and Routes of Contamination*

In the UK it is compulsory for a flock of size greater than 2,000 birds to be tested for the presence of *Salmonella* within 3 weeks prior to the date of slaughter (DEFRA, 2016). If *S.* Enteritidis or *S.* Typhimurium is found to be present in a UK flock the holding must be disinfected (DEFRA, 2016). This represents a major cost for farmers. Before slaughter the chickens are starved of food for up to 12 hours (DEFRA, 2018). They are then caught, transported to the abattoir where they are held until slaughter; the transportation and holding increases the likelihood of *Salmonella* contamination from external sources (Buncic and Sofos, 2012). Once the bird housing is empty it is recommended that it is dry cleaned, washed and then disinfected before the next flock is brought in (DEFRA, 2018). Chickens are required to have enough space to stand, turn and stretch their wings. This allows transmission of pathogens between flocks and their close proximity to one another increases the risk of horizontal transmission.

The EU has targets for the levels of *Salmonella* in *Gallus gallus* in each member state (EFSA and ECDC, 2019a). For breeders, less than 1% of flocks should be positive for *S.* Enteritidis, *S.* Typhimurium including monophasic *S.* Typhimurium, *S.* Virchow*, S.* Infantis, and *S.* Hadar (European Commission, 2011). The targets change for birds bred for meat and egg production. Less than 1% of broiler flocks and 2% of laying hens should be positive for *S.* Enteritidis and *S.* Typhimurium, including the monophasic variant (EFSA and ECDC, 2019a; European Commission, 2012). However, these are only targets and in 2018 just 16 of the 27 reporting member states met all the poultry *Salmonella* targets.

Research has been performed to identify whether there are differences in the prevalence of *Salmonella* depending on the type of farm; however, results from different studies are contradictory with one study reporting higher levels of *Salmonella* in conventional broiler farms and another in organically reared poultry (Cui *et al.*, 2005; Alali *et al.*, 2010). Once *Salmonella* has been detected in a poultry house, its removal is challenging; it can be found in litter dust after the house has been cleansed; wild birds and rodents surrounding poultry houses have also been found positive for carrying *Salmonella* (Tizard, 2004; Davies and Wray, 1996). Multiple sources of *Salmonella* contamination exist in broiler farms; a study of 65 in Spain found that 32% of delivery box liners were positive for *Salmonella*, 25% of dust samples, 20% of farming boot samples and 16% of feed from feeders (Marin *et al.*, 2011).

1.2.4.4 *Chicken Vaccination Programme*

S. Gallinarum and *S.* Pullorum cause clinical disease in poultry and were the dominant serovars in the UK poultry industry until the 1970's, upon which slaughtering and vaccination reduced their numbers (O'Brien, 2013). This left an ecological niche which was filled by *S.* Enteritidis, associated with high levels of this serovar causing human salmonellosis; in 1993 *S.* Enteritidis phage type 4 levels peaked in the UK, causing over 18,000 reported cases. A *Salmonella* vaccine was introduced in the poultry industry and a reduction in human *Salmonella* cases has been seen since 1997.

Vaccination of layer hens against *S.* Enteritidis and *S.* Typhimurium is implemented in EU member states with a *Salmonella* prevalence greater than 10% (European Commission, 2006). Vaccination against *S.* Enteritidis and *S.* Typhimurium induces immunity against *Salmonella* belonging to groups 0:9 (D_1) and 0:4 (B) respectively (Miller *et al.*, 2010). As seen post *S.* Gallinarum and *S.* Pullorum vaccination, this leaves an ecological niche which *Salmonella* containing other somatic antigens can fill, such as group 0:7 (C_1) containing serovars like *S.* Infantis.

1.2.4.5 *Antimicrobial Usage*

In the poultry industry, antibiotics are used to promote growth and prevent or treat disease (Roth *et al.*, 2018). Whilst their use as growth promoters is prohibited in Europe and the USA, they are still being used for this application in countries like China, although China has plans to reduce the use of antimicrobials in animals by 2021 (Qu, Huang and Lv, 2019).

The antibiotics approved for use varies by country, for example, colistin is approved for use in the USA, UK, Brazil, China, Poland, Germany, France and Spain whereas fosfomycin is only approved in Brazil (Roth *et al.*, 2018). In Denmark the antibiotics used in the poultry industry are aminoglycosides, sulphonamides, macrolides, penicillin's, trimethoprim and tetracyclines, the antibiotic used most with broilers (Borck Høg *et al.*, 2018). The use of antimicrobials in the poultry industry promotes the emergence and spread of AMR in *Salmonella* (Su *et al.*, 2004).

1.2.5 Host adaptation and Genetic Diversity in *Salmonella*

The host range of *Salmonella* can be very broad; infecting insects, fish, reptiles, amphibians, birds and mammals (Briones *et al.*, 2004; Millan *et al.*, 2004; Nakadai *et al.*, 2005; Musto *et al.*, 2006; Wales *et al.*, 2010) . It can also survive in plants, soil, water and protozoa (Thomason, Biddle and Cherry, 1975; Gaze *et al.*, 2003; Barak and Liang, 2008).

S. enterica subsp. *enterica* serovars can be split into 3 groups due to the range of hosts they infect: host generalist, host-adapted and host-restricted (Sanderson and Nair, 2013). Serovars that are generalists are capable of infecting a range of hosts. These serovars, such as *S.* Enteritidis, usually cause gastrointestinal disease but are also capable of causing infections in humans ranging from causing no symptoms to invasive infection (Crump and Wain, 2017).

Host-adapted serovars can be found in a smaller number of hosts but are mainly associated with one host, in which they cause severe infection; for example *S.* Dublin and *S.* Choleraesuis can both infect humans, but cause systemic infection in bovine and swine animals respectively (Sanderson and Nair, 2013). They are also able to persist in the host they are associated with by direct transmission (Kingsley and Bäumler, 2000).

Host-restricted serovars only cause infection in one host and are usually associated with severe systemic infections (Sanderson and Nair, 2013). Examples of host restricted serovars include *S.* Gallinarum which is restricted to galliformes and causes fowl typhoid; *S.* Typhi and *S.* Paratyphi which causes typhoid and paratyphoid fever respectively in humans and *S.* Abortusovis which triggers abortions in sheep.

Host adaptation initially occurs due to the occurrence of variation in the genome during infection, including genomic rearrangement, deletions, acquisition of genes through horizontal gene transfer and point mutations (Tanner and Kingsley, 2018). If these mutations are beneficial to survival, then they are maintained and transmitted. If these mutations enable colonisation of a novel niche, such as a new host, a novel site within the same host or survival in the presence of an antibiotic; then this mutation may become fixed in the population.

Plasmids can be associated with host adaptation; the *Salmonell*a plasmid virulence (SPV) gene cluster, *spv*, which contains several virulence determinants, has been found in the following *S. enterica* subsp. *enterica* serovars: Abortusequis, Abortusovis, Choleraesuis, Dublin, Enteritidis, Gallinarum, Paratyphi C, Sendai, and Typhimurium (Silva, Calva and Maloy, 2014). The majority of these serovars are host-adapted, the only host generalist serovars found with the plasmid are *S.* Enteritidis and *S.* Typhimurium. The *spv* genes play a key role in host adaptation in *S.* Dublin; they were found to be essential for *S.* Dublin to cause severe diarrhoea and systemic infection in cattle (Libby *et al.*, 1997).

1.2.5.1 *Virulence Factors*

Salmonella have acquired many virulence factors, such as a capsule, flagella, fimbriae and adhesion systems; enabling them to colonise a wide variety of hosts and evade the host immune response (Jajere, 2019). With the notable exception of *S.* Gallinarum and *S.* Pullorum, *Salmonella* possess 5 to 10 peritrichous flagella over their surface, conferring motility (Asten and Dijk, 2005).

SPIs are large chromosomal gene cassettes containing virulence genes needed for invasion and extraintestinal spread (Marcus *et al.*, 2000; Thornbrough and Worley, 2012). They can either be found throughout the *Salmonella* genus or in specific serovars, for example SPI-1 has been identified across the genus whereas SPI-2 has been identified in all *S. enterica* subspecies but not in *S. bongori* (Ochman and Groisman, 1996). SPI-1 is needed for host cell internalisation and SPI-2 for intracellular growth (Jennings, Thurston and Holden, 2017). Other examples of SPI's include SPI-6, which encodes an antibacterial amidase that induces lysis of bacteria and SPI-3 which contains *misL*, which is needed for long-term colonisation (Ilyas, Tsai and Coombes, 2017). The evolution of *Salmonella* pathogenicity is strongly linked to acquisition of SPIs through horizontal gene transfer (Langridge *et al.*, 2015). Differential expression of SPIs between serovars and evolution of genes within the pathogenicity islands has also occurred, which may be linked with host specificity (Imre et al., 2013; Eswarappa *et al.*, 2008).

Several SPIs have been discovered; their presence varies between serovars (Crump and Wain, 2017). For example, SPI-7 has only been identified in *S*. Typhi, *S*. Paratyphi C and some *S*. Dublin strains (Seth-Smith *et al.*, 2012). It contains approximately 150 genes including those encoding the surface Vi polysaccharide antigen, a type IVb pili and the *sopE* virulence factor (Nieto *et al.*, 2016). SPI-7 presence is thought to not be essential for human epithelial cell invasion but required for systemic infection.

SPI-13, whilst present in many serovars of *S. enterica* subsp. *enterica*, is largely absent in *S.* Typhi, which has SPI-8 in its place (Espinoza *et al.*, 2017). SPI-13 is important for *Salmonella* virulence as it plays a role in the nutritional fitness of the pathogen (Elder et al., 2018). It also was found to be needed for pathogen internalisation in murine macrophages but not human macrophages; conversely SPI-8 was not needed for macrophage internalisation in either human or murine macrophages, it's maintenance within the typhoidal serovars suggests involvement in another phase of the human infection process.

Another example, SPI-10, encodes the *sefA-R* chaperone-usher fimbrial operon and contains the *prpZ* gene cluster which is carried on a prophage and is involved in *S*. Typhi survival in macrophages (Faucher *et al.*, 2008; Liaquat *et al.*, 2018). Whilst the *prpZ* gene cluster is thought to be restricted to *S.* Typhi (Liaquat *et al.*, 2018), SPI-10 has been identified in *S*. Dublin, *S*. Enteritidis, *S*. Gallinarum and *S*. Paratyphi (Saroj *et al.*, 2008).

The number of virulence genes present vary between and within serovars, a study comparing selected virulence gene presence across 15 *S. enterica* subsp. *enterica* serovars found that whilst some genes were present in either 0% or 100% of strains from a serovar, several genes were present but not maintained in all strains (Cheng *et al.*, 2015). Virulence gene presence also varied significantly between serovars. As the presence of virulence genes can alter the organism's ability to cause infection, understanding the population structure of *S. enterica* and how virulence factors differ between serovars is important.

1.2.5.1.1 Genome-wide screens

Alongside the well described SPIs, other virulence factors have been identified as important in *Salmonella* survival and pathogenesis. Several different methods have been developed to identify the importance of genes in a *Salmonella* genome for survival in different environments.

Signature tagged mutagenesis (STM) is polymerase chain reaction (PCR) based technique where transposons containing unique tags are used to create a pool of mutants (Andrews-Polymenis, Santiviago and McClelland, 2009). The mutant pool is then exposed to a selective condition to identify genes that are essential for survival. Examples of virulence factors whose roles have been described using STM include *envZ*, which was found to play a role in *S*. Gallinarum infection of chickens (Shah *et al*., 2005) and fimbrial operons such as *stbC* and *sthB* which, when mutated reduced *S*. Typhimurium chick colonisation but not calf colonisation (Morgan *et al.,* 2004). Also, *manC* in *S*. Dublin isolates, when attenuated, reduced the amount of lipopolysaccharide in the outer membrane, resulting in decreased virulence and stress tolerance (Thomsen *et al*., 2003).

Transposon site hybridisation (TraSH) and transposon-mediated differential hybridisation (TMDH) are similar genomic screening methods that use microarrays to enable the analysis of greater numbers of insertions than STM. They involve transposon mutagenesis, identification of the location of the transposons in the genome;

amplification of the neighbouring regions to the transposon-insertions sites and transcription of these regions (Sassetti, Boyd and Rubin, 2001; Chaudhuri, Allen *et al.,* 2009). To allow for quantification, DNA is acquired from organisms grown in selective conditions, labelled with fluorophores and hybridised to a DNA microarray in TraSH; in TMDH, transcription is induced and the RNA is hybridised to an RNA microarray.

The use of TMDH with *S*. Typhimurium in a mouse model identified numerous genes associated with virulence including several present on SPIs or involved in aromatic amino acid or purine biosynthesis (Chaudhuri, Peters *et al.*, 2009). Others that had not previously been associated with virulence were also found such as *tolA* and *tolB*, which are involved with membrane stability and *ychK* which codes for a lipolytic enzyme. Another mouse experiment assessing the competitive fitness of *S*. Typhimurium TraSH mutants against *Enterobacter cloacae*, a commensal which prevents *Salmonella* colonisation, identified genes which reduced fitness when mutated (Ali *et al*., 2014). Examples include *sirA*, which is involved in regulating the mRNA stability of numerous virulence and metabolic genes and the *fra* locus, which was associated with a severe fitness defect when mutated and is needed for utilisation of fructose-asparagine.

Transposon directed insertion-site sequencing (TraDIS) involves the random insertion of transposons into millions of cells; utilising next generation sequencing to allow for the investigation of considerably more insertions than STM and TraSH/TMDH (Langridge et al., 2009). The mutants are then challenged and the genes needed for survival identified. Using TraDIS, the genes *hupA* and *pagP* were found to be essential for bile tolerance in *S*. Typhi, a trait that is important in *S*. Typhi carriage in the gallbladder. TraDIS has also been used to identify genes in *S.* Typhimurium that were key for causing infection in chickens, cattle and pigs as well as genes essential for infection in just one host (Chaudhuri *et al*., 2013). For example, *clpB*, *clpP* and *clpX*, involved in *rpoS* regulation which contributes to resistance to environmental stresses, were needed for survival in chickens but not cattle or pigs.

Tissue specific virulence factors have also been described using TraDIS. *S*. Typhimurium mutants in mesenteric lymph nodes from the distal ileum and the ileal wall of calves were compared (Vohra *et al.*, 2019). Attenuation in both tissues was associated with insertions in 653 genes, indicating a conserved role in pathogenesis. Conversely, 30 genes were identified that reduced fitness in mutants recovered from the ileal wall and 2 genes which reduced fitness in mesenteric lymph nodes; suggesting that these genes are niche-specific virulence factors. TraDIS has also been used to compare essential genes for different *Salmonella* serovars; for example, in a comparison of *S*. Typhi and *S*. Typhimurium, 281 genes were required by both for competitive growth in a rich media (Barquist *et al.*, 2013). Serovar-specific gene requirements were also identified; 29 genes were only required by *S*. Typhi and 56 by *S*. Typhimurium including *SifB*, a virulence effector protein, which was present in both serovars but was only essential for *S*. Typhimurium.

This illustrates that, whilst the SPIs play essential roles in *Salmonella* pathogenesis, other virulence factors, which often vary between serovars, are also important in *Salmonella* survival, pathogenesis and host specificity.

1.2.5.2 *Population Structure of S. enterica subsp. enterica*

In 2011, a comparison of MLST results of 4,257 *S. enterica* subsp*. enterica* genomes from 554 serovars identified 1092 STs, which clustered in 138 eBGs (Achtman *et al.*, 2012). The population structure of these isolates is shown in Figure 1.5. Whilst the majority of the *S.* Typhimurium and *S.* Enteritidis sequences belonged to eBG 1 and 4 respectively; most of the other serovars did not cluster into one eBG and were present in several distinct eBGs. Additionally, the majority of the serovars were found in multiple eBGs and were therefore polyphyletic. Of the 42 serovars where 15 or more sequences were included, 25 of the serovars were located in multiple eBGs or STs and 17 were present in a single eBG. For example, whilst the majority of isolates in eBG1 were *S.* Typhimurium, monophasic *S.* Typhimurium variants and serovars Farsta and Hato were also present in that eBG.

The diversity of the strains within a serovar varies considerably; *S.* Typhimurium was also found in another eBG and ST which both contained monophasic variants or other serovars (Achtman *et al.*, 2012). Conversely all 50 *S.* Typhi sequences belonged to eBG13 and no other serovars were found in that eBG. By 2017 the number of defined STs had increased markedly; 118,391 *Salmonella* genomes on Enterobase divided into 3,929 STs, which resided in 360 eBGs (Alikhan *et al.*, 2018).

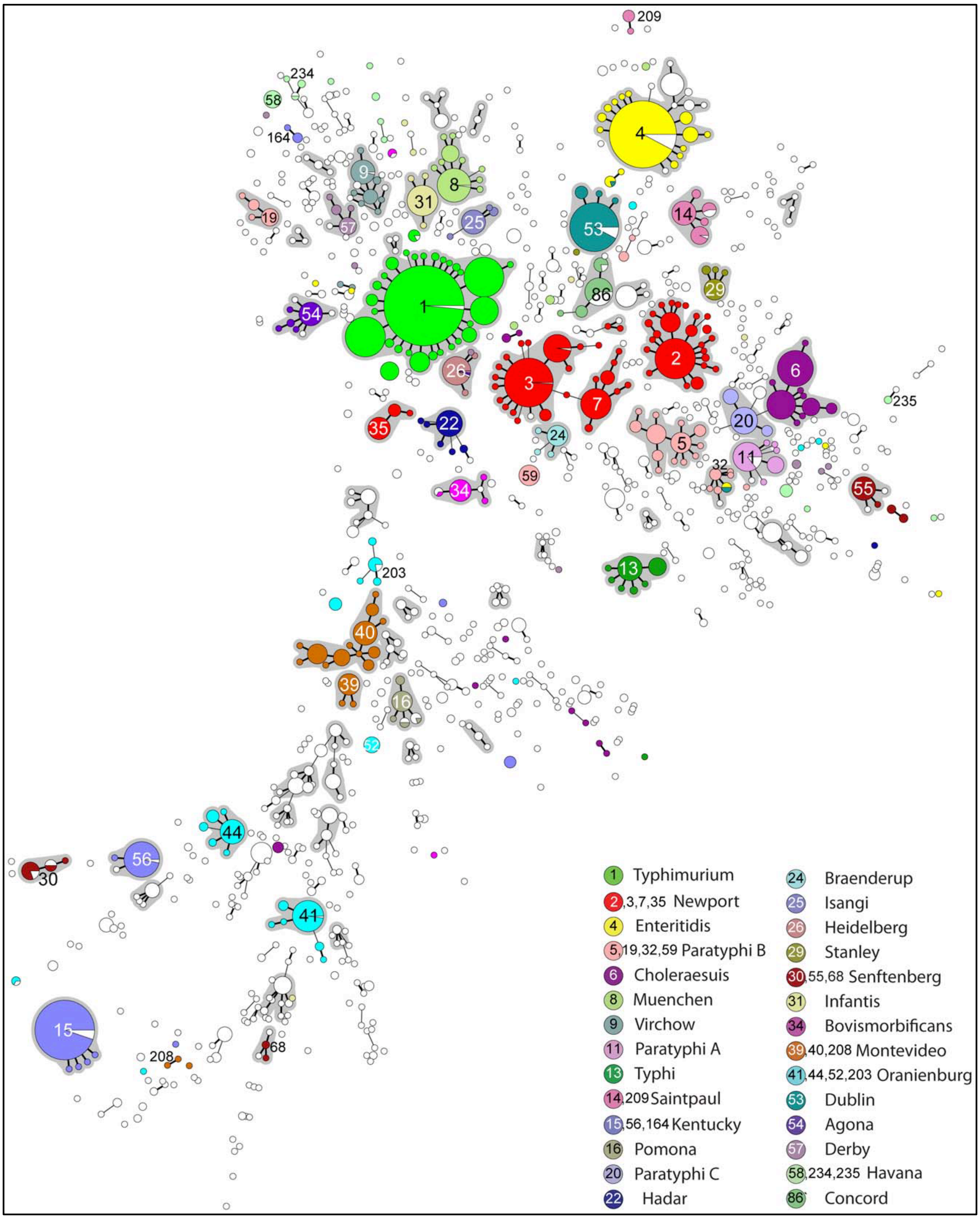


Figure 1.5 Population structure of *S. enterica* subsp. *enterica*

Minimum spanning tree of MLST data of *S. enterica* subsp. *enterica* (n=4257). Each ST is represented by a circle of a size proportional to the number of isolates it contains. A thick black line indicates an SLV and a thin black line a double-locus variant (DLV); the eBGs are designated by grey shading. If the majority of the sequences in the ST or eBG belong to one of the 28 most numerous serovars it is colour coded. Reproduced from Achtman *et al.*, 2012.

1.2.5.3 *Pan and core-genome of Salmonella*

The pan genome is defined as all the genes present in a collection of genomes; the core genome is the genes shared by 99% of the collection and the accessory genome is all the genes in the pan genome that are not core genes (Page *et al.*, 2015). Multiple studies have identified the size of the pan-genome of *Salmonella*, *S. enterica* and individual serovars. A pan-genome analysis of 35 *Salmonella* strains identified 10,015 gene families, genes grouped due to sequence similarity, with a core genome of 2,811 gene families (Jacobsen and Hendriksen, 2011). When excluding the *S. bongori* strains*,* the pan-genome size decreased to 9,161 gene families and the core genome increased to 3,224 gene families. Gene families were identified that were unique to a serovar. Interestingly, whilst *S. bongori* had 315 unique gene families, the more recently diverged *S. enterica* subsp. *arizonae* had 504. Variation was also seen within *S. enterica* subsp. *enterica*; 10 of the serovars had fewer than 100 unique gene families, including *S.* Enteritidis (29), *S.* Typhimurium (43) and *S.* Dublin (71). Conversely, 11 of the serovars had 100 or greater unique gene families including *S.* Gallinarum (135), *S.* Newport (162) and *S.* Typhi (349).

Another study comparing 72 *S. enterica* subsp. *enterica* isolates and an *S. enterica* subsp. *arizonae* isolate identified fewer core genes (2,882), despite the exclusion of other *S. enterica* subspecies, showing how the addition of more isolates decreases the number of core genes (Leekitcharoenphon *et al.*, 2012). The variation within the core genes was calculated; the majority of the genes were conserved but approximately 5% of them were highly variable.

A larger study described the pan-genome of 4,893 *S. enterica* isolates, including isolates from all six subspecies (Laing, Whiteside and Gannon, 2017). Assuming an average gene size of 1,000bp, approximately 1,500 genes were core and the pan-genome contained 25,300 genes. While genomic regions to unique each subspecies were identified; regions unique to a serovar were not, although some had significant associations with a serovar. 404 1,000bp regions were identified as being specific to *S. enterica*; *S.* Enteritidis had the largest number of these regions and *S.* Typhi the least, suggesting that *S.* Enteritidis is the most archetypal *S. enterica* genome and *S.* Typhi the most atypical.

The core genome size varies between serovars; in a comparison of 21 *S.* Dublin, 32 *S.* Newport and 37 *S.* Typhimurium isolates from humans and cattle in the USA, the overall core genome size was 3637 genes (Liao *et al.*, 2019). When the serovars were looked at individually the *S.* Typhimurium isolates had the lowest number of core genes at 4,003 and the *S.* Dublin isolates had the highest at 4,326; although the differences in size could again be due to differences in the number of sequences included. The overall pan-genome size was 7,077, ranging in size within serovars from 5,066 in *S.* Dublin to 6,433 in *S.* Typhimurium. The *S.* Typhimurium isolates were found to have increased diversity in their gene composition and the *S.* Dublin isolates had the largest number of significantly over-represented genes.

Understanding the pan genome, and how it varies within and between serovars, is important as it can provide genetic markers for serovars and sub-groups of serovars, associated with traits such as increased virulence.

1.2.5.4 *Intergenic Diversity of Salmonella*

Whilst many studies have explored the variation in genes in *Salmonella*, very little research has been done on non-coding regions. These intergenic regions (IGRs) typically comprise 10-15% of the bacterial genome and contain non-coding ribonucleic acids (ncRNAs), promoters, regulatory binding sites and terminators (Thorpe *et al.*, 2017, 2018). Switching of regulatory regions to non-homologous alternatives has been found to occur throughout the bacterial domain, with horizontal regulatory transfer observed between species and genera (Oren *et al.*, 2014). The rate of regulatory switching varies, with 10-fold higher levels occurring in *E. coli* than in *S. enterica*. Changes to the regulatory regions can have substantial phenotypic effects, for example, a promoter inversion in *Photorhabdus luminescens* changes it from a pathogen to a commensal (Somvanshi *et al.*, 2012).

IGRs have been found to play a role in stress tolerance. For example, 61 genes and 6 IGRs are essential for *S.* Typhimurium to withstand desiccation stress (Mandal and Kwon, 2017). The 6 IGRs varied in length from 132bp to 791bp; no known small ncRNAs were identified in these IGRs and a coding region was found in only one of them for a hypothetical protein.

When examining 68 *S.* Typhimurium strains, 3,846 core genes and 1,576 core IGRs were identified; a further 281 IGRs were present in different numbers in the strains (Fu *et al.*, 2015). The IGRs ranged in size from 101bp to 12,151bp. The location of SNPs within 21 of the genomes were determined and an average of 21% of the SNPs were present on IGRs; similar substitution rates were seen between core IGRs and core genes.

Small ncRNAs in bacteria are typically located in IGRs and are involved in regulating gene expression (Raghavan *et al.*, 2015). Several small ncRNAs have been identified in *Salmonella*, located in SPIs; one of these is *IsrM*, which contributes to *Salmonella* pathogenesis (Gong *et al.*, 2011). *IsrM* regulates *HilE* and *SopA* protein expression and is involved in murine colonisation, epithelial cell invasion and replication within macrophages. It has been found in several serovars including *S.* Typhimurium, *S.* Heidelberg and *S.* Saintpaul and is absent from *S.* Typhi and *S. bongori* which could indicate that it plays a role in host specificity. As with the pan genome, defining IGRs shared or unique to serovars or sub-groups of serovars, could identify markers associated with traits of interest.

1.2.6 Antimicrobial Resistance in *Salmonella*

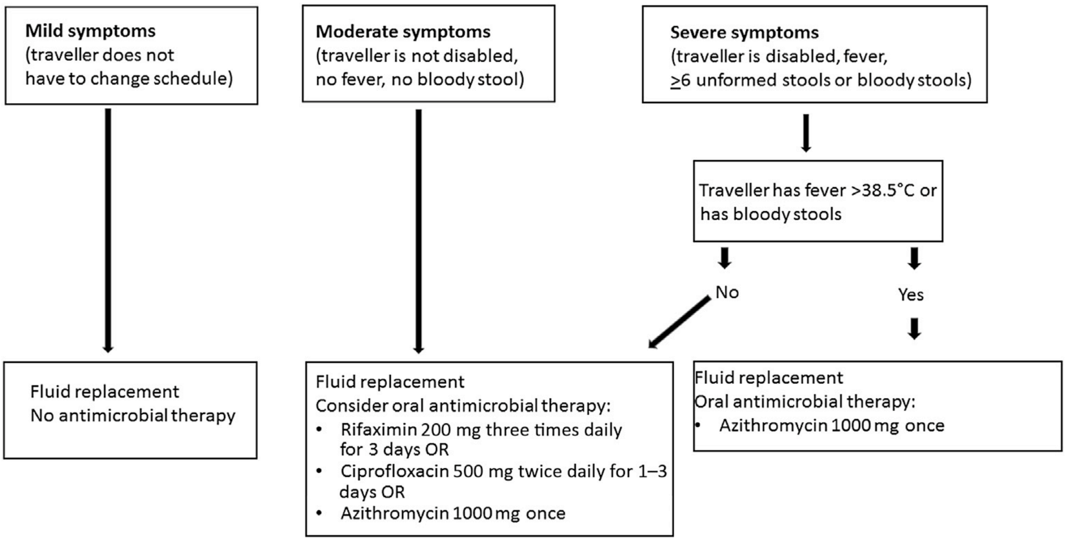
AMR in NTS is a public health concern; in 2016 the WHO added fluoroquinolone resistant *Salmonella* to the high-priority tier of antibiotic resistant bacteria requiring research and drug development (Tacconelli *et al.*, 2018). In 2002 it was estimated that in the USA, 29,379 extra cases of salmonellosis in humans occur annually due to AMR in NTS, resulting in 342 hospitalisations and 12 fatalities (Barza and Travers, 2002).

1.2.6.1 *AMR and the Treatment of Salmonellosis*

Human salmonellosis is an important economic burden in high income countries (McEntire *et al.*, 2014) where the majority of cases result in gastrointestinal symptoms, and are primarily managed with oral rehydration therapy. However, cases can develop into invasive disease (Cuypers *et al.*, 2018) which presents as febrile illness (Gordon, 2011; Feasey *et al.*, 2012), and requires specific antibiotic therapy (Figure 1.6) (Colobatiu *et al.*, 2015). Immunocompromised individuals are at a greater risk of developing iNTS; high endemic levels of HIV lead to the increased use of antimicrobials which, in turn, leads to an increase in AMR (Essack *et al.*, 2017).

Antibiotics commonly used include: chloramphenicol, ciprofloxacin, trimethoprim-sulfamethoxazole, the third-generation cephalosporin ceftriaxone and penicillins such as amoxicillin and ampicillin (Su *et al.*, 2004; Chen *et al.*, 2013; Franco *et al.*, 2015;

Hindermann *et al.*, 2017; Kongsoi, Nakajima and Suzuki, 2017; Medalla *et al.*, 2017; Brown *et al.*, 2018). AMR to these drugs is a public health concern, particularly ceftriaxone, as it is used in the treatment of severe salmonellosis (Jajere, 2019).



Mild symptoms
(traveller does not have to change schedule)
Moderate symptoms
(traveller is not disabled, no fever, no bloody stool)
Severe symptoms
(traveller is disabled, fever, ≥6 unformed stools or bloody stools)
Traveller has fever >38.5°C or has bloody stools
No
Yes
Fluid replacement
No antimicrobial therapy
Fluid replacement
Consider oral antimicrobial therapy:
• Rifaximin 200 mg three times daily for 3 days OR
• Ciprofloxacin 500 mg twice daily for 1–3 days OR
• Azithromycin 1000 mg once
Fluid replacement
Oral antimicrobial therapy:
• Azithromycin 1000 mg once

Figure 1.6 Treatment pathway for the management of NTS infection
Reproduced from Zollner-Schwetz and Krause, 2015.

1.2.6.2 *AMR levels in Salmonella*

Levels of AMR in *Salmonella* vary by serovar, geography, and source. In Australia, between 1975 and 2015, levels of AMR in human NTS were low, 83% of isolates were sensitive to all the antimicrobials tested (Williamson *et al.*, 2018). The levels of AMR varied across the top 10 serovars, with 78% of 142 *S.* Panama isolates being resistant to one or more antimicrobial compared to 1% of 2,205 *S.* Mississippi isolates.

High levels of AMR in *Salmonella* have been observed elsewhere across the planet. A study of AMR in food-producing livestock, including chickens, pigs and ducks, in the Shandong province of China between 2009 and 2012 found that over 99% of isolates had resistance to at least one antibiotic, with the number significantly increasing between 2009 and 2012 (Lai *et al.*, 2014). In Ghana, 60.6% of *Salmonella* strains isolated from poultry farms were resistant to at least one antimicrobial, as were 60% of *Salmonella* strains isolated from broiler chicken farms in Brazil (Voss-Rech *et al.*, 2015; Andoh *et al.*, 2016).

The levels of AMR also vary between *Salmonella* isolated from humans and poultry. The trend in AMR in *Salmonella* from human and chicken samples in the USA is shown in Figure 1.7.

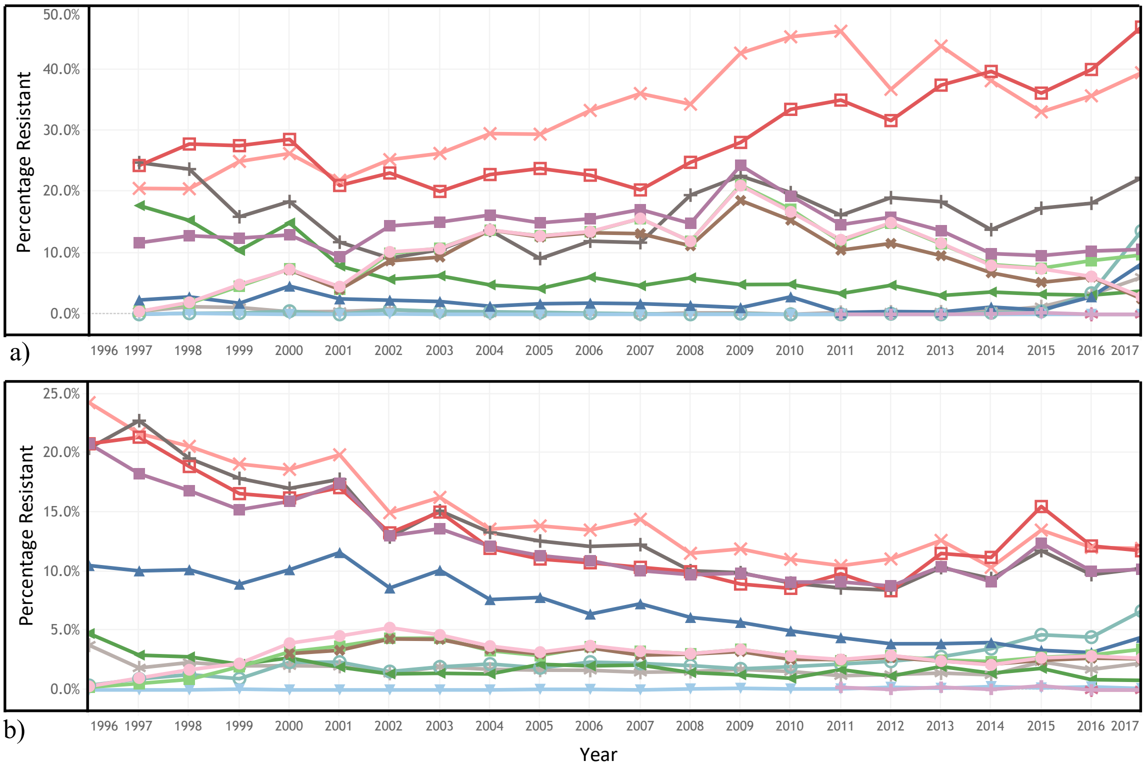


Figure 1.7 The trend of AMR in *Salmonella* in the USA

a) Percentage of isolates from retail chicken, caecal samples at slaughter and samples taken at slaughter with AMR
b) Percentage of isolates causing clinical illness in humans with AMR

Amoxicillin-Clavulanic Acid ● Ampicillin ■ Azithromycin + Cefoxitin ✖ Ceftriaxone ■
Chloramphenicol ▲ Ciprofloxacin ▼ Gentamicin ◄ Meropenem * Nalidixic Acid ○
Streptomycin □ Sulfamethoxazole-Sulfisoxazole + Tetracycline × Trimethoprim-Sulfamethoxazole *
Generated using data from Food and Drug Administration (FDA), 2019b.

A difference was seen between the two isolation sources, with AMR in the chicken isolates gradually increasing over time and the *Salmonella* isolates from humans decreasing in AMR levels between 1996 and 2012; a slight increase in AMR in the human isolates is seen following 2012 for the following antibiotics: ampicillin, chloramphenicol, ciprofloxacin, nalidixic acid, streptomycin, sulphonamides and trimethoprim-sulfamethoxazole. In the *Salmonella* isolates from chickens, an increase in the proportion of isolates with AMR between 2015 and 2017 was seen for all of the 14 antimicrobials tested, except for amoxicillin-clavulanic acid, azithromycin, cefoxitin and ciprofloxacin.

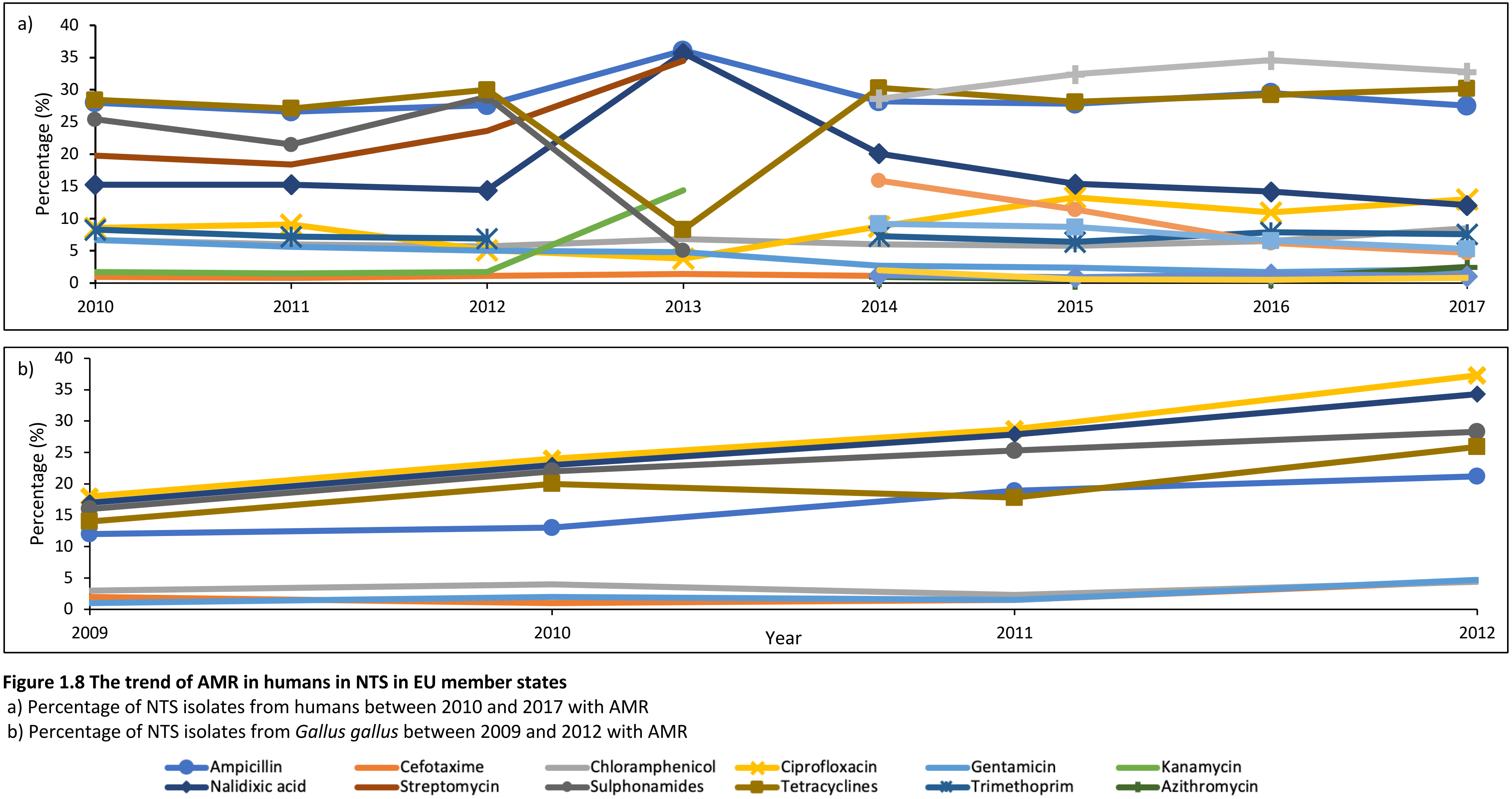


Figure 1.8 The trend of AMR in humans in NTS in EU member states

a) Percentage of NTS isolates from humans between 2010 and 2017 with AMR

b) Percentage of NTS isolates from *Gallus gallus* between 2009 and 2012 with AMR

Generated using data from EFSA and ECDC, 2011, 2012, 2013a, 2014a, 2015a, 2016a, 2017a, 2018a, 2019b

The proportion of NTS isolates with AMR from humans and *Gallus gallus* in EU member states that collect this data is shown in Figure 1.8. As with the isolates from the USA, there is a general increase in AMR in *Salmonella*, with the occurrence of resistance to azithromycin, cefotaxime, ceftazidime, chloramphenicol, ciprofloxacin, colistin, co-trimoxazole, sulfamethoxazole, tetracycline and tigecycline in NTS from humans increasing between 2010 and 2017. An increase in AMR was also identified in NTS isolated from domestic fowl between 2009 and 2012 across the EU member states that collect this data, all of the antimicrobials increased between this time period. High levels of AMR were also seen in NTS from broiler meat in EU member states, in 2014 42.6% of the samples tested were positive for ciprofloxacin resistance, 39.7% for nalidixic acid resistance and 27% for sulfamethoxazole resistance (EFSA and ECDC, 2016a).

Differences are also seen between AMR in isolates from humans and cattle; a study comparing AMR in the top 20 serovars from Northwest USA between 2004 and 2011 found a larger number of AMR profiles in the human *Salmonella* isolates than in those from cattle; which the authors hypothesised was due to either continuous evolution in the human samples or an increased diversity of sources associated with causing the human infections (Afema, Mather and Sischo, 2015).

1.2.6.3 *Antimicrobial Resistance Determinants*

Many different AMR genes have been identified in *Salmonella* chromosomes or extra-chromosomal DNA, conferring resistance to different classes of antimicrobials (Su *et al.*, 2004). For example, resistance to extended-spectrum beta-lactams, a newer group of antimicrobials associated with higher mortality rates, is a public health concern (Su *et al.*, 2004; Dhillon, R, H and Clark, 2012). *Salmonella* have acquired resistance to these antimicrobials with the production of extended-spectrum beta-lactamases (ESBLs) which, as with other AMR genes, are present on plasmids, integrons and transposons (Su *et al.*, 2004).

Another AMR method *Salmonella* have evolved is the resistance to fluoroquinolones through the acquisition of mutations in DNA gyrase genes (DNA gyrase encoded by *gyrA, gyrB* and topoisomerase IV by *parC, parE*) in the quinolone resistance determining regions (QRDRs) (Su *et al.*, 2004). For example, amino acid substitutions in *gyrA* at Ser-83 or Asp-87 are often seen in isolates resistant to nalidixic acid (Michael *et al.*, 2006). Analysis of 283 *Salmonella* strains from food, humans and animals between

1970 to 2009 in Malaysia found nalidixic acid resistance in 30.7% of strains (Thong *et al.*, 2015). 69.3% of the resistant isolates contained a missense mutation in the *gyrA* QRDR; silent mutations were identified in *gyrB*, *parC* and *parE*.

Also present in *S. enterica* genomes are cryptic resistance genes, genes that appear to confer AMR but are present in phenotypically antimicrobial sensitive isolates (Salipante, Barlow and Hall, 2003). For example, the *aac(6')-Iaa* gene is commonly present in *S. enterica* and has been identified as a cryptic gene that has no phenotypic effect (Salipante, Barlow and Hall, 2003; Vetting, M *et al.*, 2004).

An alternative mechanism that can be used by *Salmonella* to enable survival in the presence of antimicrobials is the formation of persisters upon *Salmonella* macrophage internalisation (Helaine *et al*., 2014). Persisters are a population of non-replicating cells that form due to nutrient deprivation and vacuolar acidification and are tolerant to antibiotics. *Salmonella* persisters have been observed resuming intracellular growth which could result in a relapse of infection after antibiotic treatment.

1.2.6.4 *Multidrug Resistance*

Multidrug resistance (MDR) is defined as resistance to three or more classes of antibiotic. High levels of MDR in *Salmonella* is a public health concern as MDR strains are associated with increased severity of infection (Eng *et al.*, 2015); a study of NTS in Kenya identified an association with MDR NTS causing bacteraemia, even in patients that weren't immunosuppressed (Akullian *et al.*, 2018).

The levels of MDR fluctuate over time: in EU member states in 2013, MDR in *Salmonella* from humans was 31.8%, 26.0% in 2014, 29.3% in 2015, 26.5% in 2016 and 28.6% in 2017 (EFSA and ECDC, 2015a, 2016a, 2017a, 2018a, 2019b). They also vary by geographical location. Figure 1.9 shows the proportion of isolates from humans from 14 EU member states that were MDR. 52.6% of the isolates were susceptible to all 9 of the antimicrobial classes and 28.6% of the isolates were MDR, with Greece containing the highest proportion of MDR isolates at 81.4% (EFSA and ECDC, 2019b).

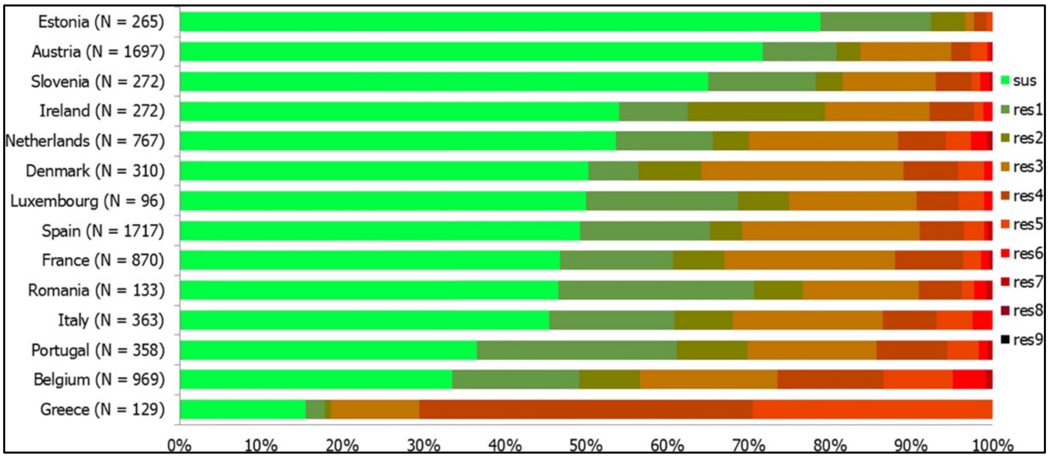


Figure 1.9 Proportion of European human *Salmonella* isolates that had MDR in 2017

Distribution of *Salmonella* isolates from 14 EU member states that were resistant to up to 9 antimicrobial classes.
sus: susceptible to all of the nine antimicrobial classes commonly tested across the EU
res1-9: resistant to 1 to 9 of the antimicrobial classes
Reproduced from EFSA and ECDC, 2019

Differences in MDR are also seen between serovars, with some serovars associated with increased incidence; in the USA in 2017, 25% of human *Salmonella* infections with MDR were caused by monophasic *S.* Typhimurium (Food and Drug Administration (FDA), 2019a). In 2016 in EU member states 76.3% of S. Kentucky isolates from humans had MDR compared to 40% of *S.* Typhimurium and *S.* Infantis isolates (EFSA and ECDC, 2018a). Comparatively low levels of MDR were seen in *S.* Enteritidis isolates from humans, with an average of 1.6% of isolates from each member state having MDR. In broilers the levels of MDR also varied by serovar, from 2.3% in *S.* Enteritidis, 62.5% in both *S.* Typhimurium and monophasic *S.* Typhimurium and 78.4% in both *S.* Infantis and *S.* Kentucky.

Furthermore, the levels of MDR vary by isolation source; in 2013 MDR ranged from 8.1% to 70.8% in isolates from broiler meat in each EU member state, higher levels ranging between 27.2% and 81.8% were observed in isolates from fattening pigs (EFSA and ECDC, 2015a). A study looking at *Salmonella* AMR levels in China between 2016 and 2017 identified that 81.1% of isolates from chicken meat and 73.2% from pork meat had MDR (Zhang *et al.*, 2018). This variation in MDR requires monitoring, to determine which serovars and sources are becoming more resistant and therefore becoming more of a concern to public health.

1.2.7 Mobile Genetic Elements in *Salmonella*

Mobile genetic elements have played a role in the evolution of *Salmonella,* host adaptation and AMR (Silva, Wiesner and Calva, 2012). Notable self-transmissible examples associated with AMR are plasmids, integrons, transposons and insertion sequences. Other routes of horizontal gene transfer are transformation, the uptake of DNA from the surrounding environment, and transduction where a viral vector, such as a bacteriophage, inserts genes into the genome (Monte *et al.*, 2019). Bacteriophages, which have been associated with the spread of virulence genes and insertion sequences, can also carry AMR genes (Silva, Wiesner and Calva, 2012). They also play a role in *Salmonell*a evolution and can be a cause of diversity between strains. Bacteriophages have been identified that can contribute to the spread of AMR genes, such as the spreading of penta-resistance genes in *S*. Typhimurium DT104 (Schmieger and Schicklmaier, 1999); however, their overall contribution to *Salmonella* AMR is not well understood (Colavecchio *et al*., 2017).

1.2.7.1 *Plasmids*

Plasmids are linear or circular strands of DNA that are independent to the chromosome (Shintani, Sanchez and Kimbara, 2015). They can be typed based on their incompatibility (Inc) to replicate in the same host as another related plasmid. Replicon (rep) typing is used to identify the replication initiation protein, this information is used to classify the plasmid into an Inc group.

Plasmids can be transferred between isolates by horizontal gene transfer with conjugation, the transfer of genetic material between organisms using the type 4 secretion system, being the dominant mechanism for the spread of plasmids (Cabezón *et al.*, 2014; Tanner and Kingsley, 2018). The rate of horizontal gene transfer is increased when *Salmonella* is stressed by the host immune response, bile and antibiotics; increasing the spread of genes linked to resisting these stresses (Tanner and Kingsley, 2018). Many plasmids ensure the maintenance of the plasmid in the daughter cells with an addiction system, whereby the daughter cell that does not inherit the plasmid is killed with toxins (Carattoli and Elena, 2009).

Mobile genetic elements, such as plasmids and integrons, often carry resistance determinants (Gupta *et al.*, 2019). Virulence genes can also be found on plasmids. Plasmids can be transferred between bacteria belonging to different species, allowing the spread of AMR between pathogens in the intestines (Su *et al.*, 2004). Plasmid acquisition can have a big impact on the virulence of *Salmonella*, for example, in the 1980's in Japan, a plasmid was acquired by *S.* Dublin which encoded resistance for ampicillin, kanamycin and nalidixic acid (Akiba *et al.*, 2007). Upon acquisition of the plasmid, the numbers of *S.* Dublin increased drastically in cattle, becoming one of the most prevalent disease-causing serovars in cattle.

1.2.7.2 *Common Plasmids in Salmonella*

The type and size of plasmid present in strains vary by serovar (Williams *et al.*, 2013). A comparison of the plasmid types in serovars from a *Salmonella* reference collection identified that some of the serovars were more likely to have a particular plasmid type, for example 19 out of 22 *S.* Paratyphi B genomes had a large untypeable plasmid. However, multiple plasmid types were also identified within serovars; whilst 13 of the 20 *S.* Typhimurium genomes had an IncFIIA plasmid, 6 other plasmid types were identified.

A study comparing plasmid types across *S. enterica*, including sequences from 266 serovars, identified that the most common type was IncFII plasmids (Worley *et al.*, 2018). The majority of the plasmid types were found in a small number of isolates, with only 5 of the 28 types present in over 10 genomes. Another common plasmid type found amongst *S. enterica* is IncA/C plasmids (Han *et al.*, 2018). These plasmids are commonly found with several resistance determinants including ESBLs. Carriage of multiple plasmids is commonly associated in isolates with IncA/C plasmids.

Serotype specific plasmids are also present in *S. enterica*; SPV is found in several serovars, all contain the *spv* gene cluster but otherwise the content of the plasmid varies between serovars (Silva, Puente and Calva, 2017); for example it varies in size from 285kb in *S.* Sendai to 50kb in *S.* Choleraesuis (Feng *et al.*, 2012). They also differ in the roles in pathogenicity they play; *S.* Typhimurium and *S.* Enteritidis SPVs were found enhance apoptosis in human monocytes whereas *S.* Choleraesuis SPV inhibited apoptosis (Huang *et al.*, 2019).

1.2.7.3 *Integrons*

Integrons are mobile genetic elements that can be located in the chromosome, plasmids and transposons and are capable of capturing genes using recombination; carrying several gene cassettes with resistance genes (Khaitsa and Doetkott, 2009; Deng *et al.*, 2015).

An integron contains an integrase gene, a recombination site, a promoter for the integrase and a promoter for the gene cassettes (Figure 1.10) (Cury *et al.*, 2016a). Gene cassettes usually lack their own promoter and are surrounded by two *attC* recombination sites. Integrons are grouped into classes depending on the variant of the *int* gene present with class 1 being the most common.

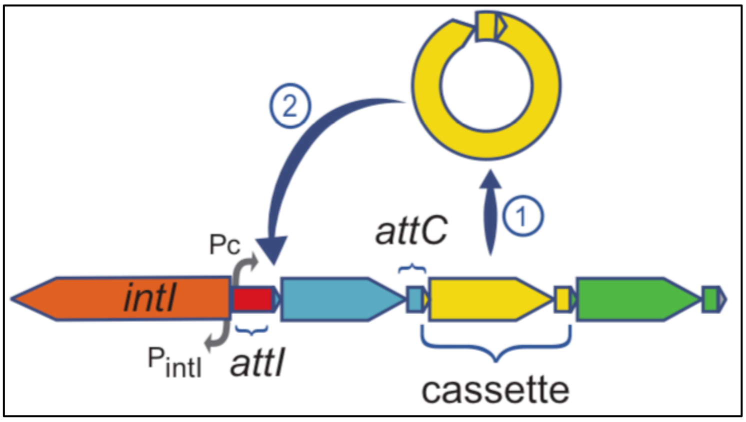


Figure 1.10 The structure of an integron

The gene cassettes can be excised (1) and integrated (2) by the integrase.

Integrase gene (*intI*)🟧 Recombination site (*attI*)🟥 Gene cassettes🟦🟨🟩

attC = Gene cassette recombination sites

P_{intI} = promoter for the integrase

P_c = promoter for the cassettes. Reproduced from Cury *et al*., 2016

Integrons are not identified frequently, after screening all the bacterium genomes available at the time, researchers observed that 9% of the 603 genomes contained integrons (Boucher *et al.*, 2007). A study of 95 *Salmonella* strains isolated from livestock and meat, representing 11 serovars, identified class 1 integrons in 14 isolates and class 2 integrons in 3 (Card *et al.*, 2016).

1.3 *Salmonella* Infantis

S. Infantis is group C_1 serovar of *S. enterica* subsp. *enterica* with the antigenic formula 6,7,14: r: 1,5 (Grimont and Weill, 2007). It has been isolated across the planet from multiple different sources. For example, a study analysing 36 *Salmonella* isolates from food in Brazil found, between 2009 and 2011, that *S.* Infantis was the most frequently identified serovar, present in both meat products, rice and chocolate (Miranda *et al.*, 2017). However, *S.* Typhimurium and *S.* Enteritidis were most frequently isolated from human samples.

S. Infantis has been found in many asymptomatic chickens worldwide but also in asymptomatic humans, suggesting a reduced pathogenicity compared to, for example, *S. Enteritidis* (Yokoyama *et al.*, 2014). Whilst research has shown that a live *S.* Enteritidis vaccine could provide some cross-protection for *S.* Infantis, no vaccines currently exist that are licensed against this serovar (Eeckhaut *et al.*, 2018; Başak and Yardımcı, 2019).

1.3.1 *S.* Infantis Occurrence in Humans

The incidence of human *S.* Infantis infection varies geographically and temporally. Between 2001 and 2007, the proportion of human *S.* Infantis cases ranged from 1.5% to 2.2% of infections caused by the top 20 *Salmonella* serovars in 37 countries associated with the WHO (Hendriksen *et al.*, 2011). A difference was seen between developing and undeveloped countries, while in developing countries a decrease in the proportion of *S.* Infantis isolates was seen between 2005 and 2007, an increase was seen in developed countries.

More recently, *S.* Infantis was the fourth most common serovar in EU member states between 2013 and 2018 (EFSA and ECDC, 2015b, 2015c, 2016b, 2017b, 2018b, 2019a). It was also in the top five most common serovars identified causing human infection in South Africa in 2009, 2010, 2013 and 2016 (Group for Enteric Respiratory and Meningeal disease Surveillance in South Africa, 2009, 2010, 2013, 2016). *S.* Infantis is seen at higher frequencies in some countries; between 2008 and 2015 it was the most common serovar in Israel, accounting for 30% of human cases (Aviv, Rahav and Gal-Mor, 2016).

Whilst *S.* Infantis is rarely the most frequent cause of human salmonellosis, the number of cases it causes is increasing. Figure 1.11 shows the incidence of the top *Salmonella* serovars in the USA between 1970 and 2016. Whilst some of the rarer serovars decreased in incidence, the incidence of *S*. Infantis has increased between 2001 and 2016 by 167% (NCEZID, 2018). In 2016 it was the sixth most common serovar in the USA behind *S.* Enteritidis, *S.* Newport, *S.* Typhimurium, *S.* Javiana and monophasic *S.* Typhimurium.

S. Infantis has been recognised as the cause of outbreaks, including one between January 2018 and January 2019 in the USA that had 129 human cases and was associated with chicken products (CDC, 2019). Another *S.* Infantis outbreak, also in the USA, occurred in May 2019, with 5 cases associated with vegetable trays (U.S. Food & Drug Administration, 2019). In 2018, in EU member states, *S.* Infantis was the cause of 1.1% of *Salmonella* outbreaks (EFSA and ECDC, 2019a).

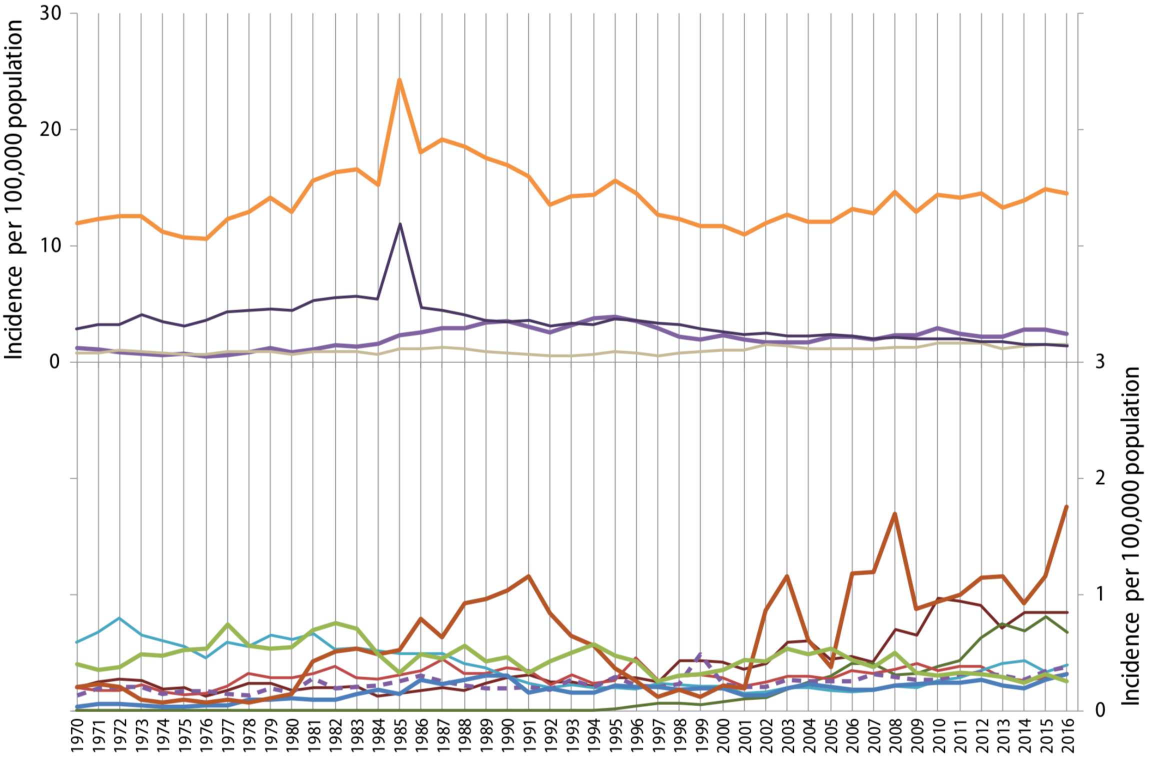


Figure 1.11 USA Human *Salmonella* incidence between 1970-2016
Incidence of human *Salmonella* cases by year of the serovars that caused over 1,000 cases in 2016. All cases were culture-confirmed.

Reproduced from NCEZID, 2018

1.3.2 *S.* Infantis Occurrence in Domestic Fowl

Historically, levels of *S.* Infantis in poultry have not been high, although in recent years they have been increasing; between 2004 and 2012, *S.* Infantis and *S.* Enteritidis were the serovars most frequently identified in *Gallus gallus*, broiler meat and eggs with *S.* Infantis levels increasing during that period (EFSA and ECDC, 2013b, 2014b). It was also the only serovar of the top five causing human infection with a significantly increasing presence in poultry. In 2013, *S.* Infantis was the serovar most frequently isolated from *Gallus gallus* and the second most dominant serovar in broiler meat (EFSA and ECDC, 2015b). Between 2014 and 2018 *S.* Infantis became the dominant serovar isolated from both *Gallus gallus* and broiler meat in EU member states, accounting for 56.7% of isolates from broiler meat in 2018 (EFSA and ECDC, 2015c, 2016b, 2017b, 2018b, 2019a). It was also present in 6.6% of isolates from turkey meat and 6.3% of those from cattle meat.

It is worth noting that the high proportion of *S.* Infantis isolates from *Gallus gallus* is often due to one of the EU member states having a particularly high result; for example, in 2013, 50% of the *S.* Infantis isolates from *Gallus gallus* were found in Romania and in 2018, 50.1% of the S. Infantis isolates from broilers originated from Italy (EFSA and ECDC, 2015b, 2019a). Concerningly it was reported that, dissimilar from previous years, in 2018 *S.* Infantis presence in *Gallus gallus* was widespread in the majority of the reporting member states; suggesting that *S.* Infantis is being disseminated from those countries with higher prevalence, across Europe.

The incidence of *S.* Infantis in poultry varies across the rest of the globe. In Japan *S.* Infantis accounted for 72.2% of isolates from ground chicken and in Iran 75% of isolates from broiler farms were *S.* Infantis (Hara-Kudo *et al.*, 2013; Rahmani *et al.*, 2013). This decreased in Pakistan where 36% of isolates from poultry carcasses were *S.* Infantis (Wajid *et al.*, 2019). *S.* Infantis was also the most common serovar in egg laying farms in New Zealand and in Ecuador 84% of isolates from broilers were *S.* Infantis (Vinueza-Burgos *et al.*, 2016; Kingsbury *et al.*, 2019). Interestingly, high levels of *S.* Infantis are not seen in food animals in the USA, in 2016 it was not among the top five serovars associated with chickens, turkeys and cattle; it was the fourth most common serovar causing infection in pigs (Morningstar-Shaw *et al.*, 2016).

1.3.3 Manifestations of *S.* Infantis Infection

In humans *S.* Infantis causes gastroenteritis, however, it has also been noted to have several other manifestations including: spondylitis, reactive arthritis, cellulitis, osteomyelitis and has been isolated in blood, showing it can causing iNTS (Kohler, 1964; Blaser and Feldman, 1981; Dembski, Patynski and Ciesla, 1995; Ekman *et al.*, 1999; Patil *et al.*, 2013; Muranaka *et al.*, 2015).

S. Infantis is also capable of asymptomatically infecting humans; a rehabilitation centre in Germany experienced multiple outbreaks of the serovar between 2002 and 2009, it was determined that these outbreaks were due to carriers in the kitchen staff and contamination in the kitchen (Miller *et al.*, 2018).

Like many other *Salmonella* serovars, *S.* Infantis is capable of colonising the chicken intestinal tract without causing clinical infection in the bird (Shahada *et al.*, 2010). However, *S.* Infantis has been identified in extra-intestinal sites in broiler chickens, such as the liver and spleen, suggesting that it is capable of causing invasive infection in poultry (Yokoyama *et al.*, 2015). It has also been reported in causing the abortion of a bovine foetus, indicating that it can also cause invasive infection in cattle (Mortelmans, Huygelen and Pinckers, 1958).

1.3.4 *S.* Infantis Population Structure

Despite the prevalence of *S.* Infantis in poultry and the obvious public health concern this presents, very little is known about the population structure of *S.* Infantis. Two eBGs have been identified within the S. Infantis population, eBG31 and eBG297, with the vast majority of isolates belonging to eBG31 (M.A. Chattaway, personal communication, 26th May, 2017).

A paper published recently identified the population structure of 105 *S.* Infantis strains, 100 belonging to eBG31 and 5 to eBG297 (Gymoese *et al.*, 2019). The majority of the sequences were isolated from Denmark, with isolates associated with travel to Asia, North America, South America and Africa included; five strains from humans in Japan were also included. Overall strains from 10 different isolation sources were used, including 56 strains from humans, 12 from swine and 8 from avian samples. The eBG31 strains were found to be polyphyletic, splitting into three lineages that separated 150 years ago, with one containing the majority of the sequences (Gymoese *et al.*, 2019).

Clustering was seen by isolation source in the lineage containing the majority of sequences; a cluster containing mainly isolates from poultry and human sources was present and was also associated with plasmid presence. The authors hypothesised that this cluster had evolved to establish itself in poultry and highlighted the need to determine host relatedness. In one of the smaller lineages the majority of sequences were of animal origin which they hypothesised was due to this lineage being less infectious or less sampled. The eBG31 and eBG297 isolates in this paper were separated on a phylogeny by a long branch, with the eBG297 STs sharing 0-2 alleles with the dominant ST in eBG31, ST32. It was hypothesised that the presence of strains serotyped as *S.* Infantis, but not belonging to eBG31, was due to recombination and that prophages had played a key role in the evolution of the population structure of *S.* Infantis. ST32 has also been identified as the dominant ST in isolates from Iran and Switzerland (Hindermann *et al.*, 2017; Ranjbar, Rahmati and Shokoohizadeh, 2018).

Other studies with smaller numbers of sequences have also observed clustering within phylogenies. Clustering by isolation source was seen in Japanese *S.* Infantis isolates from chicken meat and human samples (Yokoyama *et al.*, 2014), by plasmid presence in Switzerland (Hindermann *et al.*, 2017), and by *bla*$_{\text{CTX-M-65}}$ presence in isolates from Italy and the USA (Tate *et al.*, 2017). The Japanese study found that strains from humans did not cluster exclusively in one group, suggesting that no cluster was more virulent for humans (Yokoyama *et al.*, 2014).

Conversely, little evidence of clustering by country of isolation has been observed in the *S.* Infantis population, with no geographic signal observed in the recent paper comparing isolates from five continents (Gymoese *et al.*, 2019). Currently the paper comparing whole genome sequence data of the largest number of *S.* Infantis isolates included 264 strains from Asia, North America, Africa and South America; clustering of isolates from chickens was observed as was clustering by country of isolation, but as 63% of the sequences came from the USA this could be due to multiple sequences coming from one chicken farm (Acar *et al.*, 2019).

1.3.5 Antimicrobial Resistance in *S.* Infantis

Levels of AMR in *Salmonella* fluctuate globally; this also true for *S.* Infantis (Ozdemir and Acar, 2014; Velhner *et al.*, 2014; Papadopoulos *et al.*, 2017). *S.* Infantis AMR levels are often higher than seen in other serovars, with higher levels of MDR than *S.* Enteritidis seen in Turkey and Iran (Rahmani *et al.*, 2013; Ozdemir and Acar, 2014). In the USA between 2016 and 2017, *S.* Infantis was the serovar with the most ceftriaxone resistance, isolated from humans, chicken and turkey (Food and Drug Administration (FDA), 2019a). Also, in 2011 in EU member states, *S.* Infantis had higher levels of resistance to sulphonamides, ciprofloxacin, nalidixic acid and tetracycline than any other NTS (EFSA and ECDC, 2013a).

Historically, levels of AMR in *S.* Infantis in Europe have been high. Of 93 epidemiologically unrelated *S. Infantis* strains, isolated from humans, broilers and pigs in Germany between 2005 and 2008, 66% of the isolates were susceptible to the 17 antimicrobials that were tested; 31% had MDR (Hauser *et al.*, 2012). Similarly, 48% of isolates from humans in Greece between 2007 and 2010 were resistant to streptomycin (Papadopoulos *et al.*, 2017). Moreover, in Slovenia between 2007 and 2013, 88.5% of 87 *S.* Infantis isolates from broiler faeces had MDR, with the majority of isolates resistant to ciprofloxacin, nalidixic acid, streptomycin, sulfamethoxazole and tetracycline (Pate *et al.*, 2019).

Concerningly the levels of AMR in *S.* Infantis in Europe are rising; the proportion of *S.* Infantis isolates with AMR causing human infection in EU member states is shown in Figure 1.12. In Hungary between 2011 and 2013, 186 *S.* Infantis isolates were identified from broilers and humans; 75.8% of the broiler isolates and 60% of the human isolates had MDR (Szmolka *et al.*, 2018). The predominant AMR profile in the isolates with MDR was nalidixic acid, sulphonamides and tetracyclines. An upward trend was observed with the percentage with MDR increasing each year.

Higher levels of AMR in *S.* Infantis are present in *Gallus gallus* across the EU member states; *S.* Infantis from broilers are a large contributor as in 2014, 92.7% of isolates tested were ciprofloxacin resistant, 92.1% to nalidixic acid, 82.7% to sulphonamides and 81.3% to tetracycline (EFSA and ECDC, 2015a, 2016a). *S.* Infantis has become a significant contributor to *Salmonella* MDR in Europe, 31% of the isolates from broilers had MDR in 2014 and 2016, with 70% of isolates from broiler meat having MDR in 2016 (EFSA and ECDC, 2016a, 2018a). Due to its high prevalence within the European broiler population, clones of MDR *S.* Infantis have spread into the food chain (EFSA and ECDC, 2017a; Hindermann *et al.*, 2017).

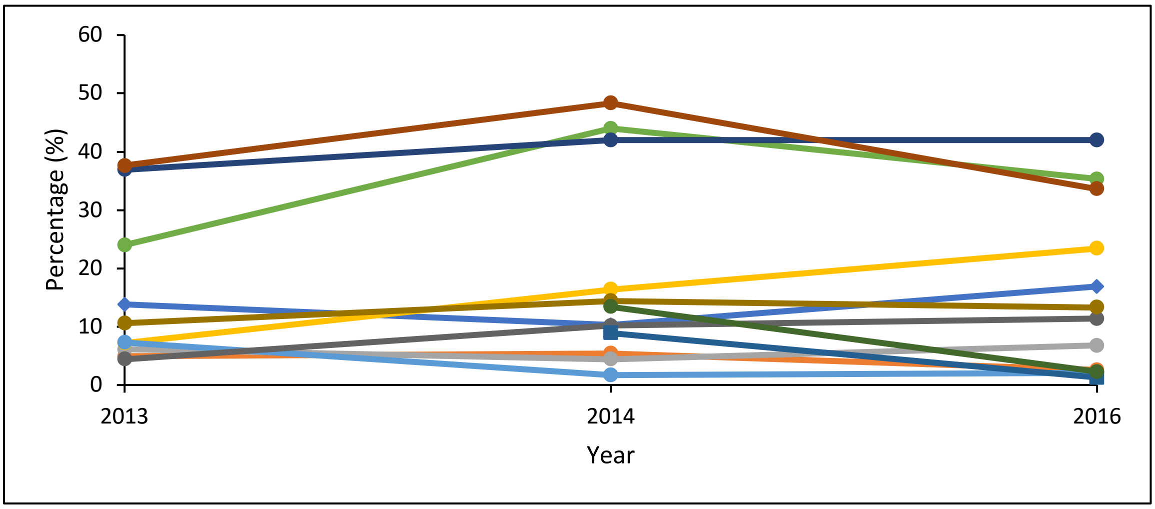


Figure 1.12 Trends in AMR in *S.* Infantis in Europe

Percentage of *S.* Infantis isolates from humans in EU member states that were resistant to antimicrobials in 2013, 2014 and 2016.

Generated using data from EFSA and ECDC, 2015a, 2016a, 2018a.

A similar upward trend in the identification of AMR in *S.* Infantis is seen in the USA (Figure 1.13). In the human isolates collected between 1996 and 2017 the levels of AMR fluctuate drastically, despite this, the highest proportion of isolates with resistance to ampicillin, ceftriaxone, chloramphenicol, nalidixic acid, streptomycin, tetracycline, trimethoprim-sulfamethoxazole was seen in the latest data point, 2017.

Unlike the human isolated *S.* Infantis strains from the USA, the AMR seen in the isolates from chickens appears to decrease between 1997 and 2007; from 2009 onwards, the levels then increase to higher levels than seen in the isolates from humans. 50% of the antimicrobials tested peaked in 2017: ampicillin, ceftriaxone, chloramphenicol, nalidixic acid, streptomycin, streptomycin, sulfamethoxazole-sulfisoxazole and tetracycline. The highest levels of resistance in 2017 was seen for nalidixic acid where 84.31% of the 408 isolates tested for it were resistant. Unlike the *S.* Infantis isolated from broilers in Europe, no ciprofloxacin resistance was reported in the *S.* Infantis isolates from the USA.

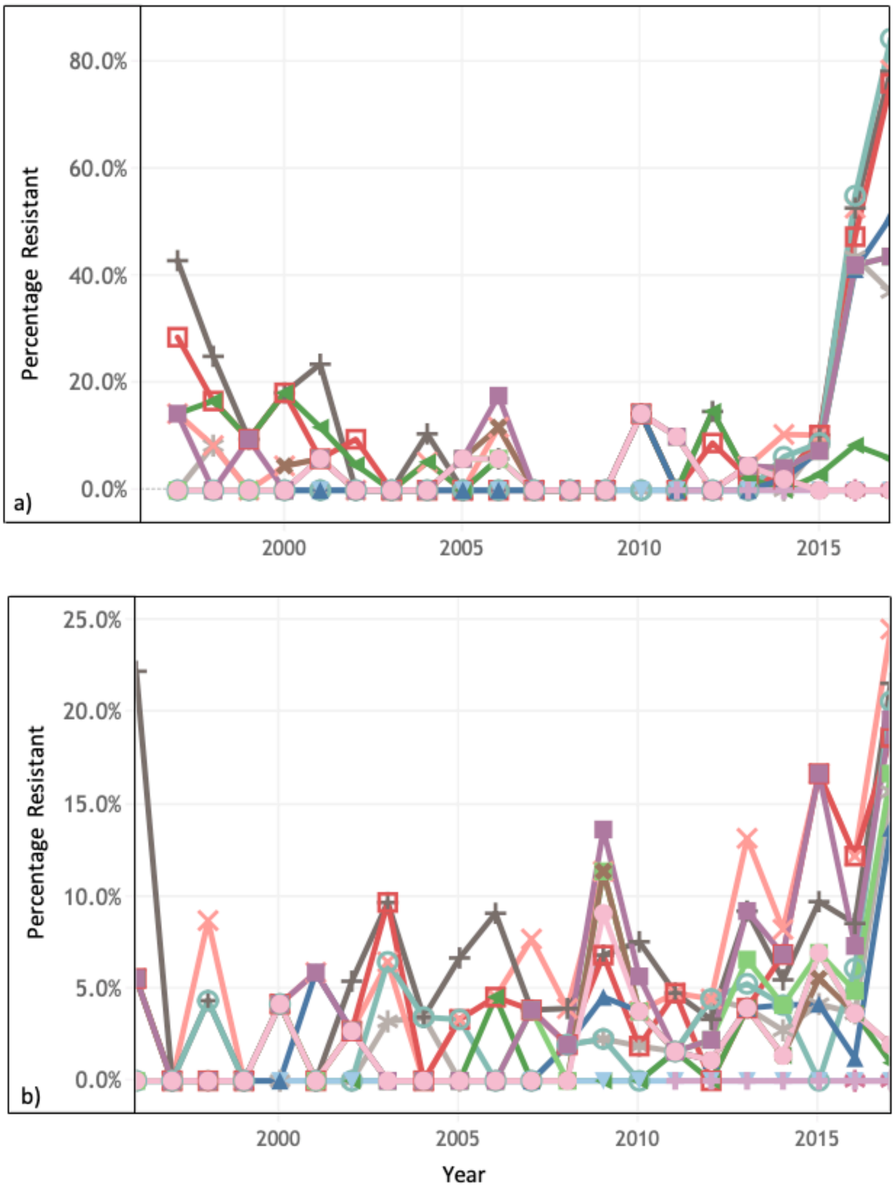


Figure 1.13 The trend of antibiotic resistance in *S.* Infantis in the USA

a) Percentage of isolates from retail chicken from 19 states and samples taken at slaughter throughout the USA with resistance to each antimicrobial between 1997 and 2017

b) Percentage of isolates causing clinical illness in humans with resistance to each antimicrobial between 1996 and 2017

Amoxicillin-Clavulanic Acid ● Ampicillin ■ Azithromycin + Cefoxitin ✖ Ceftriaxone ■ Chloramphenicol ▲ Ciprofloxacin ▼ Gentamicin ◄ Meropenem ✱ Nalidixic Acid ○ Streptomycin□ Sulfamethoxazole-Sulfisoxazole + Tetracycline × Trimethoprim-Sulfamethoxazole ✳

Generated using data from Food and Drug Administration (FDA), 2019b

High levels of AMR in *S*. Infantis have also been identified elsewhere. In Japan, in strains isolated from broiler caecal samples between 2004 and 2006; 100% of the 120 isolates were resistant to streptomycin, sulfamethoxazole and oxytetracycline (Shahada, Chuma and Dahshan, 2010). In Pakistan, 54 *S.* Infantis isolates from poultry carcasses between 2015 and 2016 were tested for antimicrobial susceptibility (Wajid *et al.*, 2019). All of the isolates had MDR, with pefloxacin resistance the most common AMR profile present. Mutations were observed in the four QRDRs of *gyr*AB and *par*CE with the highest level in *par*E at 62.5%. Also, 38 *S.* Infantis isolates were identified in hospitals in Tehran, Iran between 2008 and 2010, over 80% had MDR (Ranjbar, Rahmati and Shokoohizadeh, 2018).

Several papers have identified different ESBLs in *S.* Infantis; strains harbouring the *bla*$_{\text{CTX-M-65}}$ gene have been found in the UK, USA, Ecuador, Peru and Switzerland (Burke *et al.*, 2013; Cartelle Gestal *et al.*, 2016; Hindermann *et al.*, 2017; Tate *et al.*, 2017; Granda *et al.*, 2019). *S.* Infantis strains containing *bla*$_{\text{TEM-52}}$ and *bla*$_{\text{TEM-1}}$ have been identified in Japan between 2004 and 2006 from broiler caecal samples (Shahada, Chuma and Dahshan, 2010). The *bla*$_{\text{TEM-52}}$ gene has also been found on an IncI1 plasmid in *S.* Infantis isolates from humans and poultry in Belgium (Cloeckaert *et al.*, 2007). More recently *bla*$_{\text{TEM-70}}$, *bla*$_{\text{TEM-148}}$ and *bla*$_{\text{TEM-198}}$ have been identified on plasmids in *S.* Infantis isolates from chicken meat in Turkey (Acar *et al.*, 2019). *S.* Infantis isolates containing an ESBL gene commonly also have resistance to nalidixic acid, sulfamethoxazole or tetracyclines (Vinueza-Burgos *et al.*, 2016; Hindermann *et al.*, 2017; Granda *et al.*, 2019) .

The incidence of ESBLs in *S.* Infantis varies; in 2015, ESBLs were present in 5.3% of *S.* Infantis isolates across half of the EU member states (EFSA and ECDC, 2017a). Investigation of broiler farms in Ecuador between 2013 and 2014 isolated 62 *S.* Infantis strains, of which 55% were ESBL positive (Vinueza-Burgos *et al.*, 2016).

1.3.6 Mobile Genetic Elements in *S.* Infantis

Several studies have identified the presence of different mobile genetic elements in *S.* Infantis. Of 23 *S.* Infantis isolates from chicken meat in Turkey, 91.3% of the isolates contained plasmids (Acar *et al.*, 2019). AMR genes and virulence factors such as bacteriocins were on the plasmids, with most of the AMR genes carried in transposons or insertion elements.

Between 2007 and 2008, 2 *S.* Infantis isolates were isolated from pig faecal samples in Japan; they had MDR and, in particular, had extended-spectrum cephalosporin resistance due to the presence of a plasmid containing *bla*$_{\text{CMY-2}}$ (Dahshan *et al.*, 2010). *S.* Infantis isolates containing plasmids with ESBLs have also been found in the USA; 29 of 34 isolates from humans in 2015 carried the *bla*$_{\text{CTX-M-65}}$ gene on a plasmid (Brown *et al.*, 2018).

A study in Iran looked at the presence of integrons in *S.* Infantis strains isolated between 2009 and 2011 from chickens (Asgharpour *et al.*, 2014). 16% of the 50 isolates had resistance to at least 4 antimicrobials and *int1* was identified in 36% of the isolates, associated with resistance to nalidixic acid, streptomycin and tetracycline. Class 1 integrons were also present in 34 of 54 *S.* Infantis isolates from poultry samples in Pakistan (Wajid *et al.*, 2019). Another study in Japan, analysed integron presence in 120 *S.* Infantis isolates from broiler caecal samples isolated between 2004 and 2006 (Shahada *et al.*, 2010). All of the isolates contained a class 1 integron; they also all had MDR and carried either a 180kb plasmid or this plasmid plus another 50kb plasmid. 91% of the isolates with the extra 50kb plasmid were found to contain the *bla*$_{\text{TEM}}$ gene.

1.3.6.1 *Plasmid of Emerging S. Infantis (pESI)*

A study elucidating the cause of the emergence of *S.* Infantis in Israel between 2006 and 2007 compared strains isolated pre and post 2006 (Aviv *et al.*, 2014). Whilst the strain isolated pre 2006 was susceptible to the antibiotics tested, the emergent strain had MDR to the 12 antimicrobial classes tested. A unique megaplasmid conferring MDR was identified in the emergent strain and was named plasmid of emerging *S.* Infantis (pESI). pESI was not present in strains isolated before 2007 but 82% of strains isolated between 2007 and 2009 were positive for pESI (Gal-Mor *et al.*, 2010; Aviv, Rahav and Gal-Mor, 2016).

pESI has an IncP-1α origin of replication but is able to co-exist with other IncP plasmids and was found to be incompatible with Inc1 plasmids; suggesting that it evolved from recombination of other plasmid types (Aviv *et al.*, 2014). It is approximately 280kb in size and carries *aadA1*, *sul1* on a transposon; the *tetAR* operon on a transposon and *dfrA1* on a class 1 integron; conferring resistance to streptomycin, sulfamethoxazole, tetracycline and trimethoprim respectively. Emergent strains tested also had resistance to nalidixic acid and nitrofurantoin, but these were found to not be pESI mediated and instead located on the chromosome. The nalidixic acid resistance was due to an amino acid substitution in the QRDR of *gyrA* and the nitrofurantoin resistance due to a nonsense mutation in *nfsA*. The mer operon which encodes mercury resistance was also present, as was a yersiniabactin iron uptake system.

pESI also increases the virulence of *S.* Infantis (Aviv *et al.*, 2014). A mouse model was infected with *S.* Infantis isolates with or without pESI. pESI presence increased caecal colonisation, inflammation and tissue damage. It was also found to enable growth in the presence of mercury and hydrogen peroxide and increase biofilm formation. A more recent study identified that pESI is not restricted to *S.* Infantis; in a mouse model pESI was transferred to *E. coli* strains and gram-positive organisms in the microbiota; it persisted in the *E. coli* strains but not in the gram-positive species (Aviv, Rahav and Gal-Mor, 2016).

Since the first description of pESI, pESI-like plasmids have also been identified in *S.* Infantis from Denmark, Italy, Switzerland, Hungary, Peru, Turkey and the USA; suggesting that pESI confers a selective advantage to certain *S.* Infantis strains and could be associated with the increased incidence of *S.* Infantis (Franco *et al.*, 2015; Hindermann *et al.*, 2017; Iriarte *et al.*, 2017; Tate *et al.*, 2017; Szmolka *et al.*, 2018; Acar *et al.*, 2019; Gymoese *et al.*, 2019).

Concerningly, variants of pESI have been found containing ESBLs. The $bla_{\text{CTX-M-65}}$ gene was found in isolates with pESI from cattle, chicken and humans in the USA and from poultry meat and humans in Switzerland (Hindermann *et al.*, 2017; Tate *et al.*, 2017). Also, the $bla_{\text{CTX-M-1}}$ or $bla_{\text{CTX-M-65}}$ gene was found in pESI positive isolates from broilers, broiler meat and humans from Italy (Franco *et al.*, 2015). More recently, *S.* Infantis isolates have been identified from broilers in Italy which contained both pESI with the $bla_{\text{CTX-M-1}}$ gene and an IncX1 plasmid with *mcr*-1.1 which encodes colistin resistance (Carfora *et al.*, 2018).

1.3.7 Virulence Factors

Virulence factors, have been found in varying levels in different *S. enterica* subsp. *enterica* serovars, enabling colonisation of different hosts and causing different kinds of infection (Cheng *et al.*, 2015). The virulence markers present in 35 *S.* Infantis strains from humans or food in Brazil, isolated between 1984 and 2009 were compared; although several virulence genes were identified, *spvB*, which is commonly found on plasmids associated with virulence, was not present in any of the isolates (Almeida *et al.*, 2013).

Virulence gene presence varies by country, a study from Israel reported that *SopE* is not present in *S.* Infantis (Aviv *et al.*, 2019). Conversely, in Pakistan the virulence gene most frequently identified from *S.* Infantis isolates from poultry carcasses was *SopE* (Wajid *et al.*, 2019). Also, *S.* Infantis isolates from Turkey contained higher levels of *SopE2* than *S.* Typhimurium and *S.* Kentucky strains; different virulence gene patterns were identified in the samples from broilers and breeders, suggesting that *Salmonella* contamination of the broilers was not occurring through the breeders (Sever and Akan, 2019).

An experiment using mice as an in vivo model for infection determined that, in comparison to *S.* Typhimurium, *S.* Infantis performed better than *S.* Typhimurium at adhering to the host epithelial cells but was significantly less invasive, regardless of pESI presence, inducing less inflammation (Aviv *et al.*, 2019). This was found to be associated with a reduction in the expression of SPI-1 genes in the *S.* Infantis isolates. *S*. Infantis isolates from poultry have also been identified as being weak to moderate biofilm formers (Pate *et al.*, 2019).

1.4 Aims and Objectives

S. Infantis is causing increasing numbers of human infection globally and is identified frequently in poultry. Despite this, very little is known about the genetic diversity of *S.* Infantis, with published research including small numbers of sequences from a limited number of locations and sources. *S.* Infantis has also been reported as being associated with high levels of AMR. Whilst all the available information of AMR in *S.* Infantis is undoubtably useful, it does not provide evidence on what AMR determinants are responsible for the increase in AMR seen in *S.* Infantis isolates globally, and if the increase in AMR is occurring worldwide. Also, pESI has been identified in *S.* Infantis isolates from multiple countries; however, the levels of pESI in *S.* Infantis and how the plasmid varies is unknown; such information would determine the public health risk pESI presents.

Despite the large difference in incidence of *S.* Infantis in humans and poultry it is not known why this pathogen is so successful in one host and less so in the other. A hypothesis that would explain this disparity is that there are genetic differences between the *S*. Infantis isolates that cause infection in humans and poultry. Ascertaining if this was the case would be beneficial for public health surveillance as it could identify characteristics that have a greater risk of causing human infection.

Therefore, the objectives of this project were to:

1. Determine the population structure of a global collection of *S.* Infantis.

2. Identify the global levels of AMR in *S.* Infantis and whether resistance determinants vary by geography and source.

3. Identify the global levels of pESI in *S.* Infantis and the association of pESI presence with geography, source and the increase in AMR.

4. Establish whether there are genetic differences between *S*. Infantis isolates from human and poultry sources that explain the difference in incidence between these sources.

2. Chapter 2. Overall Methods

2.1 Sequences and Strains Used in this Study

S. Infantis sequence data or DNA was collected from as many online databases and collaborators as possible, to capture the greatest diversity in time and space.

2.1.1 Sources of Sequence Data

Sequence data were downloaded from online databases and shared by collaborators.

2.1.1.1 *GenBank and the European Nucleotide Archive (ENA)*

A literature search was carried out in August 2017 to identify papers containing *S.* Infantis sequence data. GenBank and the ENA were also searched; sequence data for 270 isolates was found from the GenBank search and 25 from the ENA search (Benson *et al.*, 2005; Leinonen, Akhtar, *et al.*, 2011). As the input for the bioinformatics pipeline was paired fastq files, all papers containing only whole genome assemblies were excluded; resulting in 182 *S.* Infantis strains with fastq files being included and downloaded from the Sequence Read Archive (SRA), GenBank, the ENA or the DNA Data Bank of Japan (Leinonen, Sugawara, *et al.*, 2011; Mashima *et al.*, 2017).

2.1.1.2 *Enterobase*

The Enterobase database was searched, up until 19th February 2018, using the qualifiers *S.* Infantis, *S.* Infantis (Predicted), Infantis and eBG31 (Alikhan *et al.*, 2018). The Enterobase database was searched for eBG297 sequences on 28.05.19 using the qualifier eBG297. Sequences that had been sequence typed by Enterobase as eBG31 or eBG297, but serotyped by the uploader as another serotype, were included. All sequences that were duplicated from the GenBank search or had been uploaded by PHE were at this point removed.

2.1.1.3 *Public Health England (PHE)*

The Gastrointestinal Bacterial Reference Unit (GBRU) began to trial whole genome sequencing on isolates from 2012, 2013 and 2014 and implemented it as routine for all pathogens they received in 2015. All of the eBG31 sequences available on 02.08.2017 were included. All eBG297 sequences available on 28.05.2019 were included.

2.1.1.4 *Animal and Plant Health Agency (APHA)*

The metadata for 62 *S.* Infantis isolates that had been sequenced, some of which were available on the SRA, was provided by Dr Liljana Petrovska-Holmes (APHA).

2.1.2 Accessing and Downloading the Sequence Data

The sequences from the ENA were downloaded directly from the web browser. All GenBank sequences and those on Enterobase which were present on the SRA were downloaded using *fastq-dump* from the *sra toolkit* with the split-3 option to output paired end fastqs (National Center for Biotechnology Information, 2014).

Some Enterobase sequences had been uploaded as an assembly to RefSeq or the Trace archive. The accession numbers for each of these were searched on GenBank and the ENA to identify any associated fastqs. These were downloaded either directly from the web browser or using *fastq-dump*.

Many samples downloaded from Enterobase did not have metadata associated with them at the time of download. Metadata can take several months to be associated with samples on Enterobase, so metadata was downloaded again on 13.02.19 and those that had acquired metadata were updated. GenBank was also searched for isolates still missing metadata using *edirect*, linking the SRA accession number to metadata in BioSample (Kans, 2019).

APHA sequences on GenBank were searched for using *edirect*, linking the strain name in the BioSample database to the SRA accession number. Again, sequences found to be associated with APHA metadata were downloaded using *fastq-dump.* Any sequences not available on GenBank were securely shared with me by Dr Liljana Petrovska-Holmes.

PHE sequence data was shared by Dr Hassan Hartman via Secure File Transfer Protocol (SFTP). Some were sent in duplicate, for those samples, the version of the sequence that had been completed most recently was included. Several of the PHE fastqs sent via SFTP were smaller than expected. I plotted the size distribution of all of the PHE sequences and any smaller than the peak of sequences (cut-off set at 40 megabytes per fastq) were investigated. All the affected isolates had been sequenced more than once by PHE and incorrect versions had been sent. The correct versions were sent and run through the pipeline again.

Following this, I investigated the size of the downloaded Enterobase fastqs. I plotted the distribution of these sizes and all those smaller than the cut-off of 60 megabytes per fastq were downloaded again. If the new download was larger than the original download it was used instead.

2.1.3 Sources of Strains

DNA of *S.* Infantis isolates that had not yet been sequenced was shared by the APHA, PHE and the National Institute for Communicable Diseases (NICD), South Africa.

2.1.3.1 *APHA*

The APHA collect *Salmonella* isolates either for surveillance or for research projects. All *S.* Infantis isolates that had not been sequenced by March 2018 from the surveillance group were selected for inclusion.

2.1.3.2 *Historical PHE (hPHE) Isolates*

The metadata for all unsequenced *S.* Infantis reported to PHE, between 2000 and 2014, was assessed and a list of isolates to sequence was generated. Those associated with foreign travel or isolated from blood, urine or chickens were given preference; ensuring that the isolates selected were evenly distributed by time and if there was foreign travel, by continent. If multiple samples came from the same source, location and time, then one isolate was chosen to represent that group. I searched for these samples in the PHE culture stores, with the assistance of Tracey Dealey and Dr Martin Day; all that were found and successfully cultured were included.

2.1.3.3 *NICD, South Africa*

A list of all *S.* Infantis reported to the NICD between January 2003 and October 2017 was provided by Dr Anthony Smith. Isolates associated with any of the following metadata were given top priority for sequencing: high levels of antibiotic resistance; associated with an outbreak; sourced from blood, pus, cerebrospinal fluid (CSF), urine and rectal swabs or non-human sources. Isolates were also selected from the smaller provinces to ensure a distribution of isolates across the whole country. This resulted in a list of 285 high priority isolates to be sequenced.

A mid priority list was created to capture isolates across each year, to ensure a distribution of strains by time. Ten isolates from the beginning, middle and end of each year were included. This list also included all isolates from infants under 1 year of age. A low priority list of 50 strains was prepared which included one isolate from 2013 and 49 from 2009. With the help of Tina Duze, Shannon Williams and Nomsa Tau, the NICD culture stores were searched and all those on the three priority lists that were found and were culturable were included, with more time being dedicated to finding those on the higher priority lists.

2.1.4 Anonymising Sequence Data

All isolates to be sequenced in-house were given a coded name, to allow both for the anonymisation of codes that could be not made publicly available and for the differentiation of data from different collaborators. All the NICD isolates were given a number and either preceded by 'SA' or followed by 'southafrica'. The PHE, hPHE and APHA isolates were given a number which was preceded by 'PHE', 'hPHE' and 'APHA' respectively.

2.1.5 Metadata Handling

The metadata for isolation source was stratified into smaller groups to enable plotting of source on the phylogenies. Appendix II Table II.1 lists the keywords used to sort the isolates into human, poultry or environmental source groups. Information about gender and age were also used as an indication that the sample was from a human. Isolates were classed as having an unknown source if that information was missing or if there was insufficient information to determine the source group.

The same process was applied to the geographic location of isolation, with all countries being categorised by continent; Appendix II Table II.2 lists the countries included in each of the categories. PHE isolates that had foreign travel information were classed as originating from that continent and isolates that did not, as originating from the UK. The year of isolation results were grouped into date ranges of varying lengths depending on the number of strains present in the group; comprising *S.* Infantis strains from 1989, 1995-1997 and in every year between 1999-2019.

2.2 DNA Extraction for Whole Genome Sequencing

All but one of the genomes in this project were short read sequenced, using Illumina whole genome sequencing. A different method was used to extract high molecular weight DNA for the long-read sequencing of the eBG31 reference isolate.

2.2.1 Short Read Sequencing

Different methods were used to extract DNA by each of the collaborators in this project.

2.2.1.1 *NICD Isolates*

I performed the DNA extraction of the NICD isolates, with the assistance of Shannon Williams, Nomsa Tau and Tina Duze. The Qiagen QIAmp DNA Mini kit (Qiagen, Germany) was used to extract the DNA directly from an overnight culture on an agar plate, following the protocol except for an incubation period of 60 minutes with proteinase K and the addition of 100µl of nuclease free water in the elution step (Qiagen Sample & Assay Technologies, 2016). 475 samples were extracted but upon receipt of the DNA at the University of East Anglia (UEA), only 450 were suitable for sequencing.

2.2.1.2 *hPHE Isolates*

The agar slopes or Dorset egg slopes containing the isolates were located in the PHE culture stores. If the agar slope was extremely dehydrated it was suspended in Difco nutrient broth and incubated, whilst shaken, at 37°C overnight. A small loop of culture from the agar slope or Dorset egg slope was inoculated in 1ml of Difco nutrient broth and another loop was used to streak a MacConkey agar plate. Both were incubated overnight at 37°C, the broth whilst shaken. Growth on the agar plate was checked for purity by Dr Martin Day and Dr Claire Maguire; if purity was confirmed, the broth was taken to the DNA extraction facility. If no growth was present the above was repeated with the addition of a blood agar plate.

Broths were taken to the GBRU DNA extraction team who, using the QiaSymphony platform (Qiagen), extracted the DNA (Walle *et al.*, 2019). They also performed DNA quantification using GloMax (Promega, USA) with Quant-iT reagents (ThermoFisher, USA) (Thermo Fisher, 2015b). Fourteen of the samples had had DNA extracted twice, those with the highest concentration from the GloMax results were carried forward.

2.2.1.3 *APHA Isolates*

DNA extraction of APHA isolates was carried out by APHA staff, in particular Dr Carmen Garcia-Pelayo. DNA was extracted using a MagMAX CORE Nucleic Acid Purification Kit (ThermoFisher) on a KingFisher Flex Purification System (ThermoFisher), following manufacturer's instructions (De Lucia *et al.*, 2018).

2.2.2 Long Read Sequencing

The reference eBG31 isolate was grown in 2ml of BHI broth at 37°C and 3mg/ml lysozyme was prepared. The next day DNA was extracted using the FireMonkey DNA extraction kit (RevoluGen, UK); following the DNA extraction for bacteria protocol, except for the incubation length with lysozyme, which was increased to 1 hour (Revolugen, 2017). RNase was added for the extraction of DNA alone.

2.3 DNA Quality Control and Quantification

Per batch, DNA quality was checked by looking at the absorbance with a spectrophotometer. All DNA was quantified by fluorescence, using a Qubit fluorometer (ThermoFisher).

2.3.1 Measuring DNA by Absorbance

For the first sequencing run, the DNA of 151 *S.* Infantis isolates was tested for purity using a NanoDrop Spectrophotometer (ThermoFisher). Briefly, the NanoDrop was cleaned with DNase/RNase free distilled water, 1µl of this was used as a blank and 1µl of DNA for each sample was measured, with cleaning being carried out between each measurement. The desired absorbance values were approximately 1.8 for 260:280 and 2 for 260:230.

For all other samples that were diluted, at least five from each DNA extraction batch were checked on the NanoDrop for consistent results. This was not carried out for the DNA from the APHA, as they shared their absorbance results.

2.3.2 Measuring DNA by Fluorescence

The DNA for the first 151 isolates was quantified with the assistance of Dr Emma Manners (UEA) using Qubit Broad-Range and High Sensitivity assay kits, following the kit protocols (Thermo Fisher, 2015c, 2015d).

2.3.2.1 *Trial Methods*

With the aim of speeding up quantification, a method using the Qubit reagents in a microplate reader was employed (Tanny, 2014). 1x TE buffer was made by mixing 1ml 10mM Tris (Qiagen) and 2µl EDTA. The 100ng/µl standard that came with the Qubit Broad-Range assay kit was then diluted to make standards at 5, 10, 20, 40, 60 and 80ng/µl. The 0 ng/µl and 100 ng/µl standards that came with the Qubit kit were also used. Qubit High Sensitivity reagent was mixed with High Sensitivity buffer at a ratio of 1:200. 198µl of the working solution was pipetted into each well of a black 96 well plate. 2µl of each standard was added to three wells and 2µl of DNA for each sample was added to a well after vortexing. The plate was vortexed, incubated for 2 minutes and the fluorescence was read on a FLUOstar Omega microplate reader (BMG Labtech, Germany), using the 485/520 setting for excitation/emission.

Subsequently, the method suggested by BMG Labtech was followed and only the standards provided in the Qubit Broad-Range assay kit were used to make the standard curve (Krumm, Gröne and Maurer, 2017). Quant-iT High Sensitivity and Broad-Range kits were then used with the prepared standards; the kit protocol was followed (Thermo Fisher, 2015b, 2015a).

2.3.2.2 *Final Method*

Any aliquots of DNA quantified with Quant-iT were requantified with Qubit and Qubit, using the 96 well format, was used exclusively for the remaining samples. If the concentration of any of the stock DNA was too low to be quantified, it was excluded.

2.3.3 Dilution of DNA for Illumina Sequencing

Stock DNA was stored in a -80°C freezer. It was diluted to 10ng/µl with DNase/RNase free distilled water, unless the concentration was close to 10ng/µl. Diluted DNA was then further diluted to the concentration required for library preparation. The majority were diluted to 0.2ng/µl; when the Quadram Institute Bioscience (QIB) Sequencing Facility was used this was amended to 0.5ng/µl. Initially the DNA was diluted into Eppendorf tubes but later into 96 well plates. Dilutions were stored at -20°C.

2.4 Whole Genome Sequencing

The DNA was checked for quality, diluted, library prepped and whole genome sequenced.

2.4.1 Illumina Whole Genome Sequencing

All of the NICD, hPHE and APHA DNA was sequenced within UEA or QIB.

2.4.1.1 *Illumina Library Prep – Standard Protocol*

The first 136 DNA samples were tagmented, indexed, amplified and cleaned up, with the assistance of Dr Emma Manners, following Nextera XT DNA Library Prep Kit protocol version 3 (Illumina, Great Chesterford) (Illumina, 2018a). After library prep, the quality of at least 2 isolates from each batch was assessed using the TapeStation (Agilent Technologies, USA). 10µl of sample buffer was aliquoted into each PCR tube, 1µl of ladder into the first tube and 1µl of DNA into the others. The TapeStation electropherogram was used to check for a defined peak. At least one of the samples from each batch had a corresponding sample containing pre-amplification DNA. The peaks of these were compared to confirm that amplification had taken place.

Libraries were quantified using the Qubit High Sensitivity assay kit and normalised to 4nM, using both the modal peak height for that batch calculated from the TapeStation and the concentration of each library to identify the amount of DNase/RNase free distilled water to add. 5µl of 128 of the libraries were pooled and the sections Denature a 4nM Library and Dilute a Denatured 20pM Library in the MiSeq System Denature and Dilute Libraries Guide were followed (Illumina, 2017). 6µl of denatured and diluted PhiX was added and the libraries were loaded onto the reagent cartridge, which was run with a mid-output flow cell on an Illumina NextSeq 500 (Illumina).

The next 152 isolates were library prepped, again following the Nextera XT DNA Library Prep Kit protocol version 3 (Illumina, 2018a). The TapeStation was again used to check for amplification. These libraries and 8 omitted from the first sequencing run were then submitted to the QIB Sequencing Facility, run by David Baker, who normalised, pooled and sequenced the libraries on a mid-output flowcell on an Illumina NextSeq 500.

2.4.1.2 *Illumina Library Prep – Amended Protocol*

The remaining 495 isolates were submitted to the QIB Sequencing Facility as DNA, diluted to either 0.2ng/µl or 0.5ng/µl. David Baker performed the following: a tagmentation mix was created for each sample using 0.9µl Illumina Tagment DNA Buffer, 0.09µl Illumina Tagment DNA Enzyme and 2.01µl PCR grade water. 3µl of this mix was mixed with 2µl of the 0.5ng/µl DNA in a chilled 96 well plate and then heated in a PCR block for 10 minutes at 55^{o}C, these proportions were adjusted for the DNA that had been diluted to 0.2ng/µl.

A PCR mastermix using reagents from the Sigma-Aldrich Kap2G Robust PCR kit (Sigma-Aldrich, USA), containing 4µl kapa2G buffer, 0.4µl deoxyribonucleotide triphosphate, 0.08µl polymerase and 6.52µl PCR grade water, was added to each well. 2µl of Illumina Nextera XT Index Kit primers (Illumina) were added to each well, followed by 5µl of the tagmentation mix. This was mixed and heated in a PCR block at 72^{o}C for 3 minutes, 95^{o}C for 1 minute and 14 cycles of 95^{o}C for 10 seconds, 55^{o}C for 20s and 72^{o}C for 3 minutes.

The libraries were quantified using either a Quant-iT High Sensitivity kit or a QuantiFluor dsDNA System kit (Promega) on a FLUOstar Optima plate reader (BMG

Labtech) (Thermo Fisher, 2015b; Promega Corporation, 2018). They were then pooled, cleaned up and double-SPRI size selected between 0.5 and 0.7X bead volumes with KAPA Pure Beads (Roche, Switzerland) (KAPA Biosystems, 2017).

The final pool was quantified using Qubit or QuantiFluor reagents with a Qubit 3.0 instrument (Thermo Fisher, 2015d; Promega Corporation, 2018). The molarity was calculated using the Agilent Tapestation 4200 (Chapter 2.4.1.1). Following the Illumina protocol, the pool was denatured and loaded at a final concentration of 1.5pM with a 1% PhiX spike (Illumina, 2018b).

These libraries, including those that needed resequencing, were spread across 5 sequencing runs, using mid and high output flow cells and an Illumina NextSeq 500.

2.4.2 MinION Whole Genome Sequencing

To long-read sequence the DNA for the eBG31 reference isolate, a Nanopore Rapid Barcoding Kit (Oxford Nanopore Technologies, UK) was used, with the assistance of Dr Gemma Langridge and Dr Emma Ainsworth, following the protocol (Oxford Nanopore Technologies, 2018) except for the following:

- A concentration of 600ng of high-molecular weight DNA was used as input instead of 400ng
- The pooled barcoded sample and AMPure beads were vortexed for 2 minutes at 1800rpm before a 5 minute incubation at room temperature
- The beads were washed with 80% ethanol

The priming and loading of the flow cell were done solely by Dr Gemma Langridge and Dr Emma Ainsworth. The running parameters for the experiment were left as default.

2.4.3 Resequencing of Failed Genomes

As the sequences passed through the analysis pipeline (Chapter 2.6), several had to be excluded due to poor quality. For those strains with DNA at QIB, resequencing took place. A poor quality genome was resequenced for one or more of the following reasons:

- It did not pass into *SnapperDB* due to low coverage (Chapter 2.6.5)
- It had poor assembly quality (Chapter 2.6.7)

- It was sequence typed as eBG31 or eBG297 but had too high QC max percentage non consensus base values for all loci to be included (Chapter 2.6.3)
- It did not have a close SLV, DLV or multiple-locus variant (MLV) when sequence typed and had high maximum consensus values; those that did not have a close SLV, DLV or MLV but had acceptable maximum consensus scores were not resequenced (Chapter 2.6.3).

All resequencing was carried out by the QIB Sequencing Facility, who either resequenced the already prepared libraries or remade the libraries. 20 had to be diluted again from stock concentration to 0.5ng/µl using a QuantiFluor dsDNA System kit (Promega Corporation, 2018). Any sequences that were still poor quality after the final sequencing run were not resequenced again due to time constraints.

2.5 Selection of Reference Sequences

Reference genomes for eBG31 and eBG297 were required for the bioinformatic analyses.

2.5.1 eBG31

The reference selected for eBG31 was SRR1968494, the reference used by PHE, to allow for continuity with their work. They chose this sequence as it was the highest quality assembly in their collection; it had the smallest number of contigs and the highest N50. We chose to long-read sequence this isolate to generate an even higher quality reference.

2.5.2 eBG297

As PHE did not have a reference for eBG297, I used their methods for reference selection, with all the eBG297 sequences available on 29.05.2019. The genomes were all assembled and quality checked (Chapter 2.6.7), the sequence with the highest N50 and smallest number of contigs was chosen to be the eBG297 reference, PHE_709.

2.5.3 Identifying Prophages in the eBG31 and eBG297 References

PHASTER (accessed 17.07.2019), with the metagenomic contigs option, was run on the references for eBG31 and eBG297 to identify putative prophages present in the reference genomes (Arndt *et al.*, 2016). For confirmation, the following key words were searched for in each putative prophage: capsid, head, integrase, plate, tail, fiber, coat, transposase, portal, terminase, protease, lysin using the gff output of *Prokka* (version 1.11) (Chapter 2.6.8) and *Artemis* (version 17.0.1) (Carver *et al.*, 2012; Seemann, 2014). The confirmed prophages that *PHASTER* reported as intact were then masked during phylogeny creation (Table 2.1).

eBG	*Prophage*	*Position*	*Length*
31	1	2551266-2583308	32Kb
31	2	3082238-3130649	48.4Kb
31	3	3387764-3415196	27.4Kb
31	4	3951065-3987469	36.4Kb
297	1	Node 26, 2405-45901	43.5Kb

Table 2.1 Prophages masked from eBG31 and eBG297 alignments
Prophages identified in the eBG31 and eBG297 reference genomes by *PHASTER* and chosen to be masked from the eBG31 and eBG297 soft-core alignments.

2.6 Bioinformatics Analysis Pipeline

All sequence data were put through a bioinformatics pipeline, illustrated in Figure 2.1.

2.6.1 Demultiplexing Illumina Output

For the first sequencing run, BaseSpace was unable to correctly demultiplex the sequence data (Illumina, 2019a). *bcl2fastq* was installed on *CLIMB* and successfully run with an edited version of the Sample Sheet created by BaseSpace (Connor *et al.*, 2016; Illumina, 2019b). For all other sequencing runs performed by the QIB Sequencing Facility, BaseSpace was used to demultiplex the Illumina output, generating 4 read 1 fastqs and 4 read 2 fastqs for each sample.

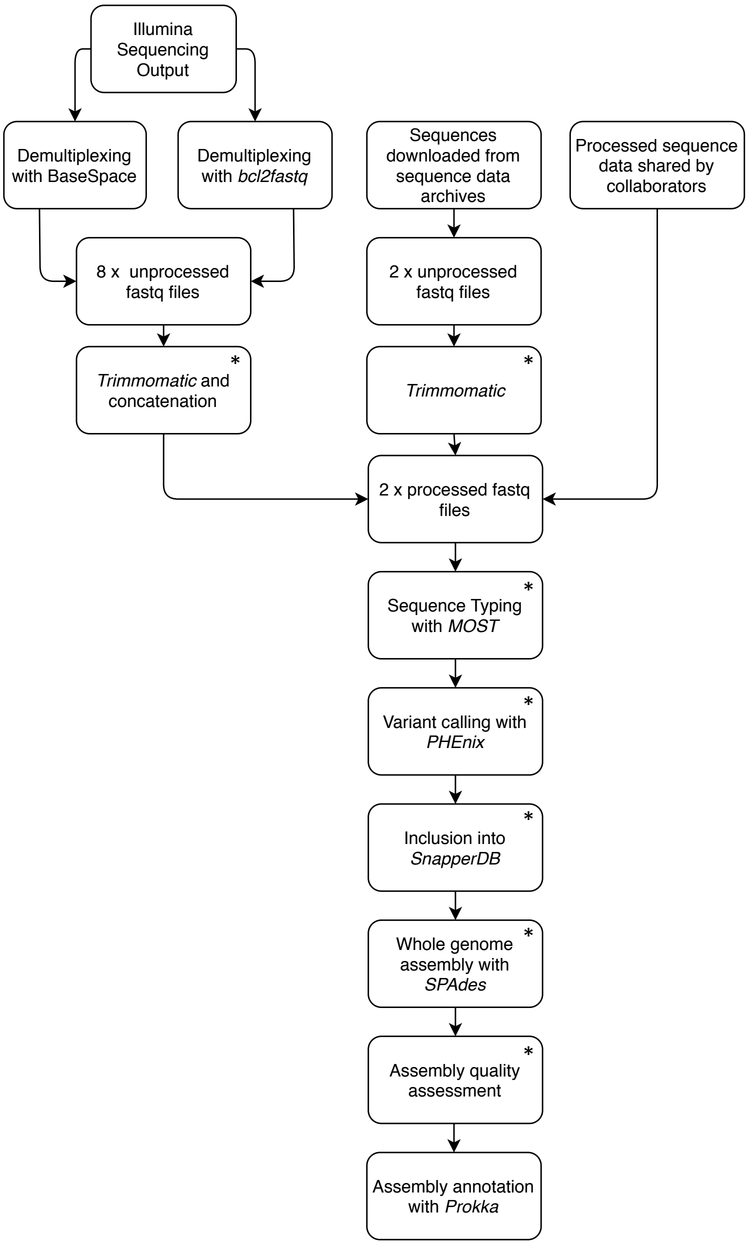

Illumina Sequencing Output
Demultiplexing with BaseSpace
Demultiplexing with *bcl2fastq*
Sequences downloaded from sequence data archives
Processed sequence data shared by collaborators
8 x unprocessed fastq files
2 x unprocessed fastq files
Trimmomatic and concatenation *
Trimmomatic *
2 x processed fastq files
Sequence Typing with *MOST* *
Variant calling with *PHEnix* *
Inclusion into *SnapperDB* *
Whole genome assembly with *SPAdes* *
Assembly quality assessment *
Assembly annotation with *Prokka*

Figure 2.1 Flowchart of bioinformatics pipeline
A flowchart illustrating the pipeline that all sequence data were run through. Steps marked with an asterisk are those that resulted in sequence exclusion from the next step, due to poor quality.

2.6.2 Sequence Data Quality Control

For all sequencing runs carried out locally, at least three sequences from each run were quality checked using *FastQC* (version 0.11.5) prior to trimming (Andrews, 2010). Poor quality data was trimmed using *Trimmomatic* (version 0.36) (Bolger, Lohse and Usadel, 2014). Initially trimming parameters were selected based on scripts used by Dr Lisa Crossman (UEA), PHE and from literature searches. *FastQC* was used to determine whether the more relaxed options (Dr Lisa Crossman) could be used or the stricter PHE options were needed. *Trimmomatic* was run with different combinations of the parameters they used, and the graphs produced by *FastQC* compared to determine the ideal parameters.

All sequences that had been downloaded or sequenced, as of November 2018, were trimmed using *Trimmomatic* with these parameters: ILLUMINACLIP:2:30:10 LEADING:30 TRAILING:30 SLIDINGWINDOW:4:15 MINLEN:50. ILLUMINACLIP trims the Illumina adaptors off the reads; LEADING trims bases at the start of reads that are below a certain quality; TRAILING does the same for the end of the read; SLIDING WINDOW trims once the average quality of the reads, in a defined window, falls below a quality threshold and MINLEN removes the read if it's under a specified length.

At this point, new trimming parameters had been validated at PHE: ILLUMINACLIP:2:30:10:8:true LEADING:30 TRAILING:30 SLIDINGWINDOW:10:20 MINLEN:50. All sequences were subsequently retrimmed with these new parameters except for the processed fastqs from PHE; these parameters were also used for new sequence data. For all the downloaded sequences, the paired output of *Trimmomatic* was used as the input for all software downstream in the pipeline. For any sequencing carried out locally, *Trimmomatic* was performed separately on the reads from each lane and the paired outputs for each read and lane were concatenated as the input for the next stage of the pipeline.

2.6.3 Sequence Typing

All of the isolates sequenced as part of my project had previously been serotyped and identified as *S.* Infantis. *MOST* (installed August 2017) was used in conjunction with the PHE *Salmonella* MLST database on all sequences to confirm that the isolates were *S.* Infantis and to determine their ST and eBG (Tewolde *et al.*, 2016; PubMLST, 2018). Originally the 2017 PHE database was used, but as all sequences needed retrimming and therefore the whole pipeline was to be re-run, an updated PHE *Salmonella* MLST database was acquired and used with MOST on all sequences (PubMLST, 2017; PubMLST, 2018). I wrote two bash scripts to filter and concatenate the results.xml files to easily compile the results for large numbers of sequences (Appendix I.1, I.2).

eBG	*Included STs*
31	32, 2283, 3277, 3756, 2146, 2780, 2937, 3815, 4196, 3870, 3854, 4366, 7759, 7760, 7761, novel1, novel2
297	603, 1823, 7731, 7732, novel3

Table 2.2 STs included in eBG31 and eBG297

STs that were defined by PHE or by Enterobase as belonging to eBG31 or eBG297

Table 2.2 lists the STs that were classed as belonging to eBG31 or eBG297. Any alternative STs identified were checked on Enterobase and then excluded from further analysis. Novel alleles were identified in 23 isolates; their sequences were uploaded to Enterobase to define the ST (ST7731, ST7732, ST7759, ST7760, ST7761). MOST also identified novel STs in both eBG31 and eBG297. These sequences were already in Enterobase and had been typed there as either ST32, ST2181 or ST1823; they were therefore classified as belonging to eBG31 or eBG297 but noted as belonging to a novel ST. The novel ST an SLV from ST32 was named novel1; the ST an SLV from ST2181 (eBG31) novel2 and the ST an SLV from ST603 novel3.

The MOST results were also used as a measure of quality; sequences that weren't of sufficient quality to pass through MOST were excluded. The traffic light parameter in the MOST results output was used initially to determine whether a sequence was good enough quality, green was marked as passed and red as failed. The QC max percentage non consensus base of each locus, an output of *MOST* that showed the highest percentage of non-consensus bases in each allele, was used to determine whether the sequences that came back as amber needed further checking or had passed. Sequences that had several alleles with a max percentage non consensus base score of over 15 were checked further. The mapping of the sequence to the alleles it had been typed as having was assessed using *Tablet* (version 1.19.05.28) (Milne *et al.*, 2013). Those with poor quality mapping were excluded.

2.6.4 Variant Calling

The software PHE use in their bioinformatics pipeline, *PHEnix* (version 1.3), was used to call the variants (Jironkin *et al.*, 2017). *PHEnix* mapped sequences to the relevant reference using *BWA* (version 0.7.12), called variants with *GATK* (version 3.8-0-ge9d806836) and filtered the VCF output with the following cut-offs: minimum depth of

10, mq score of 30 and ad ratio of 0.9 (Li, 2013; Van der Auwera *et al.*, 2013). Sequences that could not be variant called were excluded.

2.6.5 SnapperDB

SnapperDB was installed on a MacBook Pro and an eBG31 database was made using the PHE Illumina sequenced reference genome (Ashton *et al.*, 2017). The same filtering values as used in *PHEnix* were included in the config file, with the addition of average depth cut-off, 25. Again, *BWA* and *GATK* were used in the creation of the database with both given access to 4 threads.

Once the PHE reference isolate had been sequenced using long-read sequencing and polished (Chapter 2.7.1), it was used as the reference for a new eBG31 *SnapperDB* (version 1.0.6). The polished assembly was annotated using *Prokka* (version 1.13.3) (Chapter 2.5.7) and the filters and parameters used in the generation of the previous database were used in the creation of this database. With the initial database, Illumina fastqs were provided to the software to prevent ambiguous mapping; however, Dr Tim Dallman (PHE) advised that, for this version of the database, it would be better to not provide any fastqs, so the database was remade without fastqs. The eBG297 sequence that had been selected as reference, PHE_709, was used as the reference for the eBG297 database; the same filters and parameters were again used in creation of this *SnapperDB.*

The *S.* Infantis filtered variant call files were then added to their relevant eBG database. This step acted as another quality check as those with insufficient coverage did not successfully pass into the database. After sequences had been added to the database the update distance matrix function of *SnapperDB*, was run. In a database populated with a large number of sequences this became a very time-consuming process, with one particular update taking three weeks. Due to the time this was taking and the lack of back-up available during this time, *SnapperDB* was installed on *CLIMB* and a duplicate eBG31 database was created there.

The update clusters function was then run and any outliers identified were investigated by creating soft-core SNP alignments and comparing the placement of undetermined bases, which appeared as N's. Big blocks of N's were judged to be deletions or potential recombinations and sequences containing those were allowed into the database. Sequences with large numbers of N's dotted sporadically throughout the alignment were considered mixed samples and were excluded from the database. Any sequences that had over 10,000 variants from the reference were judged not to belong to that eBG and were excluded.

2.6.6 Whole Genome Assembly

SPAdes allows you to choose the length of the k-mers used to assemble the genome (Bankevich *et al.*, 2012). To determine the best k-mer combination, I trialled several different combinations: the default, no k-mer value specified; values used by Dr Gemma Langridge and PHE, 21, 33, 55 and 77; and variations of these k-mer sizes, 33, 55, 77, 99; 21, 33, 55, 77, 89; 21, 33, 55, 69. The contigs.fasta files were run through the *QUAST* website and the N50, Number of contigs and GC% were compared (Center for Algorithmic Biotechnology, 2013). The *SPAdes* fastg output for assemblies with the best quality scores were visualised using *Bandage* (installed March 2017) (Wick *et al.*, 2015).

The final method chosen to assemble the *S.* Infantis collection was *SPAdes* (version 3.13.0) with default parameters except for the k-mer selection of 21, 33, 55, 77 and the careful option to reduce mismatches and short indels.

2.6.7 Whole Genome Assembly Quality Assessment

Several different methods were used to assess the quality of the assemblies. *QUAST* (version 4.6.3) was run on all the assemblies as default but for no creation of plots (Gurevich *et al.*, 2013). As the quality of the whole assembly was of interest the minimum contig length was set to 1 and the scaffolds flag was not used as it breaks off every block of over 10 Ns's and treats that as another contig. The N50, number of contigs, longest contig and overall length were extracted from the report.

The percentages of mapped reads and properly paired reads were calculated by indexing the fasta outputs of *SPAdes* with *BWA* and mapping the trimmed reads to the assembly using *BWA-MEM.* The SAM output was sorted using *SAMtools* sort (version 1.5) and *SAMtools* flagstat was used to print the desired values (Li *et al.*, 2009). The coverage of the assembly was calculated using *SAMtools* depth on the output of *SAMtools* sort.

The results for all of these parameters were compared for the first 4,438 eBG31 scaffolded assemblies. Values for N50, longest contig, number of contigs, coverage, percentage mapped reads and percentage properly paired reads were plotted as individual histograms to determine their distribution. All of these graphs contained clear peaks containing the majority of the sequences; graphs of the distribution of the N50 and the number of contigs are provided as an example (Appendix II Figure II.1, II.2). The sequences that were included in this peak and in the direction that indicated good quality were identified; for N50, longest contig, coverage, percentage mapped reads and percentage properly paired reads the values higher than the peak were indicative of good quality. For the number of contigs, the values lower than the peak were good. For all graphs the peak and trailing end contained approximately 95% of the genomes. A sequence being amongst this 95% of sequences was therefore marked as good for that parameter. Table 2.3 shows the cut-offs for good quality. All scaffolded assemblies that were quality checked after the first 4,438 (which included all of those belonging to eBG297) were marked with these cut-offs. This process was repeated for all the contigs.fasta outputs of *SPAdes*; histograms were generated including all of the contigs.fasta quality results to identify cut-offs for these assembly files.

Quality Parameter	Range that indicated poor quality	
	scaffolds.fasta	contigs.fasta
N50	Less than 125,966	Less than 112,042
Longest contig	Less than 393,573	Less 349,737
Number of contigs	More than 358	More than 366
Coverage	Less than 35.2512	Less than 34.9924
Percentage mapped reads	Less than 99.86%	Less than 99.87%
Percentage properly paired reads	Less than 98.38%	Less than 98.34%

Table 2.3 Genome assembly quality cut-offs

Genome assembly quality parameters that were used to distinguish initially between the scaffolds.fasta and contigs.fasta assemblies that were good and poor quality.

As sequences that didn't have enough depth when mapped to the reference had already been excluded by *SnapperDB,* it was not expected that there would be any genomes that had poor coverage when the trimmed reads were mapped to the assembly. None of the sequences had a coverage of less than 25, the cut-off used by PHE, so coverage was no longer used as a quality parameter that could determine inclusion or exclusion of assemblies.

Many of the sequences had one 'bad' parameter, making it challenging to determine if that value was more indicative of poor quality than the five other 'good' values. Therefore, the values that weren't in the good 95% were further broken down for each parameter, with those just below the cut-off being graded less harshly. The results were compiled again and sequences with one 'bad' parameter that had been far from the cut-off were immediately excluded. Those that were close to the cut-off were included and those in between were judged on an individual basis.

2.6.8 Genome Annotation

Whole genome assemblies were annotated using *Prokka* (version 1.13.3). Firstly, the contig names were shortened using a Perl script provided by Dr Gemma Langridge (Appendix VI.1). *Prokka* was then run using 16 threads, with the genus specified as *Salmonella* and the options to use genus specific blast databases and searching for ncRNAs selected.

2.7 Further Bioinformatics Analyses

Alongside the bioinformatics pipeline, several other pieces of software were utilised. Those that were associated with a single chapter are described in the individual methods for each chapter; those that were used in more than one chapter are described here.

2.7.1 MinION Data Processing of the eBG31 Reference

Once the MinION run of the eBG31 reference was complete, basecalling and demultiplexing was performed using *Albacore* (Oxford Nanopore Technologies, version 2.0.1) with the MinION fast5 output and a fastq output.

Dr Gemma Langridge processed the raw fastq output. The fastq outputs were concatenated into one fastq file. *NanoFilt* (version 2.2.0) was used to trim and keep reads with a length of at least 10,000 bases and quality of 10; headcrop was also set to trim the first 50 bases (De Coster *et al.*, 2018). The quality statistics before and after trimming were compared using *NanoStat* (version 1.1.2).

2.7.1.1 *Assembly*

I assembled the trimmed fastq using *Canu* and *Unicycler* on *CLIMB* (Koren *et al.*, 2017; Wick *et al.*, 2017). *Canu* (version 1.7) was run with default settings except for useGrid=false which restricts it to running on the current machine. *Unicycler* (version 0.4.4) was run with default settings and the Illumina reads (SRR1968494) to generate a hybrid assembly. Both of these assembly methods resulted in an assembly with one contig, a marked improvement on the 105 contigs in the Illumina assembly. *MegaBLAST* (version 2.6.0) was used as default but for the parameter outfmt set to 6 to generate tabular crunch files; these were used as input for *ACT* (version 1.0) to compare the assemblies (McGinnis and Madden, 2004). Using the *Circlator* (version 1.5.3) all command to run the whole programme with 4 threads identified that the *Canu* assembly could be further processed to produce a single circular DNA sequence (Hunt *et al.*, 2015).

2.7.1.2 *Nanopolish*

Nanopolish (version 0.11.0) was installed on *CLIMB* and run on the *Canu* assembly using the Oxford Nanopore fastq and fast5 files (Loman, Quick and Simpson, 2015). The data was pre-processed using the *Nanopolish* index function, with the sequencing_summary.txt file generated during the MinION run. *BWA* was used to index the *Canu* assembly; the basecalled reads were aligned to the *Canu* assembly using *BWA-MEM* (version 0.7.17) with the flag -x ont2d for MinION data and *SAMtools* (version 1.8) was used to sort the reads. The *nanopolish_makerange.py* script broke the sequence into 50 kilobyte chunks to enable parallelisation. The output of this was used in *Nanopolish* consensus, with the methylation aware dcm option to account for the possibility that the sequence contained Dcm methylation motifs and the flag min-candidate-frequency 0.5, which gave the frequency a variant must be present to be extracted. The 50kb polished segment VCF files were converted into a single polished assembly with *Nanopolish* vcf2fasta.

2.7.1.3 *Pilon*

The polished assembly was polished again with the Illumina fastqs using *Pilon* (version 1.22) on *CLIMB* (Walker *et al.*, 2014). *BWA* (version 0.7.17) was used to index the assembly and *BWA-MEM* to align the Illumina reads to the assembly. *SAMtools* (version 1.8) was used to sort and index the aligned reads and *SAMtools* faidx to further index the assembly. Initial attempts to run *Pilon* resulted in failure, this was due to java using an excess of memory and solved by limiting the memory java had access to, to 16GB. *Pilon* was then successfully run with default settings.

2.7.1.4 *Racon*

The *Pilon* polished assembly was further polished using *Racon* (version 1.3.2) with the Illumina fastqs (Vaser *et al.*, 2017). *BWA* (version 0.7.17) and *SAMtools* (version 1.8) were again used to index the draft *Pilon* polished assembly. The Illumina read 1 was aligned to the draft assembly using *BWA-MEM* and *Racon* was run with default settings, polishing the assembly with the Illumina read 1. The output of this was indexed with *BWA* and *SAMtools* and the Illumina read 2 was aligned to it. *Racon* was then run again, polishing the assembly with the Illumina read 2.

Running *Circlator* with 4 threads and the all flag on the polished genome identified that it could no longer be processed to produce a single circular DNA sequence. The quality of the assemblies were evaluated using *MUMmer dnadiff* with default settings (version 1.3), with the value of interest in the report file being average identity of 1-to-1 alignment blocks (Kurtz *et al.*, 2004).

2.7.2 Phylogeny Creation using SnapperDB

Phylogenies of SNPs against the appropriate reference were generated using *SnapperDB.* All strains that successfully passed into each database were represented in the phylogenies. As the eBG297 database was much smaller than eBG31, all sequences were included in phylogenies generated. For all eBG297 alignments created, isolates up to 10,000 SNPs from the reference were included. For eBG31 alignments, isolates with 8,000 SNPs were included.

Approximately 200 sequences were required to generate whole genome assemblies. Due to the size of the eBG31 databases, representatives of clusters were used; a PHE Python script was amended and used to select a representative from the SNP cluster level that contained approximately 200 clusters (Appendix VI.2). Whole genome alignments were created of these cluster representatives or all eBG297 isolates. The alignments were used as input for *Gubbins* (version 2.3.1) which was run using 8 threads and the option to give the output a time-stamp and prefix (Nicholas J. Croucher *et al.*, 2015).

For the eBG31 databases the aforementioned PHE Python script was used to select a representative for each cluster from every SNP cluster level. An alignment was then made using the SNP cluster level that included the highest number of sequences that SnapperDB could successfully align. Soft-core SNP alignments of these sequences, and in the cases of smaller databases all the sequences, were then generated using *SnapperDB*, masking recombination and prophages using the output of *Gubbins* and the coordinates showing prophage presence from *PHASTER* respectively.

RAxML (version 8.2.12) was used to generate a maximum likelihood phylogeny of the soft-core SNP alignments using 8 threads (Stamatakis, 2014). The nucleotide substitution model, GTRCAT, was selected with parsimony inferences enabled and rapid bootstrapping to produce a best scoring maximum likelihood phylogeny in one program run, which was detected by the autoMRE option. The integers used as the random seed for the rapid bootstrapping and parsimony inferences option were 12345. For alignments with an outgroup sequence included, the outgroup name was given to *RAxML*.

2.7.3 Phylogeny Annotation

eBG297 phylogenies were annotated using the *iToL* colored strip file (Letunic and Bork, 2016). For eBG31 phylogenies, I used a Python script provided by Gemma Langridge, as a basis to write a script to identify which cluster each sequence belonged to so the number of isolates from each metadata group that were present in each cluster could be calculated (Appendix I.3). This information was converted to a percentage and then inputted into the *iToL* multi value bar chart annotation. The number of sequences within each eBG31 cluster and the fastbaps clusters for both phylogenies were added to the phylogenies using the *iToL* colored strip file. Clades on the phylogenies were rotated so clades in the same fastbaps clusters appeared together.

2.7.4 ARIBA Installation and Use

ARIBA (version 2.13.5) was installed on a MacBook Pro and the *ResFinder, PlasmidFinder* and *vfdb_full* databases were downloaded on 11.04.2019 using this version of the software (Zankari *et al.*, 2012; Carattoli *et al.*, 2014; Chen *et al.*, 2016; Hunt *et al.*, 2017). An additional database of the gyrase genes from *S.* Typhimurium LT2 was obtained from Dr Alison Mather. *ARIBA* (version 2.10.1) was installed on the UEA High Performance Cluster and all further *ARIBA* analyses were carried out on this platform. All ARIBA runs were performed with default parameters.

2.7.5 Genome Association Software Installation

Scoary (version 1.6.16) was installed and used on *CLIMB* (Brynildsrud *et al.*, 2016). The recursions depth was increased to 10,000 in the *Scoary* methods.py script to enable large numbers of sequences to be compared.

3. Chapter 3. Global Population Structure of *S.* Infantis

3.1 Introduction

Identifying the population structure of an *S. enterica* serovar can be beneficial as it provides information on the diversity within the serovar, and whether particular subgroups present a greater risk to human or animal health. For example, novel clades have been identified in *S.* Enteritidis, restricted to Africa and associated with invasive infection (Feasey *et al.*, 2016).

The *S.* Infantis population is comprised of two eBGs, eBG31 and eBG297, which are separated by 5 to 7 MLST alleles (Gymoese *et al.*, 2019). eBG31 is the dominant eBG, making up the majority of cases (M.A. Chattaway, personal communication, 26th May, 2017).

Some comparisons of small numbers of *S.* Infantis sequences have been performed. Clustering by geography has been reported in *S.* Infantis, the largest study compared 21 *S.* Infantis isolates from chicken meat in Turkey and 243 sequences from other continents including North America, Africa, South America and identified that the Turkish sequences formed a distinct clade (Acar *et al.*, 2019). However, a paper published more recently, including 100 strains from five continents, found no evidence of clustering by geography (Gymoese *et al.*, 2019). One cluster in this phylogeny was mainly comprised of human and poultry isolates, the authors concluded that this was a clone of *S.* Infantis that was adapted to humans and poultry. Clustering by isolation source has also been identified in *S.* Infantis; one study included 67 strains from Japanese broilers, eggs and humans and found that the phylogeny split into 5 clusters that were associated with the source of infection (Yokoyama *et al.*, 2014).

3.1.1 Collaborators in this Project

The strains and sequences used in this project were collected from several different data sources and collaborators.

The GBRU is the PHE *Salmonella* reference laboratory for England and Wales. All *Salmonella* that are identified from patients reporting to their general practitioners or to hospitals are sent to the GBRU. All the isolates are serotyped and as of 2015 every *Salmonella* they receive is whole genome sequenced.

APHA is the government agency that protects the health of animals and plants in Great Britain (Animal and Plant Health Agency, 2017). Upon detection of threat to the health of animals, including *Salmonella*, their role is to characterise and assess the threat, which includes visits to farms and testing samples, and to communicate this information to policy makers. Any *Salmonella* that are reported are serotyped; whole genome sequencing is currently only carried out for research projects.

The NICD is a public health institute in South Africa that monitors threats to public health due to communicable diseases. *Salmonella* samples are sent from regional laboratories across South Africa to the Centre of Enteric Diseases department which acts as a reference laboratory. All sequences are serotyped but currently whole genome sequencing is not carried out routinely.

Due to the collaboration with PHE in this project, several of their bioinformatic pipelines were used to enable PHE to use the project outputs. It was for this reason that SnapperDB was used, along with the PHE sequence read quality cut-offs (Ashton *et al.*, 2017). SnapperDB calculates and stores the distances between all sequences added to the database (Ashton *et al.*, 2017). It clusters the sequences on seven levels of SNP distance: 250, 100, 50, 25, 10, 5 and 0. For a sequence to be added to a cluster it needs to be within that SNP distance of any isolate in the cluster. The clusters that each sequence belongs to are used to give it a seven-digit code. This SNP address is then used to provide real-time clustering of sequences and identify outbreaks. The GBRU team have, after validation, defined a *Salmonella* outbreak as sequences belonging to the same 5SNP cluster.

3.1.2 Aims and Objectives

Previous research has shown instances of clustering by source or geographical location within *S.* Infantis. However, the global diversity of *S.* Infantis is currently unknown, as is whether clustering by geography, source or year of isolation is seen in large groups of sequences.

The aims and objectives of this chapter were therefore to:

- Determine the genetic diversity within *S.* Infantis and how it varies by isolation source, year and origin
- Identify any difference in population structure between eBG31 and eBG297
- Calculate the genetic distance between sub-groups of *S.* Infantis

3.2 Specific Methods for Determining Population Structure

3.2.1 Metadata Distribution of *S*. Infantis

The classification of the different metadata categories is detailed in Chapter 2.1.5. The percentage of isolates belonging to each metadata group was calculated using all sequences in the eBG31 and eBG297 collections. In figures North America is referred to as N.America and South America as S.America.

3.2.1.1 *Isolation Source Distribution*

The sequences belonging to the human, poultry and environmental source groups were stratified into sub-groups of interest. The human isolates were grouped into faeces, blood and urine; the poultry isolates into chickens, chicken meat, eggs, duck, turkey and quail; and the environmental group into cattle, pigs and animal feed. Appendix II Table II.3 contains the key words used to stratify isolates into subgroups.

3.2.1.2 *ST Distribution*

The allelic profiles for all sequences in the *S.* Infantis collection were imported into *PHYLOViZ,* generating a minimum spanning tree of MLST alleles within *S.* Infantis (Ribeiro-Gonçalves *et al.*, 2016).

Some STs within eBG31 were grouped, allowing clearer annotation of the phylogenies. ST32, ST2146 and ST2283 were maintained as individual ST's. The following STs were grouped and named Other: ST2780, ST2937, ST3277, ST3756, ST3815, ST3854, ST3870, ST4196, ST4366, ST7759, ST7760, ST7761, novel1 and novel2. The STs within eBG297 were not grouped.

3.2.1.3 *Phylogeny Creation for eBG31 and eBG297*

The reference sequences for each eBGs *SnapperDB* were added to the other eBG database to be used as an outgroup; in the case of the eBG31 reference, the Illumina short-read sequenced version of the genome was used.

The method used to generate the phylogenies is detailed in Chapter 2.7.2. Representatives of each 50SNP and 25SNP cluster were used when generating the eBG31 whole and soft-core genome alignments respectively. The soft-core alignments for both eBGs were made twice, with the inclusion of the outgroup in one, requiring the permitted number of SNPs from the reference to be increased to 30,000.

The phylogenies containing the outgroup were each rooted to the outgroup. The most ancestral node in those phylogenies was identified and used as the root in the phylogenies generated without the outgroup.

3.2.1.4 *Identifying Clusters Within the Phylogenies*

Initially *hierBAPS* (installed 30.07.2019) with *MATLAB* (version 8.4) and *rhierBAPS* (version 1.1.2) with *R* (version 3.4.4) and *ape* (version 5.3) were used on *CLIMB* (Cheng *et al.*, 2013; MathWorks, 2014; Connor *et al.*, 2016; Paradis and Schliep, 2018; R Core Team, 2018; Tonkin-Hill *et al.*, 2018). While the eBG297 phylogeny was successfully analysed with both hierBAPS and rhierBAPS, the eBG31 phylogeny was too large for both pieces of software.

An alternative *R* package, *fastbaps* (version 1.0.0) was installed using the package *devtools* (version 2.0.1) and used with *R* (version 3.5.1) and *ape* through *R Studio* (version 1.1.463) (RStudio, 2018; Wickham *et al.*, 2018). This was run with the eBG31 and eBG297 soft-core alignments, with default settings and with the variance of the Dirichlet prior determined by *fastbaps* at 0.009. *Fastbaps* was also used to calculate and produce a heatmap of the bootstrap results of the clustering. These results were exported and the location of the sequences with low bootstrap values identified on the phylogeny using *iToL's* colored strip file utility.

3.2.1.5 *Calculating the Distance Across the Phylogenies*

The median pairwise SNP distances were calculated within and between the sequences in each phylogeny when grouped by continent, source, year, ST and fastbaps cluster. For the eBG31 phylogeny, a representative sequence was labelled with the metadata of all of the sequences it represented, for example, if a 25SNP cluster contained African and European sequences it was labelled as 'AfricaEurope' and was included in the comparisons for both of those continents. When calculating the distance between different metadata samples the clusters that contained both states were included with a distance of 0.

MEGA7 (version 7180411-i386) was used with the soft-core alignment used to create the phylogeny; the pairwise distances matrices were created with default settings but for the Model/Method which was changed to 'No. of differences' (Kumar, Stecher and Tamura, 2016). The distance matrices were exported and the median and range calculated within and between each metadata type. Box plots of the results were generated using *R* and *RStudio*.

3.2.2 SNP Cluster Analysis

The SNP clustering results were downloaded from *SnapperDB* (version 1.0.6) for eBG31 and eBG297 (Ashton *et al.*, 2017). The number of clusters in each SNP cluster level, number of isolates in each cluster and the frequency of each cluster size were calculated. Sequences belonging to outbreaks were defined as members of 5SNP clusters that contained more than 1 sequence. Sporadic cases were sequences that were the only member of a 5SNP cluster.

All sequences belonging to low frequency 100SNP clusters were annotated on the relevant eBGs soft-core SNP phylogeny to identify their location using *iToL* and the *iToL* color strip annotation file.

3.2.3 Non-Alignment Based Distance Calculation

The contigs.fasta output of *SPAdes* (Chapter 2.6.6) was uploaded to *CLIMB* for all good quality eBG31 and eBG297 sequences (Chapter 2.6.7) (Bankevich *et al.*, 2012). The Illumina short-read version of the eBG31 reference was used. The assemblies were renamed to include either their eBG, isolation source or continent of isolation. *Mash* (version 2.1.1) sketch was run on the assemblies using 4 threads and with a sketch size of 10,000 (Ondov *et al.*, 2016). *Mash* dist was run on the output of this with 4 threads, outputting the distances between all of the assemblies in the comparison.

Awk and sed commands were used to pull out the sequence name, metadata type and *Mash* distance from the *Mash* output. *RStudio* (version 1.2.1335) and *R* was used with the *R* package *data.table* (version 1.11.8) to import the distance matrices, ensuring that the correct number of results were in the output file (Dowle and Srinivasan, 2018).

Summary statistics for each file were then exported from *R*. Box plots were generated using *R*, *RStudio* and the package *ggplot2* (version 3.1.0) (Wickham, 2016). Significance between the *Mash* distances of the human/poultry and human/environmental comparisons and the comparisons between and within eBG31 and eBG297 were tested for using the Mann-Whitney U test with *R*, *RStudio* and the package *data.table*.

3.3 Results

S. Infantis sequence data was downloaded from GenBank, Enterobase, DDBJ and ENA. Data was also shared by PHE and the APHA. DNA was extracted from *S.* Infantis isolates by the NICD, the APHA and PHE.

3.3.1 Metadata of *S.* Infantis

The number of *S.* Infantis isolates from different continents, sources, and year groups was calculated and the results for each eBG compared (Appendix VI Table VI.1).

3.3.1.1 *Overall Number of S. Infantis Included in This Study*

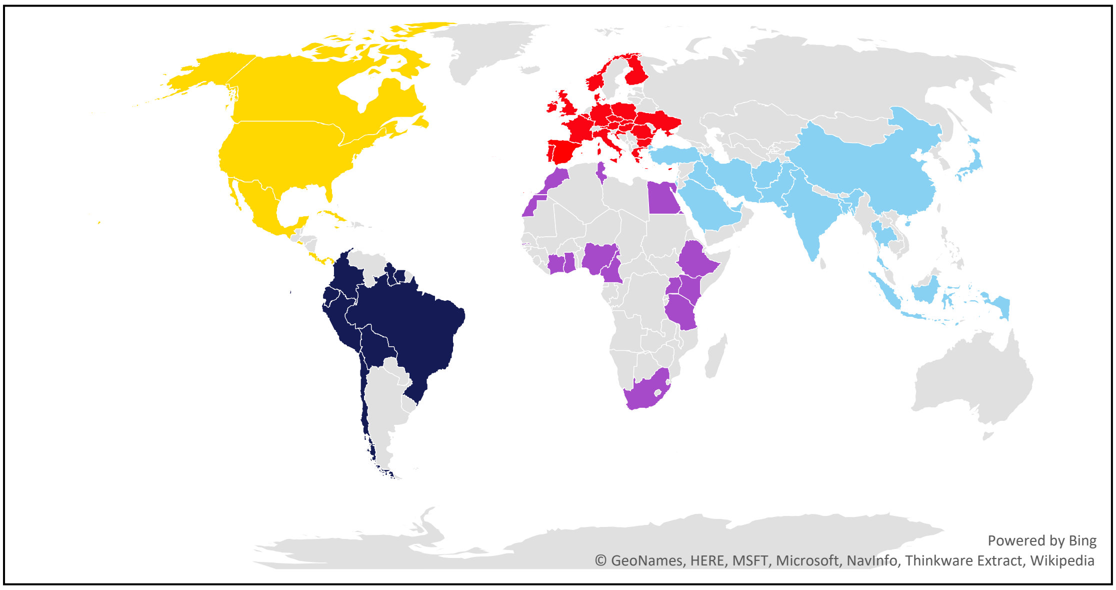
All DNA extracted and sequenced for this project was from isolates that had been serotyped as *S.* Infantis; however, many were sequence typed and found not to belong to either eBG31 or eBG297. Of the 4,739 sequences that were *S.* Infantis, 69 were not of sufficient quality to pass into SnapperDB and were excluded from the phylogenies and SNP-based clustering analyses (Table 3.1). A further 100 had low quality assemblies and were excluded from the non-alignment based distance calculations.

Collaborator	**No. sequenced**	**No. Infantis**	**No. eBG31**	**No. eBG31 in db**	**No. eBG297**	**No. eBG297 in db**
PHE	653	653	620	620	33	31
hPHE	192	181	173	173	8	8
APHA	128	124	122	122	2	2
APHA online	62	56	56	54	0	0
NICD	450	403	273	273	130	129
eBG31 Enterobase	3400	3301	3301	3244	0	0
eBG297 enterobase	21	21	0	0	21	14

Table 3.1 Number of *S.* Infantis strains and sequences from each data source and collaborator. The number of isolates that had DNA extracted; the number of sequences shared, downloaded or generated; the number of these that were *S.* Infantis from each of the collaborators/data sources and the number that belonged to either eBG and successfully passed into either eBGs SnapperDB (db).

3.3.1.2 *Global distribution of S. Infantis*

The strains included in this study originated from 70 different countries which were grouped by continent (Figure 3.1).



Powered by Bing
© GeoNames, HERE, MSFT, Microsoft, NavInfo, Thinkware Extract, Wikipedia

Figure 3.1 *S.* Infantis countries of isolation
S. Infantis sequence data from 70 countries were included. Generated using Microsoft Excel's map function.

Africa 🟪 Asia 🟦 Europe 🟥 N. America 🟨 S. America ⬛ Unknown □

The number of strains included from each continent is represented in Figure 3.2. Thousands of sequences were uploaded to Enterobase by the United States government: 43.2% (1207/2795) of the North American sequences were uploaded by the Centre for Disease Control and 25.2% (703/2795) by the Food and Drug Administration.

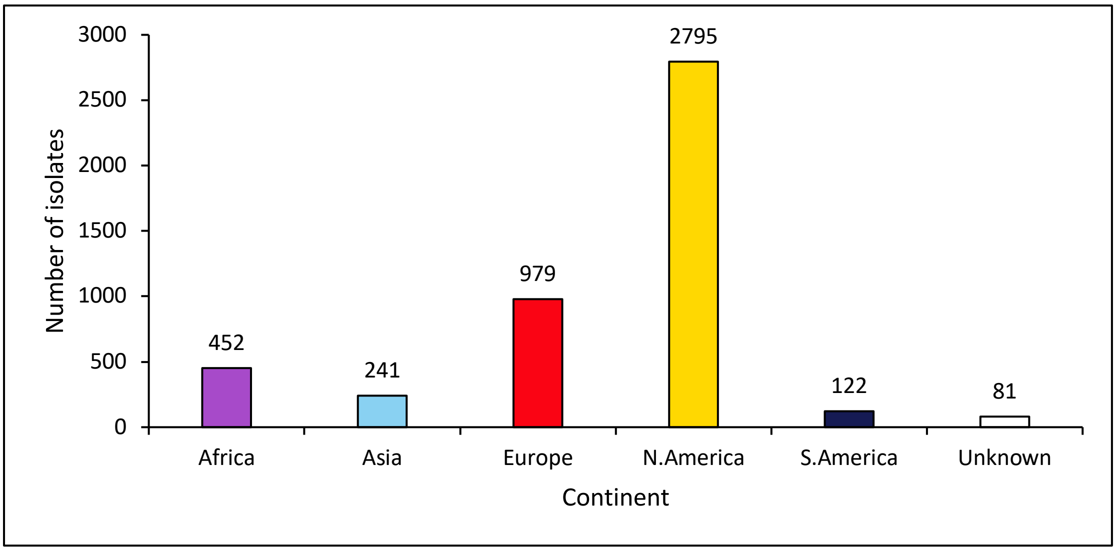


Figure 3.2 Number of *S.* Infantis isolates from each continent.

Africa 🟪 Asia 🟦 Europe 🟥 N. America 🟨 S. America ⬛ Unknown □

3.3.1.3 *Isolation Source Distribution of S. Infantis*

S. Infantis was isolated from a wide variety of different sources. These were grouped into four categories: human, poultry, environmental and unknown (Figure 3.3).

Of the *S.* Infantis isolates from a human source, 74.9% (1264/1687) were found in faeces or rectal swabs, 3.7% (63/1687) in blood and 5.0% (84/1687) in urine.

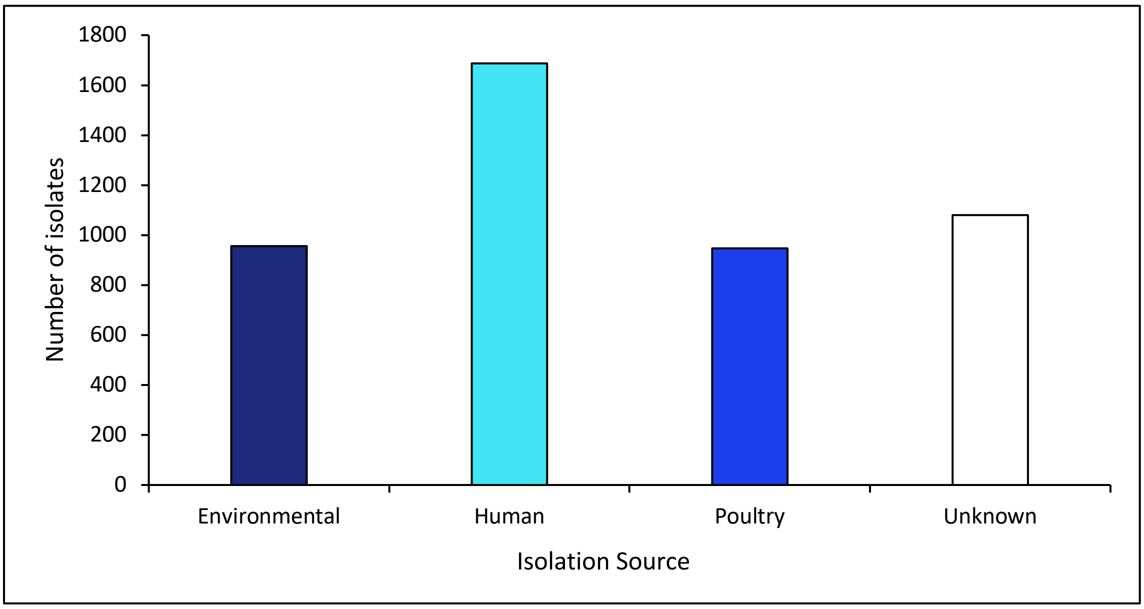


Figure 3.3 Distribution of isolation sources.
Environmental ■ Human ■ Poultry ■ Unknown □

The isolates from poultry sources, when stratified, included 21.6% (205/947) from chickens, 70.8% (671/947) from chicken meat, 1.2% (11/947) from eggs, 2.3% (22/947) from duck, 1.8% (17/947) from turkey and 0.4% (4/947) from quail.

The isolates from environmental sources represented all non-human and non-poultry sources. *S.* Infantis was frequently identified in livestock, food and animal food. 30.9% (295/956) of the environmental isolates were found in pigs, 14.1% (72/956) in cattle and 7.2% (69/956) in animal food. *S.* Infantis was also identified in several other animals including horses, dogs, camels and lizards.

3.3.1.4 *Year Distribution of S. Infantis*

Strains were isolated over three decades; the oldest in 1989 and the most recent in 2019. Figure 3.4 shows the number of strains isolated from each year. Due to the range of years covered, isolates were grouped into time periods that would represent at least 3% of the total number of sequences.

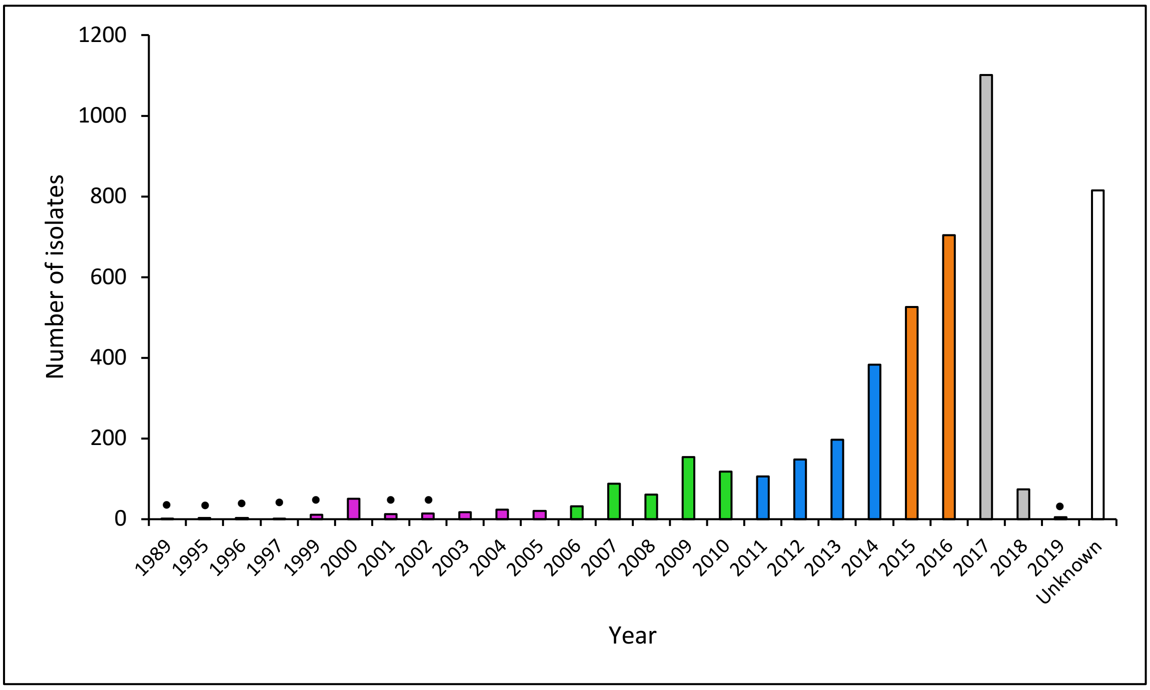


Figure 3.4 Number of *S.* Infantis isolates per year

● Between 1 and 15 sequences were isolated from that year group

1989-2005 🟪 2006-2010 🟩 2011-2014 🟦
2015-2016 🟧 2017-2019 ⬜ Unknown □

The number of isolates sequenced each year increased in-line with the decreased cost of whole genome sequencing. The number dropped post 2017 as eBG31 sequences were no longer included from online databases after February 2018. The temporal distribution of *S.* Infantis isolates from each source and continent was determined (Appendix III Figure III.1,III.2). A small peak in human cases was observed in 2000 and 2009, associated with a group of cases from Japan and South Africa respectively.

3.3.1.5 *Sequence Type Distribution of S. Infantis*

A minimum spanning tree of all the STs identified in the *S.* Infantis sequences was generated, illustrating the number of SLVs between each ST (Figure 3.5). The predicted founder STs of eBG31 and eBG297, ST32 and ST603 respectively, share no alleles, with the closest related STs, ST32 and ST1823 sharing only 1.

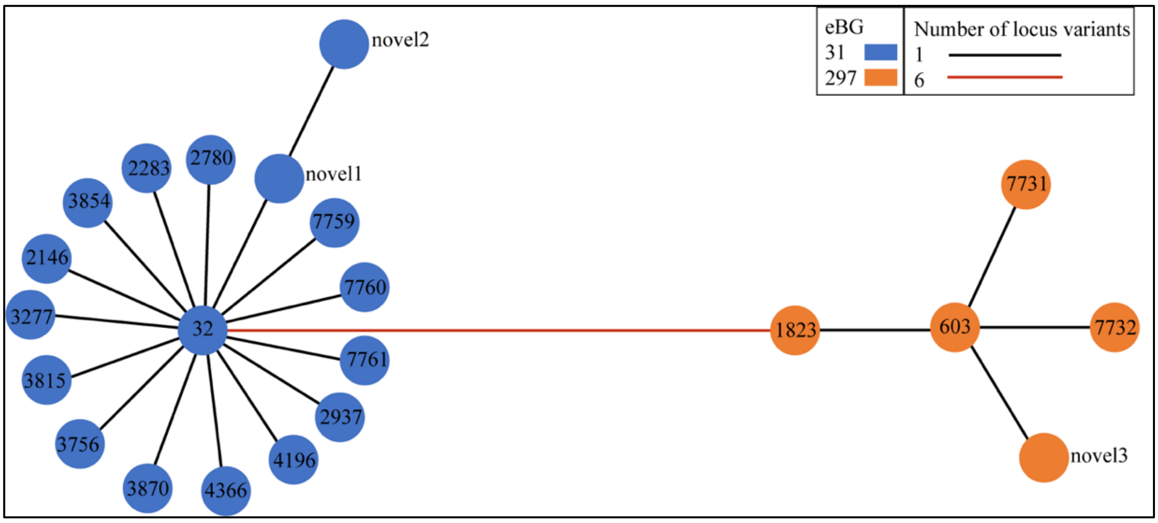


Figure 3.5 Minimum spanning tree of *S.* Infantis MLST alleles.
The number of locus variants between the *S*. Infantis STs are plotted. 6 of the 7 MLST alleles were different between ST32 and the closest eBG297 ST, ST1823.
eBG31 🟦 eBG297 🟧

The majority of the *S.* Infantis sequences, 96.1% (4486/4670), belonged to eBG31; eBG297 only accounted for 3.9% (184/4670). The distribution of the STs in eBG31 and eBG297 was calculated (Appendix III Figure III.3). In the eBG31 population ST32 was the dominant ST, accounting for 98.2% (4406/4486) of the sequences. ST2283 and ST2146 were the next most frequently identified, responsible for 33 and 26 of the cases respectively. The other STs were all identified less than 5 times.

In eBG297 there was also a dominant ST, ST603, which comprised 85.9% (158/184) of the sequences. Two novel alleles were discovered, which when typed were named ST7731 and ST7732; they were identified 8 and 12 times respectively, all from South Africa.

The years of isolation of eBG31, eBG297 and the STs within both were plotted (Appendix III Figure III.4, III.5). The earliest eBG31 isolate was from 1989 and the earliest eBG297 isolate from 2003. Both ST32 and ST603 were present in all year groups; ST7731 and ST7732 were identified in every year group but 2017-2019, correlating with when the South African strains were isolated. The majority of the other eBG31 STs were only identified in one year group.

3.3.1.6 *Association between Geography, Source and eBG*

The proportion of isolates from each source group varied between continents (Figure 3.6).

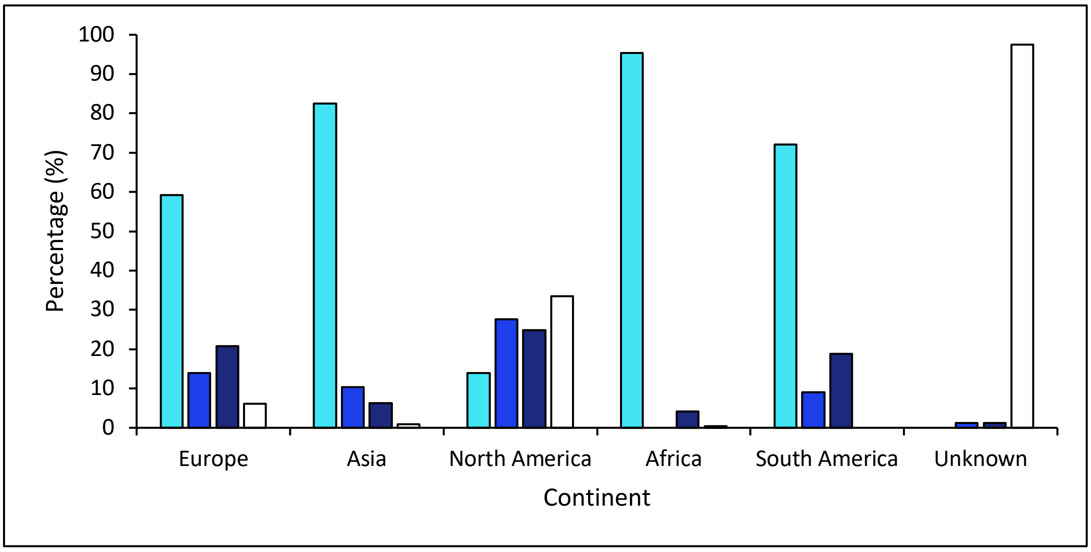


Figure 3.6 Source distribution within each continent
Percentage of *S*. Infantis isolates from each continent that were isolated from each source.

Environmental ■ Human ■ Poultry ■ Unknown □

In all continents but North America, the majority of *S.* Infantis strains were isolated from humans. In North America, poultry and environmental sources accounted for the majority of the isolates. However, in Europe, Africa and South America, more isolates had been sequenced from environmental sources than poultry sources.

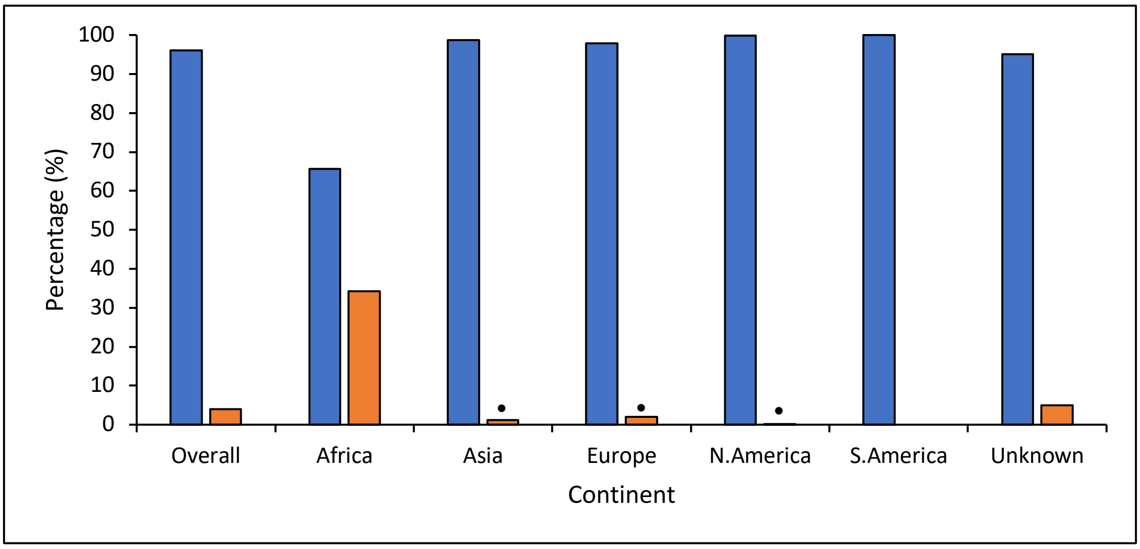


Figure 3.7 Distribution of eBG per continent
Africa n=452, Asia n=241, Europe n=979, N.America n=2795, S.America n=122, Unknown n=81. • Values are greater than 0 and less than 3%
eBG31 ■ eBG297 ■

Figure 3.7 demonstrates a clear difference in continent distribution by eBG. Whilst eBG31 was present in high numbers in every continent, eBG297 was more common in Africa, with 84.2% (154/183) of all eBG297 isolates being either isolated from there or from someone who had travelled there. Converse to the overall global eBG distribution, in Africa 34.1% (154/451) of the sequences belonged to eBG297.

The percentage of isolates from each eBG that had been isolated from each source group was calculated (Figure 3.8). Whilst similar numbers of sequences were present for each source type in the eBG31 population this was not seen in eBG297. The majority, 91.3% (168/184), of sequences were isolated from humans, with no sequences being isolated from poultry.

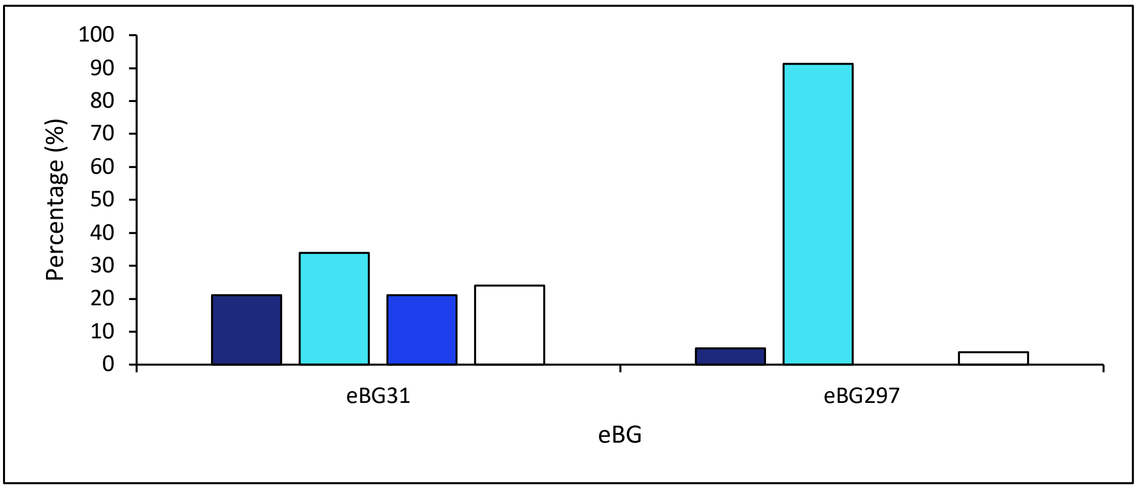


Figure 3.8 Distribution of eBG by source
eBG31 n=4486, eBG297 n=184

Environmental ■ Human ■ Poultry ■ Unknown □

The distribution of STs across each continent and source was also assessed to determine whether some were more prevalent in different locations (Appendix III Figure III.6, III.7). Whilst the predominant ST in eBG31, ST32, was identified from every continent, all other STs, excluding those with an unknown origin, were isolated from a single continent. In eBG297, ST603 was identified in all continents but South America. ST1823, despite its low frequency, was also identified in both North America and Africa. A similar pattern was observed in ST distribution across isolation sources; ST32 was identified in all source groups and ST603 in human and environmental sources. The majority of the other STs from both eBGs were identified exclusively from one source group, although in low numbers, with humans being the most common source.

3.3.2 Phylogenetic Structure of the eBG31 Population

In order to identify the genetic diversity within eBG31, a soft-core SNP phylogeny was generated masking recombination and prophages, including a representative from each 25SNP cluster in eBG31 (Figure 3.9). *Fastbaps* was then used to determine the structure of the phylogeny.

Single representatives of clusters were used as SnapperDB could not handle the entire set of eBG31 sequences. When determining associations across the phylogeny, data for all isolates within the representative clusters were taken into account.

831 25SNP cluster representatives were present in the phylogeny. The inner ring on Figure 3.9 displays the number of sequences represented by each leaf on the tree. 64.5% (n=536) of the leaves were the only member of the 25SNP cluster they represented, 15.9% (n=132) represented clusters containing two sequences and 10.6% (n=88) represented clusters containing between 3 and 5 sequences. However, there were also clusters containing large number of sequences, with 7 containing over 50 sequences. The two largest clusters contained 709 and 874 sequences.

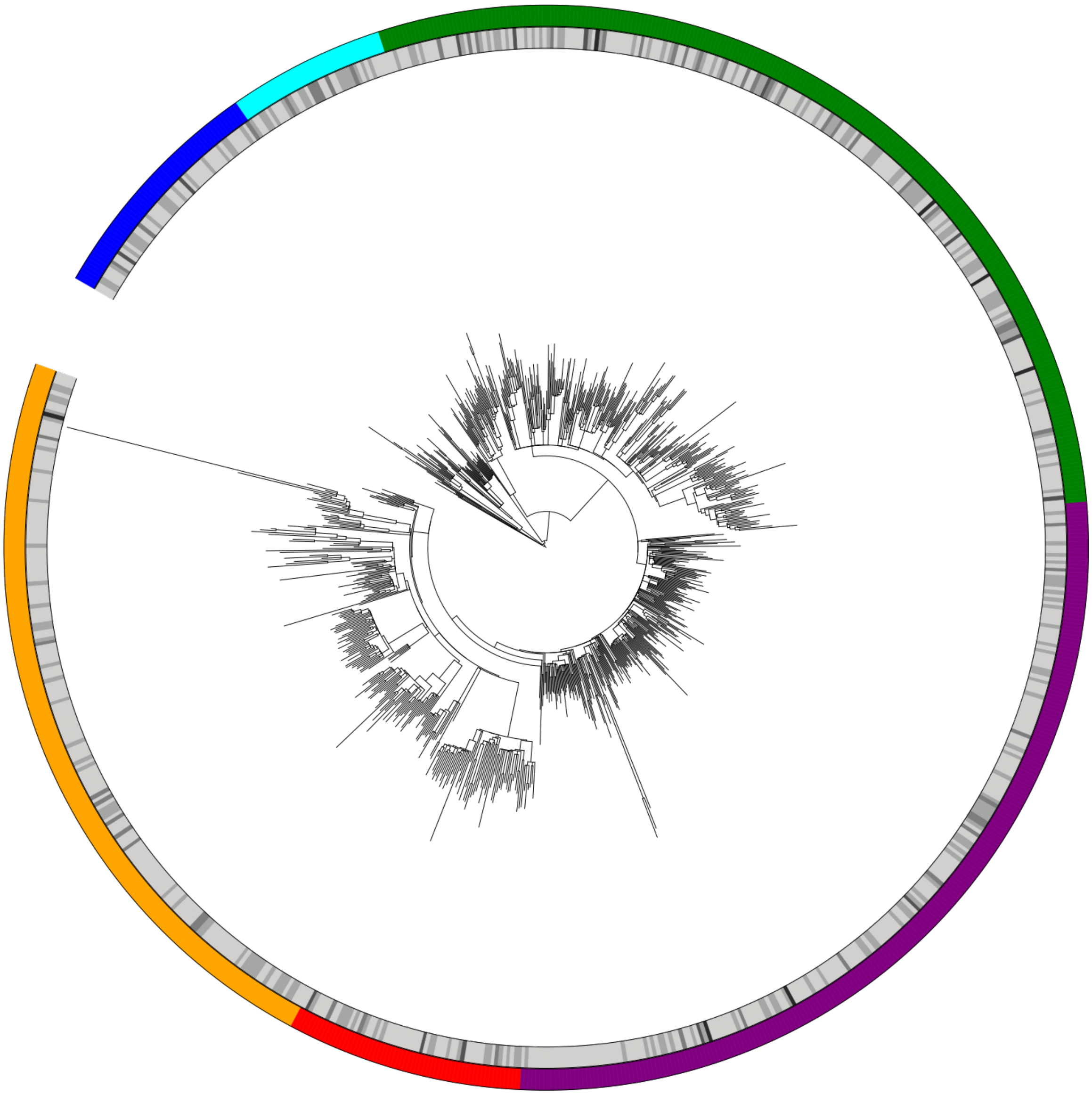


Figure 3.9 eBG31 phylogeny annotated with fastbaps cluster
Soft-core SNP Maximum Likelihood phylogeny of 831 25SNP cluster representatives of eBG31.

Inner ring, Number of sequences in 25SNP cluster: ■ 1 ■ 2-5 ■ 6-20 ■ 21-50 ■ >50

Outer ring, fastbaps cluster: 1 🟥 2 ■ 3 🟦 4 🟪 5 🟧 6 🟩

An electronic version of the figure is available in Appendix VI Figure VI.1

3.3.2.1 *Clusters within the eBG31 phylogeny*

6 clusters were identified within the eBG31 phylogeny using *fastbaps* (outer ring of Figure 3.9), containing varying numbers of sequences (Table 3.2) (Appendix VI Table VI.2).

Cluster	Number of 25SNP clusters	Total number of sequences
1	60	207
2	40	121
3	60	187
4	232	1298
5	194	1209
6	245	1463

Table 3.2 Number of sequences in each eBG31 fastbaps cluster
The number of 25SNP cluster representatives in each fastbaps cluster and the total number of sequences they represented.

The bootstrap values of the identified clusters were viewed to check the stability of the clusters (Appendix III Figure III.8). Clusters 1, 2 and 3 had bootstrap values of 100%. Clusters 4, 5 and 6 had the majority of their bootstrapping results at 100% but also contained sequences with lower values. For Clusters 4 and 5 these sequences were either on very long branches or closely related to sequences in other clusters; however, upon measuring branch length they were more closely related to the node they had been assigned by the software. Cluster 6 contained 2 large monophyletic clades and all the low bootstrapping results for this cluster were between these two clades. The clusters predicted by *fastbaps* were therefore deemed robust.

The median pairwise SNP distances within and between fastbaps clusters were calculated (Appendix III Figure III.10). The lowest distance was observed within Cluster 4 (78, range 15-256) and the highest within Cluster 3 (160, range 13-275). The distance between the fastbaps clusters was in general larger than the distance within the clusters, with the largest seen between Cluster 1 and 3 (301, range, 251-429).

Appendix III Figure III.11 depicts the phylogeny as a cladogram to allow for visualisation of the structure of the tree. The tree is rooted to the most ancestral node. Closely related to this are two other ancestral groups, one of which shares a common ancestor to the rest of the *S.* Infantis population (Figure 3.9, Cluster 2 and 3). The majority of the sequences are closely related to other sequences in the phylogeny; however present in each cluster is at least one sequence with a noticeably long branch length.

3.3.2.2 *Sequence Type Distribution Across the eBG31 Phylogeny*

The ST distribution across the eBG31 soft-core phylogeny was identified (Appendix III Figure III.12). ST32, the most prevalent ST in eBG31, was distributed throughout the phylogeny. All of the sequences belonging to the next most prevalent STs, ST2146 and ST2283, were within a single representative sequence for either ST. The other STs were distributed throughout the phylogeny, making up 50% of a cluster with ST32 in 2 cases. Where there was more than one instance of an ST from the 'other' group, they also were within a single representative cluster on the phylogeny, either alone or with ST32 isolates.

3.3.2.3 *Year Distribution Across the eBG31 Phylogeny*

The eBG31 soft-core SNP phylogeny was annotated with the year group of isolation to identify any temporal correlation with the structure of the phylogeny (Figure 3.10).

Every fastbaps cluster contained sequences from every year group but for 1989-2005, which was not present in Cluster 1 (Appendix III Figure III.9). The most common year group did vary by cluster; Clusters 1 and 2 were mainly comprised of sequences from 2015-2016 making up 48.8% (101/207) and 35.5% (43/121) of sequences in Clusters 1 and 2 respectively. Also, the majority of sequences (46%, 557/1209) in Cluster 5 were isolated in 2017-2018.

The median pairwise SNP distance within and between year groups was calculated to determine whether strains isolated from a year group were distinct from strains from other time points (Appendix III Figure III.13). The distance within the year groups did not vary greatly, the medians all fell between a range of 125-181. There was also little variation of distance between year groups.

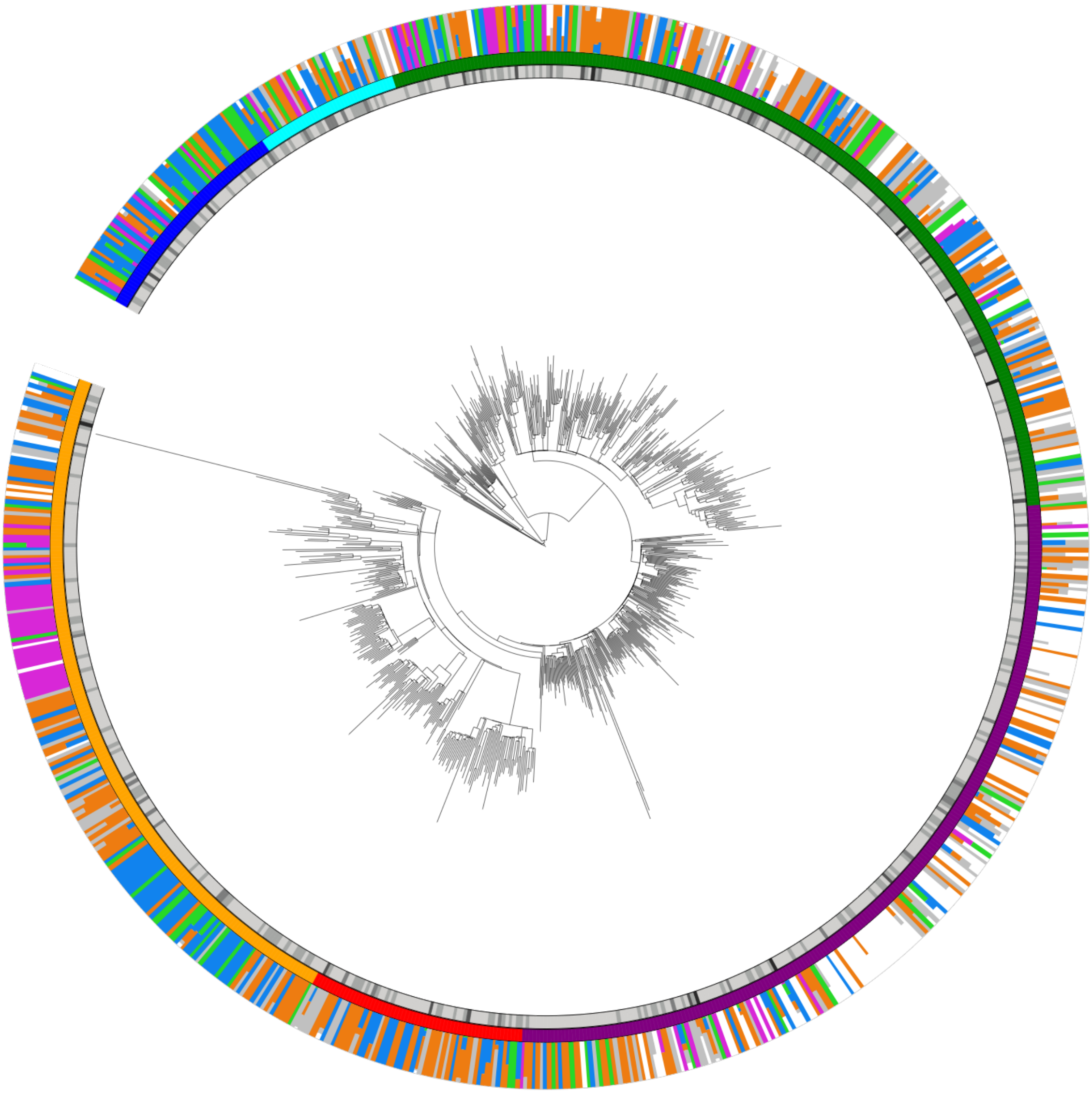


Figure 3.10 eBG31 phylogeny annotated with year

Soft-core SNP Maximum Likelihood Phylogeny of 831 25SNP cluster representatives of eBG31. The outer ring is annotated with the percentage of isolates in each 25SNP cluster that were isolated from each year group.

Inner ring, Number of sequences in 25SNP cluster: ⬜ 1 ⬜ 2-5 ⬛ 6-20 ⬛ 21-50 ⬛ >50

Middle ring, fastbaps cluster: 1 🟥 2 🟦 3 🟦 4 🟪 5 🟧 6 🟩

Outer ring, Year Group: 1989-2005 🟪 2006-2010 🟩 2011-2014 🟦 2015-2016 🟧 2017-2018 ⬜ Unknown ⬜

3.3.2.4 *Source Distribution Across the eBG31 Phylogeny*

In order to identify how isolates from different sources were distributed over the eBG31 soft-core phylogeny it was annotated with source group (Figure 3.11). There were many 25SNP clusters containing sequences that were isolated from different sources. However, there was also several examples of sequences clustering by isolation source. Small clusters of isolates from poultry or environmental sources were seen but the most striking clustering was seen with the human isolates.

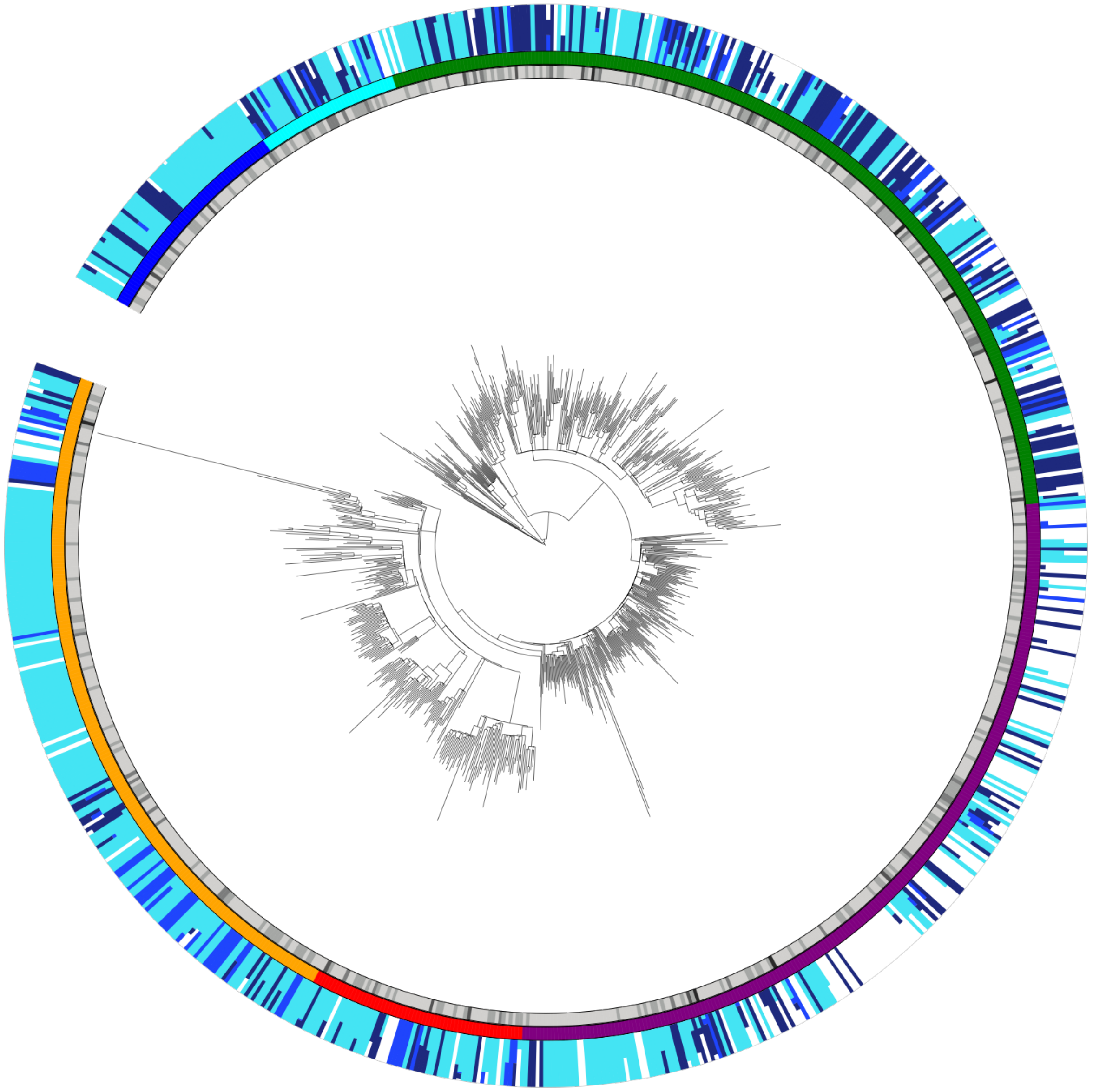


Figure 3.11 eBG31 phylogeny annotated with source
Soft-core SNP Maximum Likelihood Phylogeny of 831 25SNP cluster representatives of eBG31. The outer ring is annotated with the percentage of isolates in each 25SNP cluster that were isolated from each source group.

Inner ring, Number of sequences in 25SNP cluster: ■ 1 ■ 2-5 ■ 6-20 ■ 21-50 ■ >50

Middle ring, fastbaps cluster: 1 🟥 2 🟦 3 🟦 4 🟪 5 🟧 6 🟩

Outer ring, Isolation Source: Environmental ■ Human 🟦 Poultry 🟦 Unknown □

The most common source group varied between fastbaps clusters (Figure 3.12). In both Clusters 1 and 3 the majority of sequences were isolated from humans, at 64.7% (134/207) and 80.2% (150/187) respectively. Conversely, in Cluster 5 the most prevalent isolation source was poultry (48.1%, 582/1209).

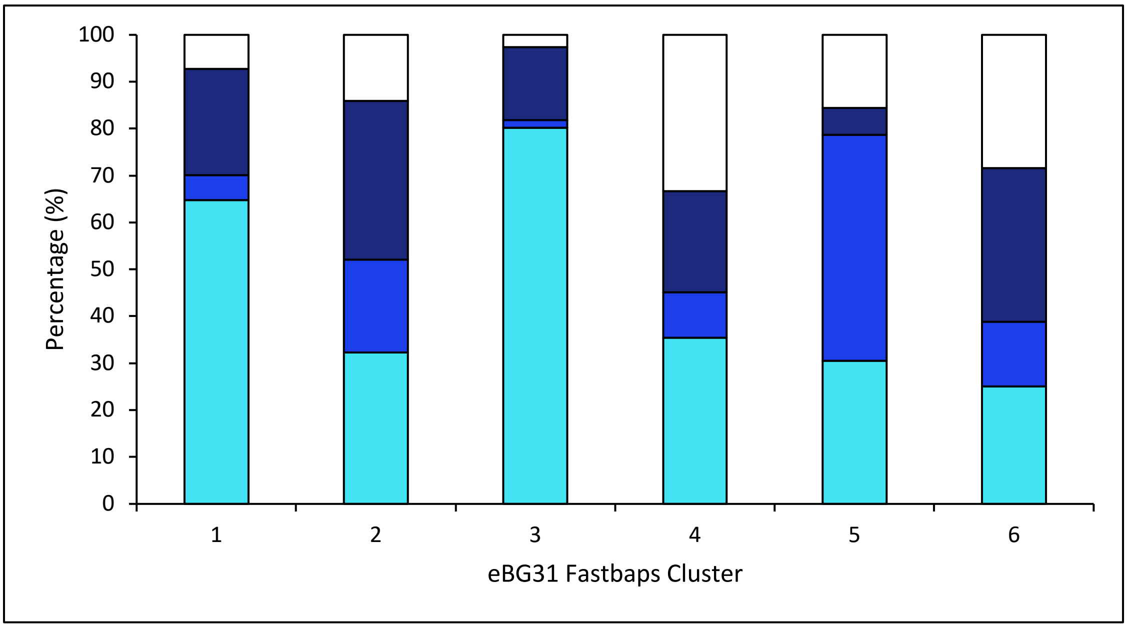


Figure 3.12 Source distribution within each eBG31 fastbaps cluster
Percentage of isolates from each eBG31 fastbaps cluster that were isolated from each isolation source.

Environmental ■ Human ■ Poultry ■ Unknown □

To identify the distance of isolates from different sources on the phylogeny the median pairwise SNP distance within and between each isolation source was calculated (Appendix III Figure III.14). Surprisingly, considering the clustering that is apparent upon looking at the phylogeny, there was very little variation in the median distance either within or between the isolation sources. The median distance between source groups ranged from 157 to 162.

3.3.2.5 *Continent Distribution Across the eBG31 Phylogeny*

The geographical distribution of eBG31 across the phylogeny was determined (Figure 3.13). There did appear to be some clustering by the continent of isolation on the phylogeny; Clusters 4 and 6 were largely composed of sequences from North America and Clusters 1 and 5 from Europe or Asia.

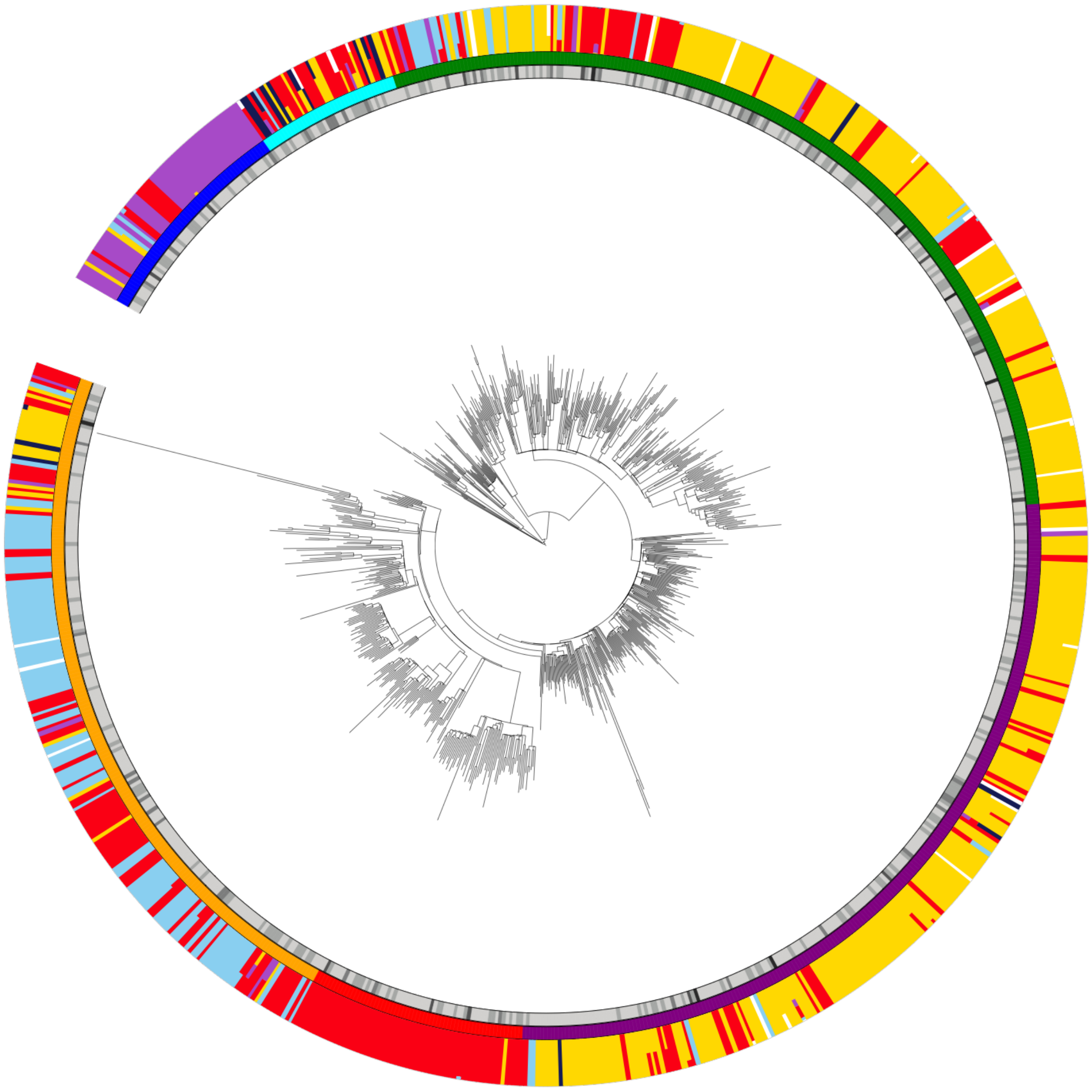


Figure 3.13 eBG31 phylogeny annotated with origin
Soft-core SNP Maximum Likelihood Phylogeny of 831 25SNP cluster representatives of eBG31. The outer ring is annotated with the percentage of isolates in each 25SNP cluster that were isolated from each continent group.

Inner ring, Number of sequences in 25SNP cluster: ░ 1 ▒ 2-5 ▓ 6-20 █ 21-50 █ >50

Middle ring, fastbaps cluster: 1 🟥 2 🟦 3 🟦 4 🟪 5 🟧 6 🟩

Outer ring, Continent of Isolation: Africa 🟪 Asia 🟦 Europe 🟥 N. America 🟨 S. America ⬛ Unknown ⬜

The percentage of sequences in each cluster from each continent is shown in Figure 3.14. Cluster 1 was almost entirely composed of sequences that were isolated from across Europe (98.6%, 204/207). The majority of the sequences in Cluster 2 were also isolated in Europe at 62.8% (76/121). However, the predominant continent changed for the other clusters; 80.2% (150/187) of the sequences in Cluster 3 were isolated from Africa, of which 1 originated from Ethiopia, 1 from Tunisia and the remaining 148 from South Africa. North American isolates dominated Clusters 4, 5 and 6. The high North American percentage in Cluster 5 can be attributed to a 25SNP cluster which contained 771 North American, 9 European, 90 South American and 4 Unknown isolates. Excluding this cluster, the percentages from each continent are as would be expected when looking at the phylogeny of representatives, with 45.7% (153/335) of the sequences isolated in Europe and 36.1% (121/335) of the sequences isolated from Asia.

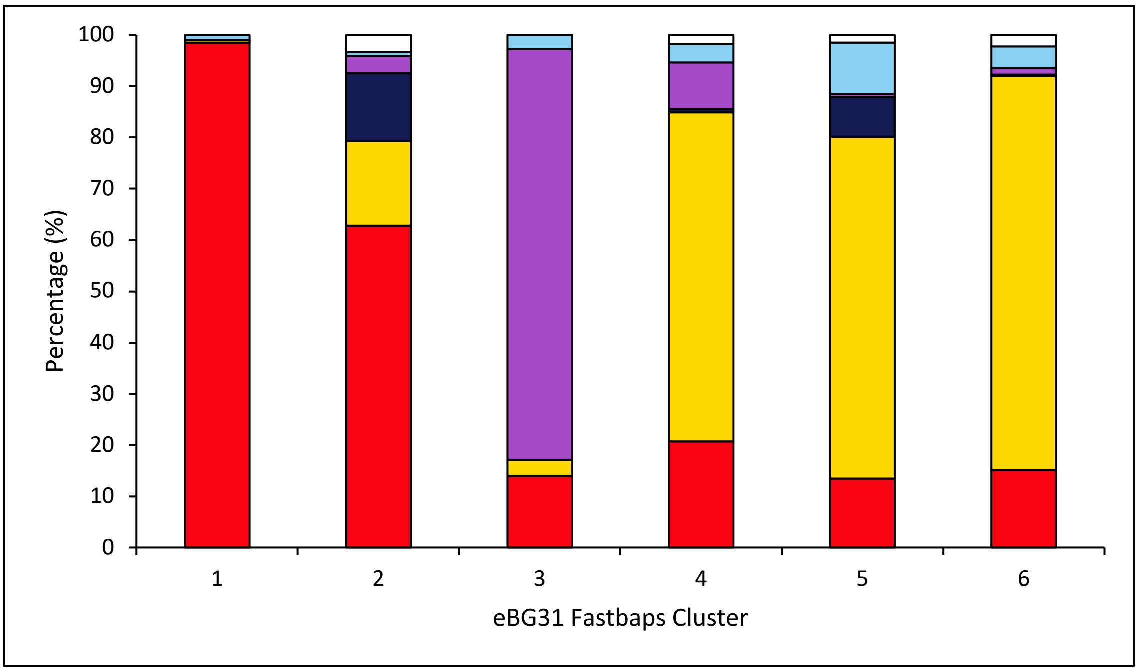


Figure 3.14 Continent distribution within eBG31 fastbaps clusters
Percentage of isolates from each eBG31 fastbaps cluster that were isolated from each year group.

Africa 🟪 Asia 🟦 Europe 🟥 N. America 🟨 S. America ■ Unknown □

To determine how distant isolates from each continent were to each other, the median pairwise SNP distance within and between each continent was investigated (Appendix III Figure III.15). The distance varied within the continents, ranging from 118 amongst the North American isolates to 189 in the African isolates. When comparing the distance between continents the North American isolates were frequently the isolates most closely related to isolates from other continents and the African sequences were consistently the most distantly related.

3.3.3 Phylogenetic Structure of the eBG297 Population

To ascertain the genetic diversity within eBG297, a soft-core SNP phylogeny was generated, masking recombination and prophages (Figure 3.15).

The phylogeny was rooted to its most ancestral node (Figure 3.15, Cluster 2). All other sequences in the eBG297 population shared a common ancestor, splitting into 2 clades, one smaller (Figure 3.15, Cluster 1) and the other containing the majority of the eBG297 sequences (Figure 3.15, Clusters 3, 4, and 5). Unlike the eBG31 phylogeny there were no clear outliers on long branches.

Five clusters were identified within the eBG297 phylogeny using *fastbaps* (Table 3.3) (Appendix VI Table VI.2). All clusters were monophyletic clades, with Clusters 3, 4 and 5 sharing a common ancestor (Figure 3.15).

Cluster	Number of sequences
1	6
2	33
3	16
4	71
5	57

Table 3.3 Number of sequences in each eBG297 fastbaps cluster

The bootstrap results were assessed for cluster stability (Appendix III Figure III.16). Clusters 1, 2 and 3 all had perfect bootstrap results. Clusters 4 and 5 had lower bootstrap results, this was due to sequences being closely related to other clusters or, in the case of Cluster 5, due to two distinct clades being present within the cluster.

The eBG297 soft-core SNP phylogeny contained 2,943 SNPs. The median pairwise SNP distance within each fastbaps predicted cluster was calculated (Appendix III Figure III.18). Compared to the eBG31 median pairwise SNP distances there was less variation within the clusters in the eBG297 phylogeny, with only Cluster 3 having a higher median pairwise SNP distance within the cluster at 103.5 (range, 0-116) than the eBG31 cluster with the lowest distance (Figure 3.9 Cluster 4). Whilst Clusters 3, 4 and 5 were closely related to one another, Clusters 1 and 2 had a higher distance from other clusters than seen in the eBG31 phylogeny.

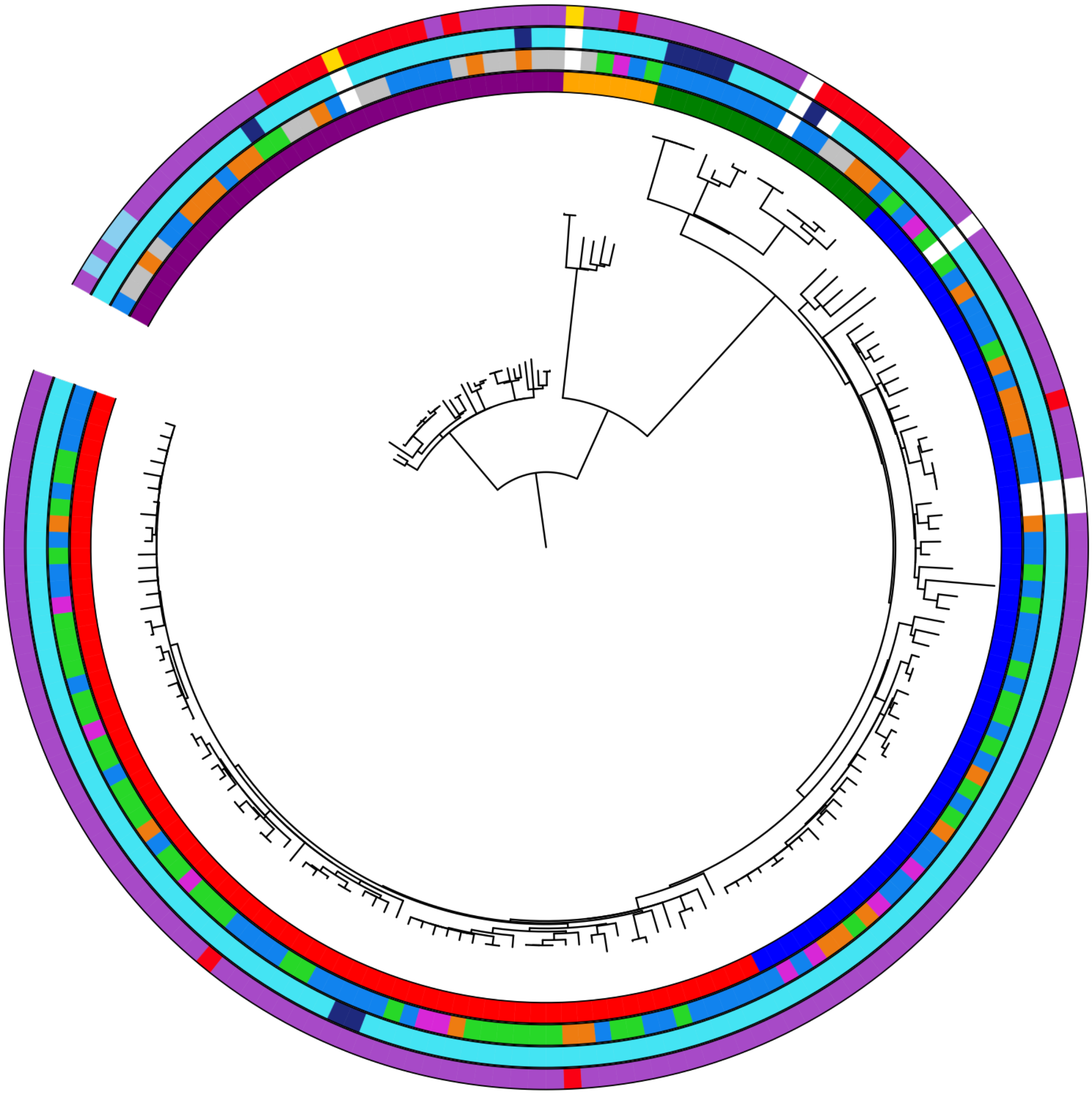


Figure 3.15 eBG297 phylogeny
Soft-core SNP Maximum Likelihood Phylogeny of 183 eBG297 isolates

Inner ring, fastbaps cluster: 1 🟧 2 🟪 3 🟩 4 🟥 5 🟦

Second ring, year group: 2003-2005 🟪 2006-2010 🟩 2011-2014 🟦
2015-2016 🟧 2017-2019 ⬜ Unknown □

Third ring, source: Environmental 🟦 Human 🟦 Unknown □

Outer ring, continent: Africa 🟪 Asia 🟦 Europe 🟥 N. America 🟨 Unknown □

The phylogeny annotated with individual rings is available in Appendix III Figures III.19, III.20, III.21.

3.3.3.1 *ST Distribution Across the eBG297 Phylogeny*

The eBG297 soft-core SNP phylogeny was annotated to determine the ST distribution (Appendix III Figure III.22). The dominant ST, ST603, was present in all clusters except Cluster 1. The STs novel to this project, ST7731 and ST7732, were present only in Cluster 4. Both ST1823 and novel3 were found exclusively in Cluster 1.

3.3.3.2 *Year Distribution Across the eBG297 Phylogeny*

To determine whether there was a temporal association with phylogeny structure, the eBG297 phylogeny was annotated with year group (Figure 3.15). Each year group was distributed across the tree, every year group being present in at least 3 of the 5 clusters. The majority of strains (65.3%, 32/49) isolated between 2006 and 2010 were found in Cluster 4. Also, 80% (12/15) of the strains isolated between 2017-2019 were found in Cluster 2.

A minimal difference in the median pairwise SNP distance was seen within or between year groups; however, the median distance was very high for all year group comparisons with 2017-2019 (Appendix III Figure III.23).

3.3.3.3 *Source Distribution Across the eBG297 Phylogeny*

The eBG297 phylogeny was annotated with isolation source to determine how isolates from different sources were distributed across the phylogeny (Figure 3.15).

As 91% (167/183) of the isolates in the eBG297 phylogeny were isolated from humans, the phylogeny was dominated by that isolation source. Despite this, none of the clusters were comprised completely of sequences from human sources. Isolates from environmental sources were present in 3 of the clusters, most noticeably in Cluster 3 where they made up 31.3% (5/16) of the sources in the cluster.

A higher median pairwise SNP distance was observed between human and environmental isolates (133, range 15 to 422) and within environmental isolates than seen within human isolates (Appendix III Figure III.17).

3.3.3.4 *Continent Distribution Across the eBG297 Phylogeny*

With a view to identify the geographical distribution across the phylogeny, it was annotated with the continent of isolation (Figure 3.15). Similar to the eBG31 population, all of the clusters contained isolates from more than one continent. As the majority of the eBG297 sequences (84.2%) were isolated from Africa, that continent dominated the phylogeny, with African sequences present in every fastbaps cluster.

As shown in Figure 3.16, although in much fewer numbers, sequences isolated from Europe were also present in each fastbaps cluster, accounting for 30.3% (10/33) of isolates in Cluster 2 and 37.5% (6/16) in Cluster 3.

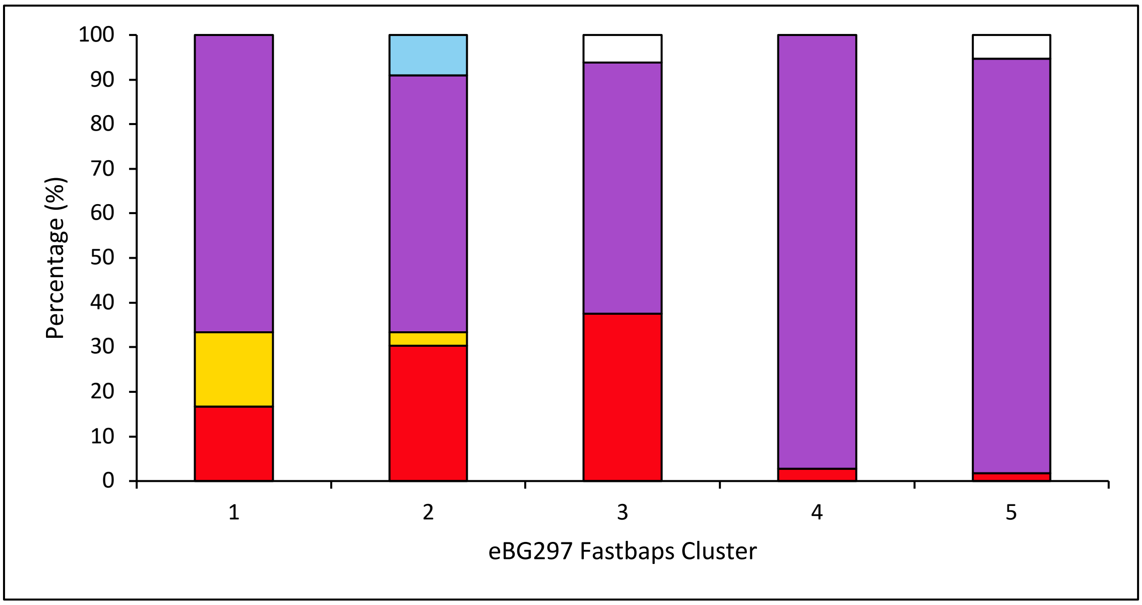


Figure 3.16 Continent distribution within eBG297 fastbaps clusters
Percentage of isolates from each eBG297 fastbaps cluster that were isolated from each year group.

Outer ring, continent: Africa 🟪 Asia 🟦 Europe 🟥 N. America 🟨 Unknown □

Whilst both African and European sequences were distributed throughout the phylogeny, the median pairwise SNP distance within the African isolates (85, range 0-431) was lower than within the European isolates (358.5, range 0-390) (Appendix III Figure III.24). When compared to isolates from other continents, the African isolates had a consistently high median pairwise SNP distance, between 350.5 and 368.5. The isolates from other continents were less distant from each other, although in the case of Europe vs. North America only slightly less at 319.

3.3.4 SNP-based Clustering Within eBG31

To establish the number and size of clusters within the eBG31 population, the SNP addresses of the 4486 sequences in the database, including the reference, were exported from SnapperDB and the number of clusters within each of the seven SNP thresholds compared.

Every sequence fell into the same 250SNP cluster, meaning all the sequences in the eBG31 population were a maximum of 250 SNPs distant from another sequence. There were 22 100SNP clusters, 1 containing 99.2% (4448/4486) of the sequences. 14 of the 100SNP clusters contained only 1 isolate. Each of the fastbaps clusters in the eBG31 phylogeny contained at least one sequence with a very long branch length between it and its most recent common ancestor (MRCA) to another sequence. The long outlier in Figure 3.9 Cluster 5 was one of the isolates belonging to a unique 100SNP cluster, potentially explaining its long branch length. The other 100SNP clusters either represented some of the outliers in the other clusters or didn't look especially distant from others on the tree.

At the 50SNP cluster level there were 208 clusters, 2 containing the majority of sequences with 2,709 and 897 sequences in them; 120 of these clusters contained only 1 sequence. 831 25SNP clusters were present, described above in greater in detail (Chapter 3.3.2). There were 1831 10SNP clusters, 74.9% (1371/1831) of which were single sequence clusters. The largest 10SNP cluster contained 483 sequences.

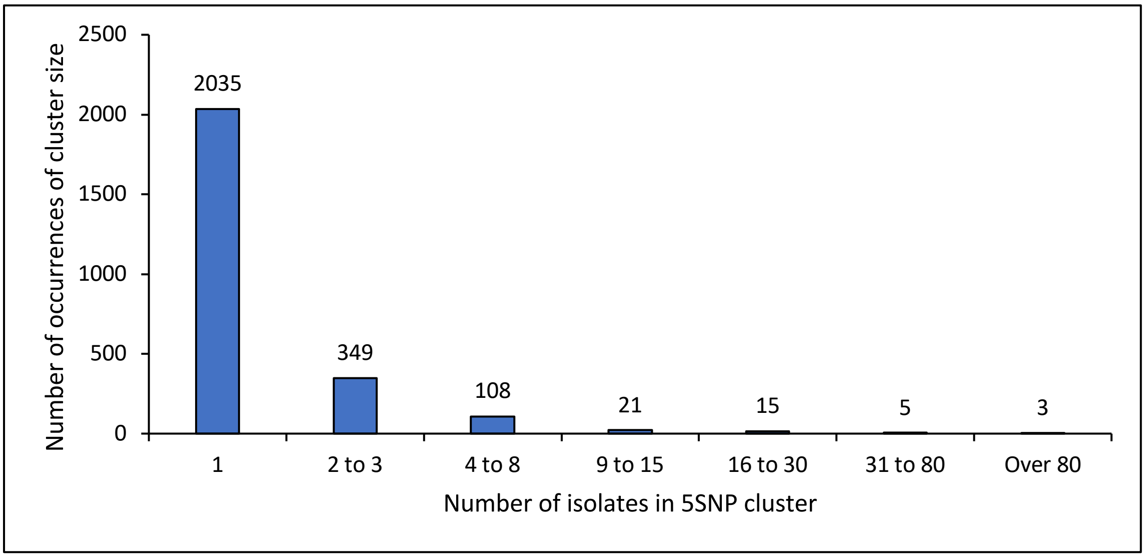


Figure 3.17 Distribution of 5SNP cluster sizes in eBG31
The number of occurrences of each 5SNP cluster size is plotted, with the majority of clusters containing only one isolate.

At the 5SNP cluster level, the level PHE define as an outbreak for *S. enterica*, there were 2536 clusters. 80.2% (2035/2536) of these clusters contained 1 isolate and were therefore sporadic cases. The 501 “outbreak” clusters, containing 2 or more cases, accounted for 54.6% (2451/4486) of cases. Figure 3.17 shows the variation in the number of isolates belonging to outbreaks; 3 of the clusters contained over 100 sequences.

At the 0SNP cluster level there were 3806 clusters; 92.2% (3508/3806) of these contained 1 sequence. Several clusters contained multiple isolates that were identical to each other; the size of the remaining clusters varied, with the majority, 65.8% (196/298) containing 2 sequences but the largest containing 82 sequences.

3.3.5 SNP-based Clustering Within eBG297

To identify the number and size of clusters within the eBG297 population, the SNP addresses for all 184 sequences, including the reference, in the eBG297 SnapperDB were compared. Unlike the eBG31 population, there were 6 different 250SNP clusters in the eBG297 population.

There were 14 100SNP clusters with 66.3% (122/184) of sequences belonging to one of the clusters; 4 of the clusters contained only 1 isolate. The more ancestral groups within eBG297 (Figure 3.15, Cluster 1, 2, 3), were comprised entirely of sequences that didn’t belong to the dominant 100SNP cluster. At the 50SNP cluster level there were 37 clusters of sequences. The largest cluster contained 33.7% (62/184) of the sequences and 21 of the clusters contained 1 sequence.

At the 25SNP cluster level the number of clusters increased to 95, with the largest containing only 16 of the sequences. 71.2% (68/95) of these clusters contained only 1 sequence. The proportion of clusters containing 1 sequence increased at the 10SNP cluster level; 120 of the 144 clusters (83.3%) contained a single sequence.

At the 5SNP cluster level, 85.9% of the 149 clusters contained 1 sequence and were therefore considered to be sporadic cases. 37.6% of the sequences were in the remaining 21 outbreak clusters, which contained between 2 and 10 sequences. At the 0SNP cluster level there were 163 clusters, 17 of which contained more than 1 sequence.

3.3.6 Whole Genome Distances with *S.* Infantis

A non-alignment based method, *Mash*, was used to compare the distance within and between the metadata groups in both the *S.* Infantis eBGs. This allowed for the distance between the whole genomes to be compared as opposed to the soft-core genome. Furthermore, it enabled comparison between the eBGs. The input for this method was assembled sequences; the sequences were assembled and assessed for quality. 95/4486 eBG31 and 5/184 eBG297 sequences that had been included in the phylogenies were excluded from this analysis.

3.3.6.1 *Distances within S. Infantis based on Isolation Source*

The Mash distance within and between eBG31 and eBG297 sequences isolated from human, poultry and environmental sources was calculated (Figure 3.18).

For all of the comparisons within and between isolates based on metadata group, the minimum Mash distance was 0, except for the distance between eBG31 environmental and poultry sources, which was just above at $2.4\text{x}10^{-6}$. The median Mash distances were higher in the comparisons amongst source groups in the eBG31 population when compared to the eBG297 results.

The eBG31 human isolates appeared to be more closely related to environmental isolates, with a significantly lower Mash distance seen between these groups (0.0016) than between human and poultry isolates (0.0021) (p-value $< 2.2\text{x}10^{-16}$). Furthermore, a higher median Mash distance was seen between eBG31 human and poultry isolates than observed within isolates from humans (0.0018).

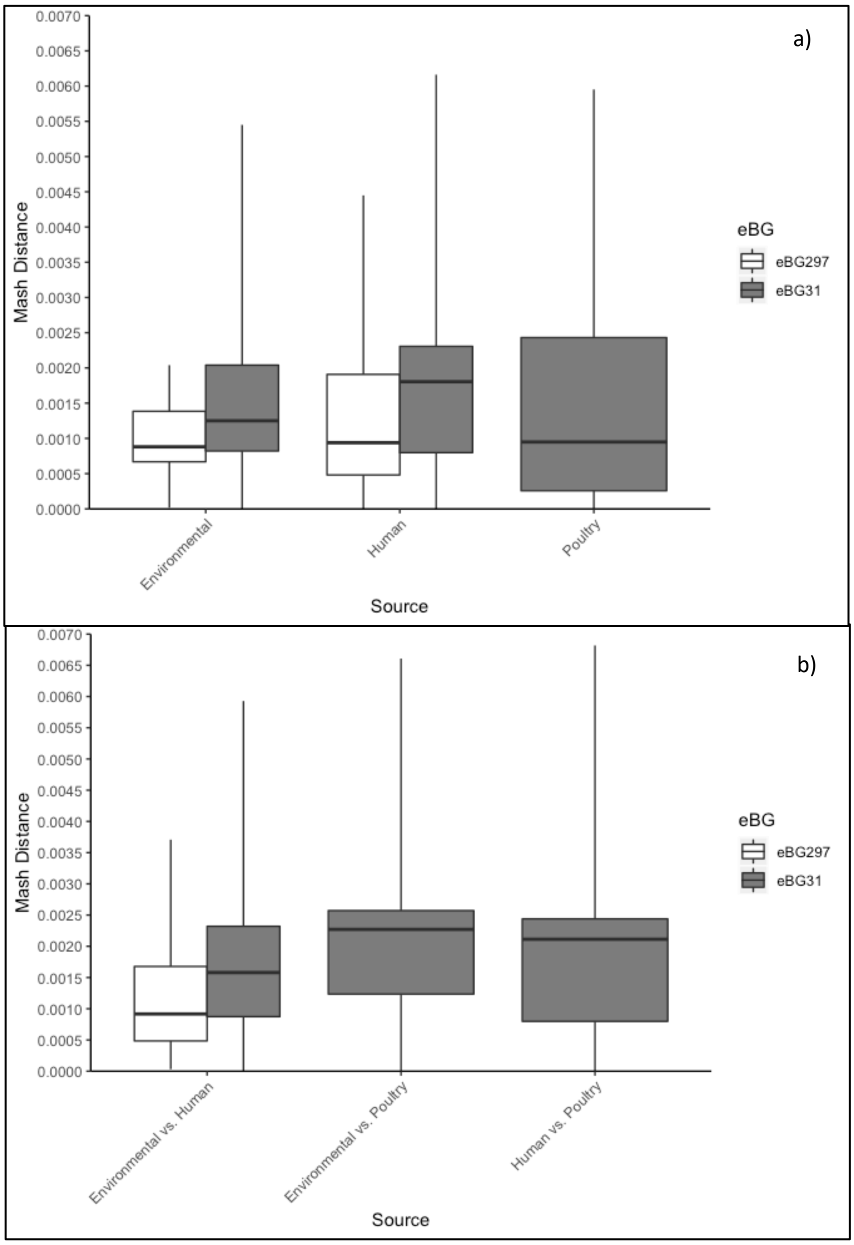


Figure 3.18 Genetic variation within and between *S.* Infantis isolated from different sources

Box and whisker plot showing the minimum, 1st quartile, median, 3rd quartile and maximum Mash distances across each eBG31 and eBG297 isolation source.

a) Distance between isolates within each source
b) Distance between isolates from different sources

3.3.6.2 *Distances within S. Infantis based on Continent of Isolation*

The distance within and between strains that had been isolated from each continent was calculated using Mash (Figure 3.19). The median distances within the African isolates from either eBG were comparable. The median Mash distance was higher within eBG31 Asian and European isolates than observed in eBG297 isolates. Conversely, a higher median Mash distance was seen in eBG297 North American isolates than in eBG31, but only 2 isolates were in the eBG297 comparison.

The median Mash distance between eBG31 continents did not vary greatly, with the smallest distance seen between isolates from Asia and South America at 0.0010 and the largest distance between Africa and South America at 0.0023. A similar variation in distance was seen between eBG297 isolates from different continents, the smallest was between Asia and Europe at 0.0010 and the largest between Africa and North America at 0.0024.

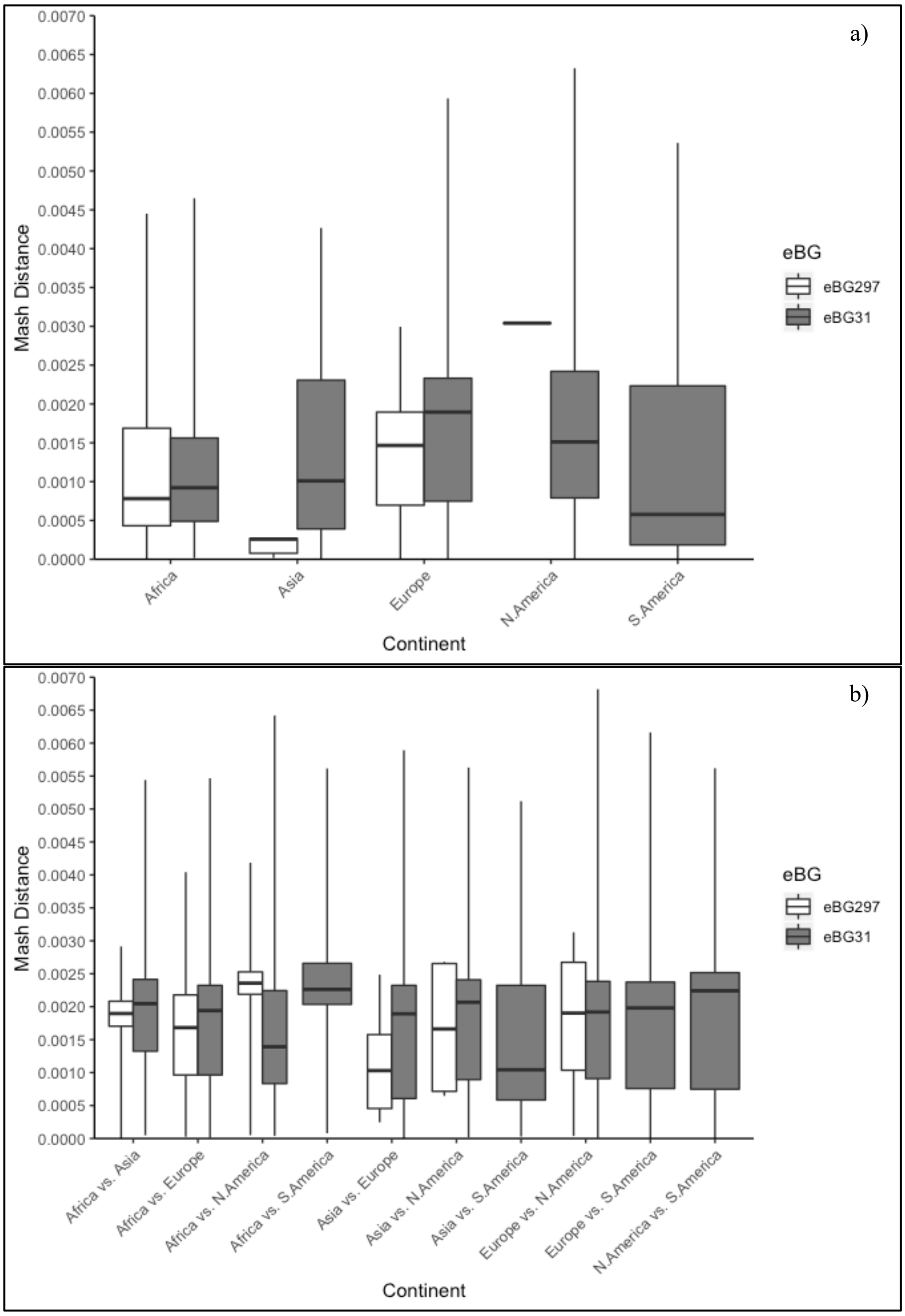


Figure 3.19 Genetic variation within and between *S.* Infantis isolated from different continents

Box and whisker plot showing the minimum, 1st quartile, median, 3rd quartile and maximum Mash distances across each eBG31 and eBG297 continent of isolation.

a) Distance between isolates within each continent
b) Distance between isolates from different continents

3.3.6.3 *Distances within S. Infantis based on eBurstGroup*

The overall distance within and between each eBG was calculated to identify the genetic distance between isolates from eBG31 and eBG297 (Figure 3.20). The median Mash distances within eBG31 and eBG297 were comparable at 0.00174065 and 0.00099543 respectively. The median Mash distance between isolates belong to eBG31 and eBG297 was 0.00713482, 4.1x the distance within eBG31 and 7.2x the distance within eBG297. The distribution of both the eBG31 and eBG297 Mash distances were significantly different to the distribution of the distances between the eBGs (p-value < 2.2×10^{-16}).

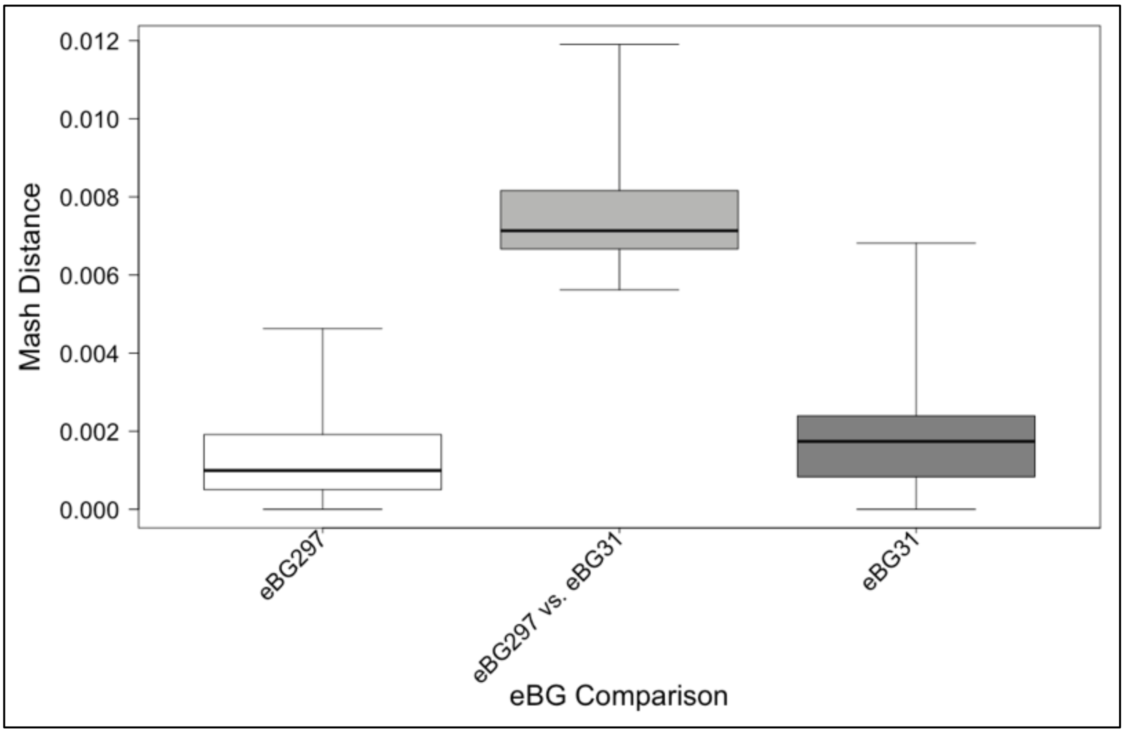


Figure 3.20 Genetic variation within and between eBG31 and eBG297
Box and whisker plot showing the minimum, 1st quartile, median, 3rd quartile and maximum Mash distances within and between each *S*. Infantis eBG.

3.4 Discussion

Whilst globally the proportion of *S.* Infantis isolates that belonged to eBG297 was 3.9%, this increased substantially to 34.1% in Africa. As the majority of eBG297 sequences were from Africa this indicates that, similar to clades observed in *S.* Enteritidis, this eBG is a distinct lineage of *S.* Infantis that is associated with this continent (Feasey *et al.*, 2016). However, the median Mash distance between eBG31 and eBG297 was 4.1x greater than the distance within the eBG31 population and 7.2x greater than the distance within the eBG297 population. This suggests that these two eBGs are too genetically distinct to belong to the same serovar; to resolve this hypothesis the distance within and between other *Salmonella* serovars would need to be analysed for comparison.

Unlike previous reports, there was a strong geographical signal within the eBG31 soft-core SNP phylogeny (Gymoese *et al.*, 2019). Each of the six clusters in the phylogeny was dominated by sequences from a specific continent. Whilst none of them exclusively contained isolates from one continent, this information would be of use for epidemiologists during outbreak investigations as it would indicate the potential continent of origin.

The median pairwise distances across the phylogeny indicate that whilst the North American sequences were the closest related to the isolates from other continents, the African isolates were consistently the most distant. This could suggest that there are lineages within the eBG31 population that are associated with Africa. When looking at whole genome comparisons the African sequences joined the South American sequences as having the lowest median Mash distance. This indicates that isolates from these continents had a smaller accessory genome than seen in isolates from the other continents. As similar distances were seen between continents across the eBG297 phylogeny and when compared using Mash, this suggests that the eBG297 isolates also have a smaller accessory genome than seen in eBG31.

In EU member states in 2018, *S.* Infantis was most frequently identified in broilers and broiler meat (EFSA and ECDC, 2019a). Concordantly, in this *S.* Infantis collection, broiler meat was the animal product associated with the highest number of isolates.

Whilst *S.* Infantis is very common in poultry, none of the eBG297 sequences were isolated from poultry. Furthermore, none of the eBG31 African sequences were isolated from poultry; as eBG297 is most prevalent in Africa it is possible that it is found in poultry, but this was not captured by the samples included. However, although unlikely, it is possible that the reason eBG297 was not found in poultry is that it is not capable of colonising domestic fowl and has adapted to other niches.

Previous research has identified evidence of clustering by isolation source, with human and poultry isolates clustering together (Yokoyama *et al.*, 2015). Concordantly, clustering by isolation source was visible in the eBG31 phylogeny for human, poultry and environmental sources. In particular, clustering of human isolates was most apparent; the majority of the sequences in 2 of the 6 fastbaps clusters were isolated from humans. It is possible that there are strains, including those in these human dominant clusters, that have adapted to become more virulent to humans. However, the median pairwise SNP distances between isolates from humans and the other sources did not support this hypothesis as there was very little difference between the distance within isolates from humans and between these isolates and isolates from other sources.

When the accessory genome was included in the comparison, there was more variation in the distances between sources. The isolates that caused infection in humans were found to be more diverse than those infecting poultry. Unlike the other source types, there were no sequences isolated from poultry that were identical to sequences from environmental sources, suggesting a lack of direct transmission between these source types. Interestingly, despite there being a large amount of variation amongst the human isolates, the distance between human and poultry isolates was greater. This concurs with the interpretation of the clustering within the phylogeny and could indicate that different members of the *S.* Infantis population cause infection in poultry and human hosts.

As not all of this data was collected by public health teams as part of surveillance it cannot be used to estimate the global prevalence of *S.* Infantis; the increase in *S.* Infantis sequences available with time just indicates that more sequencing is occurring. Small clusters were seen of more current sequences within the eBG31 phylogeny, which could indicate that lineages of *S.* Infantis have evolved recently that are more successful. However, it is probable that their appearance is due to the increased sequencing being performed by public health institutes.

The number of clusters within the eBG31 and eBG297 SnapperDB's were different. Whilst all of the sequences in the eBG31 population clustered into the same 250SNP cluster, there were 6 250SNP clusters within the eBG297 population. This could be indicative of a greater diversity within eBG297 but could also be explained by the smaller number of sequences in the database. As more sequences are added to the database, clusters merge if, upon the addition of a new sequence, they are now within that cluster threshold of one another. As there are fewer sequences in the eBG297 database the likelihood of a related sequence already being in the database is reduced, increasing the number of clusters.

Historically *S.* Infantis has been associated with outbreaks in humans, for example, one was investigated between January 2018 to January 2019 in the United States that had 129 cases and was associated with chicken products (CDC, 2019). In eBG31, 54.6% of the sequences were outbreak associated, however fewer of the eBG297 sequences were outbreak associated at 37.6%. eBG31 is therefore more outbreak driven than eBG297.

It was identified that in an *S.* Enteritidis eBG SnapperDB, 50% of the cases fell into 58 of the 2,302 5SNP clusters (Dallman, 2018). It was deemed that targeting those clusters would be an effective strategy to reduce the number of infections. Whilst targeting specific 5SNP clusters is a good strategy for some *S. enterica* serovars, it would not work for eBG31 as 54.6% of the cases fell into 501/2536 of the 5SNP clusters. Targeting specific 10SNP clusters would be more effective for eBG31 as 50.1% of the cases fell into 126/1831 of the 10SNP clusters. The strategy could be more successful in the eBG297 population as 37.6% of the cases fell into 21/149 5SNP clusters; however, due to the low number of cases associated with this eBG this is not currently necessary.

3.4.1 Conclusions

To conclude, eBG31 is the most prevalent eBG in *S.* Infantis globally although higher levels of eBG297 are seen in Africa. The sequence data also indicates that there is a distinct lineage of African sequences within the eBG31 population. Further evidence of a strong geographical signal was observed in the eBG31 with clustering by continent present. Clustering by source was also observed, particularly for human isolates.

eBG31 was found to be more diverse and outbreak driven than eBG297. As the Mash distance between the eBGs was greater than the distance within either eBG, I propose that both eBG31 and eBG297 should not be classified as *S.* Infantis.

4. Chapter 4. Antimicrobial Resistance and Mobile Genetic Elements in *S.* Infantis

4.1 Introduction

AMR is a public health concern in *Salmonella*; the WHO reports fluroquinolone-resistant *Salmonella* as being high priority for research (Tacconelli *et al.*, 2018). Increasing levels of MDR are also concerning as *S. enterica* with MDR are associated with causing infection with an increased severity; in Kenya, iNTS infection was found to be associated with MDR in NTS (Eng *et al.*, 2015; Akullian *et al.*, 2018).

Worryingly, *S.* Infantis is associated with higher levels of AMR and MDR, often higher than seen in other serovars (EFSA and ECDC, 2013a; Food and Drug Administration (FDA), 2019a). For example, in 2017, a higher percentage of *S.* Infantis isolates from humans had high level resistance to ciprofloxacin (minimum inhibitory concentration ≥ 4mg/L) than *S.* Derby, *S.* Enteritidis, *S.* Typhimurium and monophasic *S.* Typhimurium isolates in 5 EU member states (EFSA and ECDC, 2019b). The levels of AMR in *S.* Infantis are also increasing, in *S.* Infantis isolates collected from humans in the United States between 1996 and 2017, the levels of resistance to the following antimicrobials was highest in the latest date point: ampicillin, ceftriaxone, chloramphenicol, nalidixic acid, streptomycin, tetracycline, trimethoprim-sulfamethoxazole (FDA, 2019b).

High levels of MDR have been found in *S.* Infantis in EU member states, with isolates from broilers being a large contributor; in 2016, 31% of *S*. Infantis strains from broilers and 70% of isolates from broiler meat had MDR (EFSA and ECDC, 2018a). Higher levels of MDR are also observed with *S.* Infantis; in Japan between 2004 and 2006, 120 isolates from broilers were isolated, all of which had MDR (Shahada *et al.*, 2010). This was associated with integron and plasmid presence as all isolates had an 180kb plasmid and a class 1 integron. Integrons are not always as frequently identified in *S.* Infantis, in Iran only 36% of isolates from chickens contained a class 1 integron (Asgharpour *et al.*, 2014).

Also of concern is that several different ESBLs have been identified in *S.* Infantis, with $bla_{CTX\text{-}M\text{-}65}$ being the most frequently reported, present in strains from Ecuador, Peru, Switzerland, the UK and USA (Burke *et al.*, 2013; Cartelle Gestal *et al.*, 2016; Hindermann *et al.*, 2017; Tate *et al.*, 2017; Granda *et al.*, 2019). Mutations in the QRDRs of *gyrA*, *gyrB*, *parC* and *parE*, predicted to confer resistance to quinolones, have also been identified in *S*. Infantis isolated from poultry carcasses in Pakistan; 62.5% of isolates had a mutation in the QRDR of *par*E (Wajid *et al.*, 2019).

S. Infantis used to be viewed as a serovar associated with an absence of plasmids (Rychlik, Gregorova and Hradecka, 2006), however, in the past 6 years several papers have reported the presence of a novel megaplasmid in *S.* Infantis isolates. This megaplasmid, pESI, was first described in isolates from Israel in 2008; currently the oldest isolate known to contain pESI was identified in Japan in 2000 (Aviv *et al.*, 2014; Gymoese *et al.*, 2019).

The Israeli pESI is approximately 280kb long and contains several mobile genetic elements with resistance genes for aminoglycosides, trimethoprim, sulphonamides and tetracyclines conferred by *aad*A1, *dfr*A, *sul*1, *tet*A respectively. pESI has also been associated with ESBLs in both humans and broilers, in particular with *bla*$_{\text{CTX-M-65}}$ in isolates from Switzerland and *bla*$_{\text{CTX-M-1}}$ from Italy (Franco *et al.*, 2015; Hindermann *et al.*, 2017).

4.1.1 Aims and Objectives

Whilst high levels of AMR in *S.* Infantis have been described globally, little is known about the genes conferring this resistance and how they vary globally and between hosts. The occurrence and variation of pESI and other mobile genetic elements in *S.* Infantis is also not well understood.

This chapter therefore had the following aims:

- Identify the occurrence of AMR and MDR in the *S.* Infantis population
- Identify the occurrence of plasmids, in particular pESI, in the *S.* Infantis population
- Determine the genetic diversity of pESI
- Identify the occurrence of integrons in the *S.* Infantis population

4.2 Specific Methods for AMR, Plasmid, and Integron Detection

The *S.* Infantis sequence collection was assessed for the presence of genes or mutations encoding AMR resistance, plasmids, pESI plasmid variant and integrons. The Illumina sequenced eBG31 reference was used in all analyses.

4.2.1 Using ARIBA to detect AMR, Plasmids and Gyrase Mutations

ARIBA (version 2.10.1) was run on the *S.* Infantis collection with the *ResFinder*, *PlasmidFinder* and gyrase databases (Chapter 2.7.4) (Zankari *et al.*, 2012; Carattoli *et al.*, 2014; Hunt *et al.*, 2017). For the gyrase database, *ARIBA* summary was run with the preset all option to output all results; for the *ResFinder* and *PlasmidFinder* databases the cluster_cols match option was used to output only the match column.

4.2.2 Analysing the ResFinder Output

The AMR determinants were grouped into the classes of antibiotic resistance they conferred, with the cryptic resistance gene *aac6* ignored for all calculations of AMR or MDR. The *ARIBA* refquery command was used with the ResFinder database and each of the AMR gene clusters identified, to confirm that every gene in each cluster conferred resistance to the same antibiotic class. Table 4.1 lists the different genes identified and

Antimicrobial Class	*ARIBA gene cluster*
Aminoglycosides	aac_3__II-, aac_3__IV-, aac_3__VIa, aac_6___Iaa, aadA-, aadA-_1, aadA1+, aadA13, ant_2____Ia, ant_9__Ia, aph_3___IIa, aph_3___Ia, aph_3____Ib, aph_4__Ia, aph_6__Ic, aph_6__Id,
Beta-lactams	bla-, bla-_3, bla-_7, blaCARB_-_1, blaCTX_M_-, blaCTX_M_-_2, blaCTX_M_-_5, blaSHV_-, blaTEM_-
Chloramphenicols	catA1, catA2, cat_1, cat_pC194_, cml-, floR, mdf_A_
Fosfomycins	fosA3, fosA7
Lincosamides	lnu_C_, lnu_F_, lnu_G_
Macrolides	erm_B_, mef_B_, mph_A_
Polymyxins	mcr_1
Quinolones/Fluoroquinolones	oqxA, oqxB, qnrB-, qnrD-, qnrS-
Sulfonamides	sul1, sul2, sul3
Tetracyclines	tet_A_, tet_B_, tet_C_, tet_D_, tet_G_
Trimethoprims	dfrA-, dfrA-_2, dfrA10, dfrA12, dfrA14, dfrA15, dfrA1_1, dfrA8, dfrB3

Table 4.1 AMR genes and classes
Gene clusters identified by *ARIBA* in the *S*. Infantis isolates and the antimicrobial class they encode resistance for.

the class that they belonged to. MDR was defined as resistance to three or more antibiotic classes.

4.2.3 Analysing the PlasmidFinder Output

31 of the 46 plasmid clusters found by ARIBA were grouped by incompatibility type (Table 4.2). The other 15 were kept as single plasmid types. The ARIBA refquery command was used to confirm that every plasmid group contained plasmid types that belonged to that group. All did with the exception of pSL483 which contained two: pSL483.1 and pXuzhou21.1.

Plasmid Group	*Plasmid Cluster*
IncF	IncFIA, IncFI-, IncFIB_-, IncFIB_AP001918_, IncFIB_K_, IncFIB_pHCM2_, IncFII_-, IncFII_SARC14_, IncFII_S_, IncFII_p-, IncFII_p96A_, IncFII_pECLA_
IncH	IncHI1A, IncHI1B_-, IncHI2, IncHI2A
IncI	IncI-, IncI.Gamma_1_AP011954, IncI1.1_Alpha_AP005147, IncI2, IncI2.1_Delta_AP002527, IncI2.1__KP347127
IncL/M	IncL_M-, IncL_M_pMU407_
IncN	IncN, IncN-
IncX	IncX1 IncX1_1, IncX3, IncX4, IncX4_1

Table 4.2 *ARIBA* plasmid clusters

31 of the 46 plasmid clusters identified by *ARIBA*, grouped by incompatibility type.

4.2.4 Analysing the Gyrase Mutations

From the literature, the QRDRs of the *Salmonella* gyrase genes (DNA gyrase encoded by *gyrA* and *gyrB* and topoisomerase IV by *parC* and *parE*) were defined as mutations found between amino acids 67 to 106 in *gyrA*, 415 to 470 in *gyrB*, 47 to 133 in *parC* and 450 to 528 in *parE* (Eaves *et al.*, 2004; Michael *et al.*, 2006). A mutation in *parC*, T57S, was present in every isolate but previous research has shown this mutation is not associated with quinolone resistance so it was not investigated further (Lunn *et al.*, 2010).

As the four gyrase genes are essential for *Salmonella*, any sequences that came back as negative for their presence were investigated as a quality control step. The *ARIBA* report files for *gyrA, gyrB, parC* and *parE* were looked at individually to determine if the gene was present but fragmented or interrupted. These sequences were also mapped to the eBG31 reference genome with *BWA-MEM* (version 0.7.12) and sorted using *SAMtools* sort (version 1.5) (Li *et al.*, 2009; Li, 2013). *Artemis (*version 17.0.1) was then used with the BAM file output and the *Prokka* gff output (Chapter 2.6.8) to visualise the location and coverage of the missing gyrase gene (Carver *et al.*, 2012; Seemann, 2014).

All had coverage so the report file for each sequence was checked to see what numbered flag *ARIBA* reported; the ARIBA flag function was used to identify what each flag meant. Those reported as missing gyrase all had an interrupted or fragmented copy of the gene; the results for all sequences were used.

4.2.5 Identification of pESI presence

All *S.* Infantis sequences were screened to determine whether they contained pESI and how the plasmid varied.

4.2.5.1 *Pipeline Development for the Identification of pESI*

As pESI is not present in the PlasmidFinder database, several methods were trialled to identify pESI in the isolates (Carattoli *et al.*, 2014). Attempts using *ARIBA* were unsuccessful as the software was developed for genes and so couldn't handle this data type.

The eBG31 reference and the Israeli pESI reference (ASRF01000099-ASRF01000108) were concatenated to prevent common genetic motifs mapping to the pESI reference when the plasmid was not present. This will henceforth be referred to as the pESI pseudomolecule. In trials the Illumina sequenced eBG31 reference was used, once the method had been finalised the long-read sequenced reference was used.

Sequences were mapped to the pESI pseudomolecule using *BWA-MEM* (version 0.7.12) and *SMALT* (version 0.76) and the average coverage for each contig was determined (Li, 2013; Ponstingl and Ning, 2014). The presence of other plasmids confounded these results, with high levels of mapping occurring in small areas of pESI. It was therefore decided that the coverage needed to be visualised using a heatmap. *SMALT* was selected as the mapping software with seed set to 5 to enable random mapping; reducing mapping to conserved regions.

4.2.5.2 *Heatmap Generation for eBG31 isolates*

To facilitate the identification of pESI across all the genomes, nine were initially chosen: three that had pESI, three that had high coverage at some points but did not have pESI, and three without pESI. Initially the mean and total coverage of the pESI reference was calculated for these sequences, but it was not possible to distinguish between presence and absence with these results. The number of zeros in the coverage output was calculated and this did allow for differentiation; therefore, a bash script was written which counted the number of zeros in the eBG31 collection (Appendix I.4). Sequences with 50,000 bases with zero coverage or less were predicted to contain pESI; those with between 50,000 and 150,000 bases with zero coverage were marked as maybes and those with 150,000 or higher were marked as negative for pESI. These predictions were used to group sequences when making heatmaps.

I wrote a bash script (Appendix I.5) to produce a pESI coverage matrix for each sequence using *SMALT* (version 0.7.6) and *SAMtools* (version 1.5), which was then transposed (Appendix I.6). Separate lists of the yeses, maybes and nos were made in the format '(genome_1, genome_2, … genome_n);' containing approximately 200 genomes, to be imported as trees for the heatmaps into *R* (version 3.5) using the *ape* package (version 5.2) and *R Studio* (version 1.1.463) (Paradis and Schliep, 2018; R Core Team, 2018; RStudio, 2018). Each set of approximately 200 genomes also had their corresponding coverage matrices merged using the Unix join command.

I wrote a bash script to generate R scripts for each of the heatmaps. *R* (version 3.5)*,* with the packages *data.table* (version 1.11.8) and *phytools* (version 0.6-63), was then used to generate heatmaps for each of the lists (Revell, 2012; Dowle and Srinivasan, 2018) (Appendix I.7). Due to the large range of values for coverage the heatmap was not decipherable; a value that signified presence was chosen. The mapping for one genome was visualised using *Artemis* (version 17.0.1) to observe what depth values signified presence or absence of pESI (Carver *et al.*, 2012). Heatmaps were made with presence being read coverage higher than 10, 15, 20 and 30. The value of 20 was chosen to determine presence of a read.

pESI presence/absence in all the heatmaps was judged by eye, with coverage of the majority of the plasmid needed for a positive result. Those that had a questionable result were visualised on a smaller heatmaps containing approximately 100 sequences, this time sandwiched between pESI negative sequences. The script used to calculate the number of zeros (Appendix I.4) was amended to count the number of bases which had a coverage below 20, the cut-off used in the heatmap generation, for the sequences that could still not be resolved. If the number below 20 was greater than 50% of the total length of the plasmid then pESI was classified as not present.

4.2.5.3 *pESI Presence in eBG297*

The eBG297 reference, PHE_709, was concatenated with the Israeli pESI reference to create a pseudomolecule for eBG297. This was indexed with *SMALT* (version 0.76) and the script used to generate coverage matrices for the eBG31 sequences (Appendix I.5) was amended to use the new pseudomolecule (Ponstingl and Ning, 2014). The heatmap was then judged by eye to determine eBG297 pESI presence/absence.

4.2.5.4 *Annotating the pESI Reference*

The concatenated pESI reference was run through *ResFinder* to identify the location of any AMR genes. *Prokka* (version 1.11) was used with the parameters described in Chapter 2.6.8 to annotate the individual contigs of the pESI reference; the gff output was then viewed with *Artemis* to verify resistance gene location (Carver *et al.*, 2012; Seemann, 2014). The contig names and the '>' contig identifier were removed from the pESI reference to make it a concatenated sequence. *Integron Finder* on the Galaxy Pasteur website was then used on this concatenated sequence with local detection to identify the location and class of any integrons (Hyatt *et al.*, 2010; Eddy, 2011; Nawrocki and Eddy, 2013; Cury *et al.*, 2016a). The *ResFinder* results were used to identify the location of AMR genes in the integrons detected.

4.2.5.5 *Integron A and Integron B Presence in pESI*

The presence/absence of the two integrons in the Israeli pESI reference was determined in the pESI positive isolates. The integron identified on the ASRF01000104.1_contig_55 contig of the pESI assembly is referred to as Integron A and the integron on ASRF01000099.1_contig_4 as Integron B. The script used to calculate the number of zeros (Appendix I.4) was amended to count the coverage of the integrons. Another version was created which counted the number of bases which had a coverage below 20, the cut-off used in the heatmap generation.

These numbers were compiled; sequences that had fewer zero's than 25% of the total length of the integron and those where every base in the integron had less than 20 reads coverage were marked as containing and not containing the integron respectively. For sequences that were between these cut-offs, heatmaps were made using the scripts described in Chapter 4.2.5.2, but pulling out 'ASRF01000104', bases 1,384 to 3,246 for Integron A and 'ASRF01000099', bases 255 to 2719 for Integron B. The sequences were then judged by eye for integron presence. For those that were difficult to judge, absence was defined as sequences with high numbers of zeros or 50% of the length of the integron having a read depth below 20.

4.2.5.6 *Extended-Spectrum Beta-Lactamase (ESBL) Containing pESI*

The ESBL *bla*$_{\text{CTX-M-65}}$, previously found on pESI, was searched for in the pESI positive and *bla*$_{\text{CTX-M-65}}$ positive isolates.

To identify if the *ARIBA*-identified *bla*$_{\text{CTX-M-65}}$ genes were located on the pESI plasmid, the Israeli pESI reference could not be used as it was not positive for *bla*$_{\text{CTX-M-65}}$ (Hunt *et al.*, 2017). Assemblies of pESI containing an ESBL were published by Tate *et al.*, 2017 (Tate *et al.*, 2017). I downloaded the four plasmid assemblies (CP016407, CP016409, CP016411, CP016413) and used *BRIG* (version 0.95) to compare them to the Israeli pESI reference and each other (Alikhan *et al.*, 2011). CP016407 was chosen as the ESBL-containing pESI reference as it had the least missing from the original pESI reference, the least missing from the other American sequences and also the other American sequences were missing more from it. *ResFinder* confirmed that it contained *bla*$_{\text{CTX-M-65}}$ (Zankari *et al.*, 2012). Integron finder on the Galaxy Pasteur website was used with the ESBL-containing pESI reference, using the local detection parameter, which identified that *bla*$_{\text{CTX-M-65}}$ was not on an integron (Cury *et al.*, 2016b).

Initially heatmaps were made using the ESBL-containing pESI as the reference (Chapter 4.2.5.2), these heatmaps did not provide sufficient evidence as they didn't show the location of the gene and the coverage seen either side of it in the heatmap, so therefore didn't show that the gene was on pESI. The method used by Franco *et al.*, 2015 to identify whether *bla*$_{\text{CTX-M-1}}$ was on pESI was then trialled (Franco *et al.*, 2015). I downloaded an *E. coli bla*$_{\text{CTX-M-1}}$ gene (DQ915955) and used *Nucleotide BLAST* (version

2.9.0+) to blast one of the genomes in their paper (ERR1014108) against *bla*$_{\text{CTX-M-1}}$. The contig that contained the gene was then blasted against the Israeli pESI reference. There were three hits to the pESI reference which were 86, 76 and 55 bases long. I did not believe that this was sufficient evidence that *bla*$_{\text{CTX-M-1}}$ was on pESI in this isolate. Furthermore, this method was not suitable for large numbers of sequences.

To overcome these issues, I took an alternative approach. The eBG31 long-read sequenced reference was concatenated with the ESBL-containing pESI reference. I wrote a script which blasted the scaffolded assembly of each sequence against this new reference using *Nucleotide BLAST* (version 2.9.0+) and identified any regions of the query genome that matched against the area in the reference that contained *bla*$_{\text{CTX-M-65}}$ (Appendix I.8) (McGinnis and Madden, 2004). The lengths of these matched areas were then compiled; *bla*$_{\text{CTX-M-65}}$ was 876 bases long, any that were over double the length of the gene were classified as being on pESI. Sequences that were below this cut-off or did not have a match for *bla*$_{\text{CTX-M-65}}$ were checked further to ensure the gene was not on pESI. They were blasted against both the pESI pseudomolecule and a pseudomolecule containing pESI and *bla*$_{\text{CTX-M-65}}$ using *Nucleotide BLAST* (version 2.9.0+), using the megaBLAST option and the parameter outfmt 6 to generate crunch files. The assemblies of these sequences were also uploaded to the *ResFinder* website (accessed 11.09.19) to identify the location of *bla*$_{\text{CTX-M-65}}$ (Zankari *et al.*, 2012). *ACT* (version 17.0.1) was then used to visualise where the areas surrounding *bla*$_{\text{CTX-M-65}}$ in the query genome mapped to in the reference; if large portions of the surrounding area mapped to pESI then *bla*$_{\text{CTX-M-65}}$ was defined as being on pESI (Carver *et al.*, 2005).

4.2.5.7 *Determining the Phylogenetic Structure of pESI*

In order for the new eBG31 pESI pseudomolecule to be used as the reference, a new *SnapperDB* was generated*;* the creation of the database and the VCFs was done with the same parameters as in Chapter 2.6.5 (Ashton *et al.*, 2017). Prophages were detected in pESI using the method described in Chapter 2.5.3. The coordinates of the only intact prophage were: ASRF01000101.1_contig_16: 49072-58872.

A soft-core SNP alignment was generated using the method described in Chapter 2.7.2. Representatives of each 25SNP cluster were used in the whole genome alignment generation and of each 5SNP cluster for the soft-core SNP alignment. The prophage was masked alongside recombination and the eBG31 reference, resulting in an alignment of pESI variation. The resulting soft-core SNP phylogeny was rooted to its most ancestral node. The scripts used to annotate the eBG31 phylogeny (Chapter 2.7.3) were amended and used to create multi value bar chart annotation files for use with *iToL* (Letunic and Bork, 2016). They were also used to annotate the eBG31 soft-core SNP phylogeny (Figure 3.9) with pESI presence and pESI variant.

4.2.6 Integron Presence

To identify the number of integrons found in the *S.* Infantis collection, *Integron Finder* (version 2) was installed on the UEA HPC (Cury *et al.*, 2016a). It was run using the scaffolds.fasta output of *SPAdes* (Chapter 2.6.6) with the local detection and promotor-attI parameters selected (Bankevich *et al.*, 2012). Only sequences that had been determined to have good assembly quality were included (Chapter 2.6.7). As the number of complete integrons present in each genome was of interest, I wrote a bash script to extract this number from each summary file (Appendix I.9). The *int* gene for each integron was pulled out using grep and awk commands to identify the class of each integron.

4.2.7 Analysis and Presentation of the Results

Comma-separated value (CSV) files were generated for eBG, origin, source, year group and ST. CSV files were also generated for: the presence of resistance determinants to each antibiotic class; complete antibiotic sensitivity; MDR; presence of each plasmid group, including each variant of pESI and number of integrons. *Scoary* (version 1.6.16) (Chapter 2.7.5) was used on *CLIMB* with the CSV files, the --no_pairwise option to not perform pairwise comparisons and the option --p 1.0 to not exclude by p-value and report all genes (Brynildsrud *et al.*, 2016). The results that were of interest were Number_pos_present_in and Number_pos_not_present_in. The two population proportions z test calculator was used calculate whether proportions were significantly different (Stangroom, 2019).

An alternative to Venn diagrams, the UpSet plot, was chosen to represent the number of shared and unique AMR and plasmid clusters. *R* (version 3.5.1) was used with *RStudio* (version 1.1.463) and the package *UpSetR* (version 1.4.0) (Lex *et al.*, 2014). The original *ARIBA* output, not grouped by class or type, was used as input; the variants of pESI were not included in the plasmid UpSet plots created.

4.3 Results

4.3.1 Antimicrobial Resistance

In order to investigate the presence of antibiotic resistance in *S.* Infantis, the *S.* Infantis collection was screened for AMR gene presence. *ARIBA* was run on the sequences with the *ResFinder* database to identify AMR gene presence and with a database of *gyrA*, *gyrB*, *parC* and *parE* genes to identify mutations within the QRDRs.

4.3.1.1 *Overall AMR Statistics*

Across *S.* Infantis 62 clusters of AMR genes were identified, encoding resistance to 11 different classes of antibiotics: aminoglycosides, beta-lactams, chloramphenicols, fosfomycins, lincosamides, macrolides, polymyxins, quinolones, sulphonamides, tetracyclines and trimethoprims (Appendix VI Table VI.3). 25 different mutations were also identified in the QRDRs of *gyrA*, *gyrB*, *parC* or *parE*.

60.2% (2810/4670) of the sequences did not contain any currently known AMR genes. 51 of these sequences were predicted to be resistant to quinolones due to mutations in the DNA gyrase and topoisomerase IV genes, resulting in 59.1% (2759/4670) of the *S.* Infantis isolates being predicted to be putatively susceptible to all known classes of antibiotic.

Whilst observed in both eBGs, resistance to macrolides was low, with 2% of eBG297 and 0.3% of eBG31 isolates containing resistance genes for the class. Resistance to lincosamides and polymyxins was only seen in 0.4% and 0.02% of eBG31 isolates respectively. The percentage of isolates resistant to macrolides, lincosamides or polymyxins were therefore not included in comparisons of resistance in antimicrobial classes but are shown in Appendix IV Table IV.1.

MDR, defined as resistance to 3 or more classes of antibiotic, was identified in the *S.* Infantis population. 37.3% (1740/4670) of the isolates had MDR; 19.1% (332/1740) of these sharing a common antibiotic resistance class profile with resistance to aminoglycosides, quinolones, sulphonamides, tetracyclines and trimethoprims.

4.3.1.2 *AMR Variation by eBG*

To identify whether there was a difference in AMR occurrence between eBG31 and eBG297 the presence of resistance to antibiotic classes was determined and plotted in Figure 4.1.

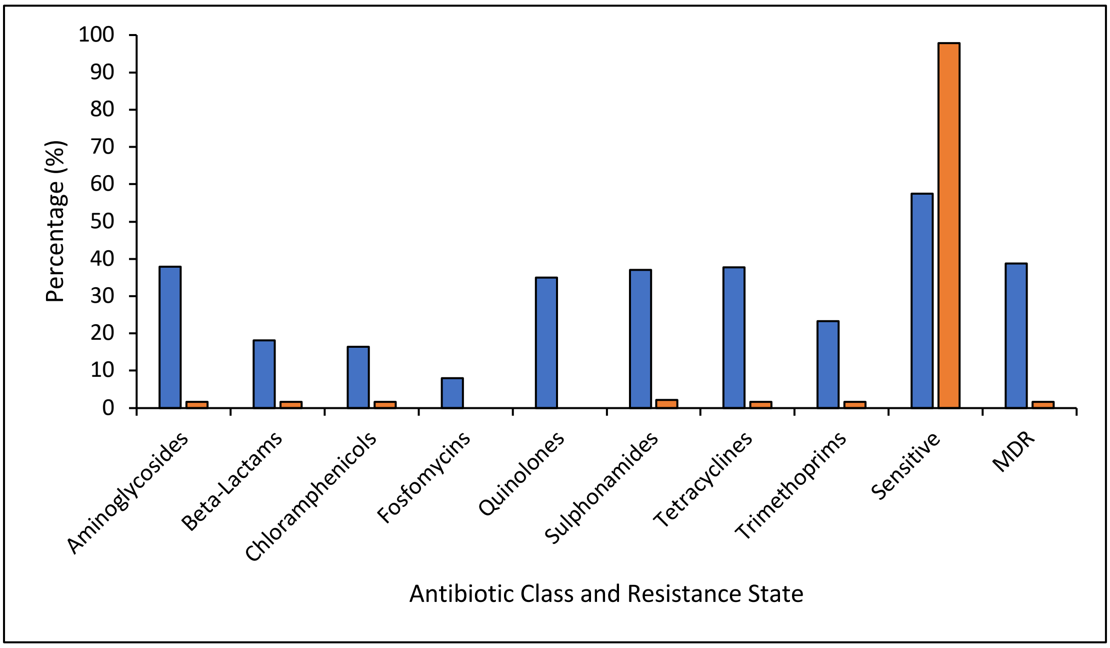


Figure 4.1 Difference in AMR between eBG31 and eBG297
eBG31 (n = 4486)🟦 eBG297 (n = 184)🟧

A significant difference was seen in the number of sensitive isolates; 57.5% (2579/4486) of eBG31 isolates compared to 97.8% (180/184) of eBG297 isolates were putatively susceptible to all known antimicrobials ($p < 0.00001$).

MDR was significantly more common in eBG31 with 38.7% (1737/4486) of the isolates positive versus 1.6% (3/184) of eBG297 isolates ($p < 0.00001$). Genes involved in resistance to aminoglycosides, quinolones, sulphonamides and tetracyclines were most frequently identified in eBG31; 31.5% (1414/4486) of the isolates harboured resistance to these four classes, either exclusively or in combination with other classes.

Whilst the eBG31 population contained 53 gene clusters unique to that eBG, eBG297 did not contain any (Figure 4.2). Only 9 AMR gene clusters were found in both eBGs: *aac*(3)II, *aph*(3')-Ib, *aph*(6)-Id, *bla*$_{\text{CTX-M}}$, *dfrA*14, *floR*, *mphA*, *sul*2 and *tet(A)*.

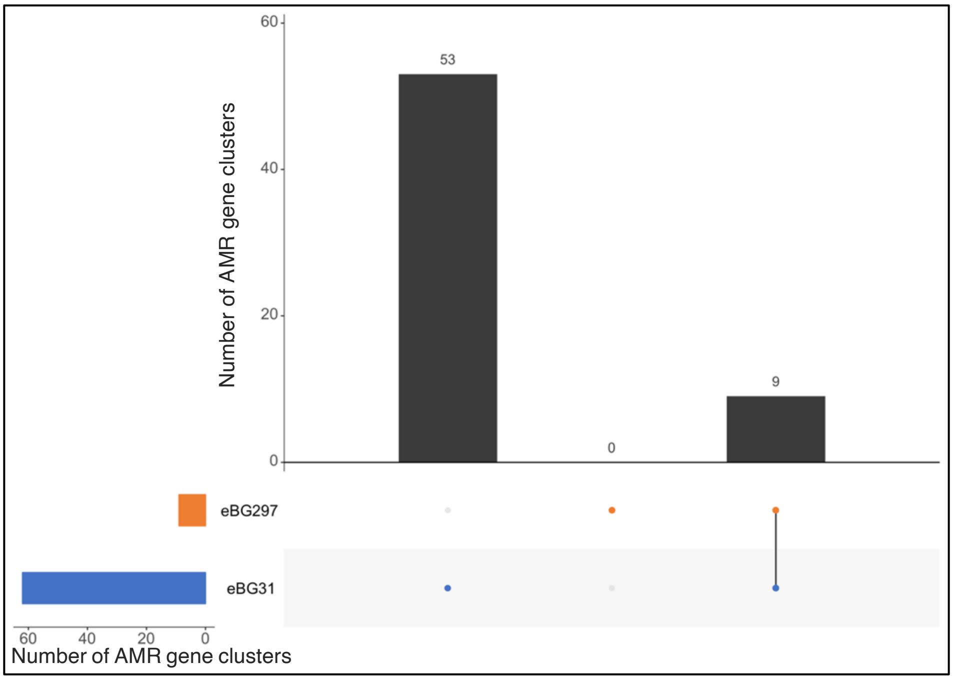


Figure 4.2 Distribution of AMR gene clusters between eBG31 and eBG297
UpSet plot showing the number of AMR gene clusters unique to and shared between eBG31 and eBG297 isolates.
eBG31 🟦 eBG297 🟧

4.3.1.3 *AMR Variation by Isolation Source*

The distribution of AMR across the *S.* Infantis isolation sources was calculated to determine whether the levels of resistance varied by source (eBG31, Figure 4.3).

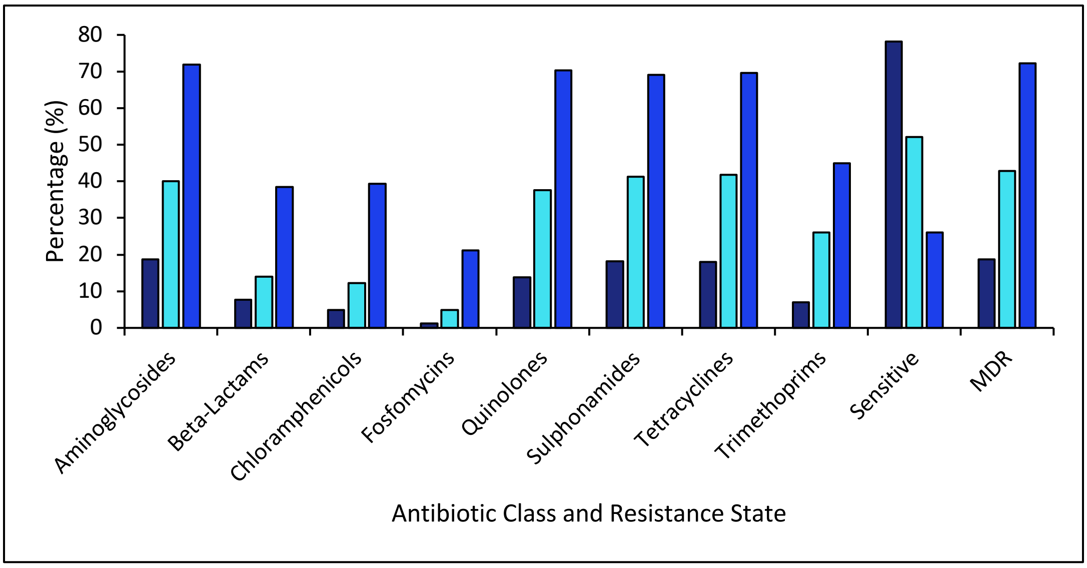


Figure 4.3 eBG31 AMR distribution by isolation source
Shown as a percentage of isolates from each source: environmental (n=947); human (n=1519) and poultry (n=947).
Environmental 🟦 Human 🟦 Poultry 🟦

For all reported antibiotic classes, the number of isolates from poultry with AMR determinants was significantly larger than the other sources (p<0.00001). 65.9% (624/947) of eBG31 poultry isolates were resistant to aminoglycosides, quinolones, sulphonamides and tetracyclines compared with 32.4% (492/1519) of eBG31 human isolates and 12.7% (120/947) of eBG31 environmental isolates. The highest levels of MDR were seen in eBG31 isolates from poultry sources at 72.2% (684/947); this were significantly higher than the levels in isolates from human and environmental sources (p<0.00001). The lowest levels of MDR were seen in the environmental isolates at 18.7% (177/947).

The eBG297 isolates from environmental, unknown sources and 97.6% (164/168) of isolates from humans were sensitive to all the antibiotic classes. Only 1.8% (3/168) of the eBG297 isolates from humans had MDR.

The distribution of AMR gene clusters across sources in eBG31 is shown in Figure 4.4. 40.3% (25/62) of the AMR gene clusters found in eBG31 isolates were found in isolates from all sources. Interestingly, despite having the largest amount of AMR, only 3.2% (2/62) of the gene clusters were found exclusively in isolates from poultry. A larger percentage, 22.6% (14/62), of AMR gene clusters were found to be unique to eBG31 isolates from humans. Notably, more clusters were shared between human and environmental isolates than human and poultry isolates.

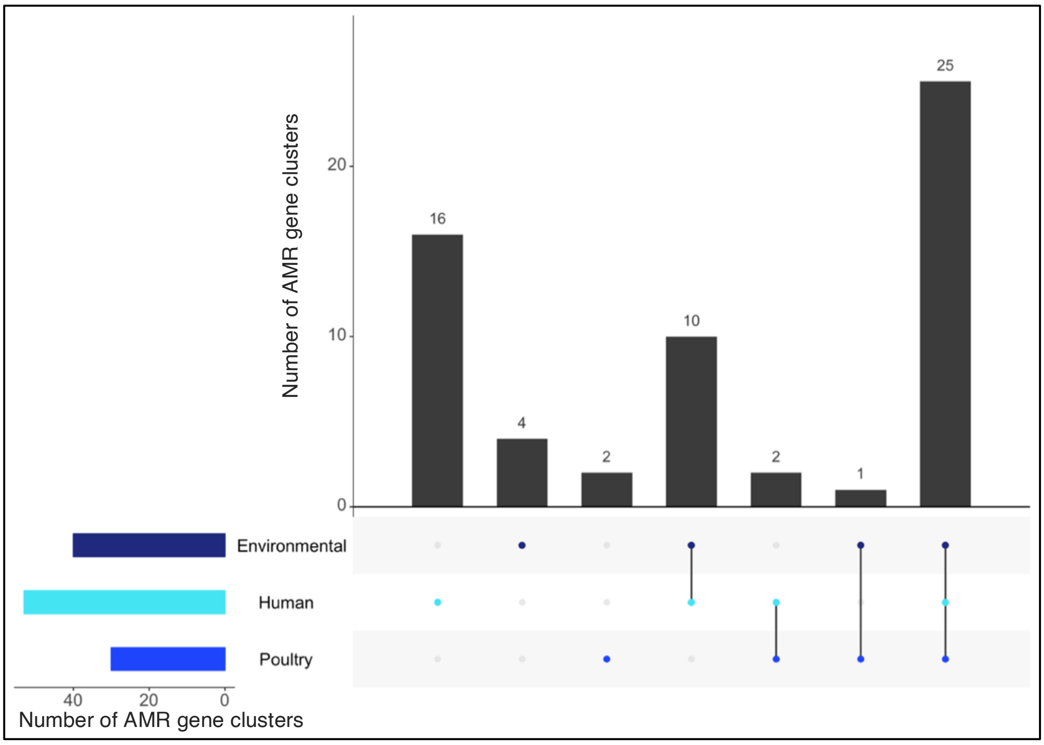


Figure 4.4 Distribution of AMR gene clusters between eBG31 isolation sources
UpSet plot showing the number of AMR gene clusters unique to and shared between eBG31 isolation sources.
Environmental ■ Human ■ Poultry ■

4.3.1.4 *AMR Variation by Origin*

In order to determine the geographical distribution of AMR in *S.* Infantis, the geographical differences in frequency of AMR and MDR in *S.* Infantis isolates were compared (Figure 4.5). Levels of resistance to antibiotic classes varied by continent of eBG31 isolation. Isolates from South America had the highest levels of AMR, with the largest percentage of resistant isolates for all classes. The isolates from that continent and from returning travellers also had the lowest number of isolates susceptible to all antibiotics (15.6%, 19/122) although this was closely followed by Asia with 18.5% (44/238). Conversely, low levels of AMR were identified in the African isolates, with 78.1% (232/297) sensitive to all antimicrobials.

A common AMR profile is visible with resistance to aminoglycosides, quinolones, sulphonamides and tetracyclines seen at similar levels within isolates from each continent, excluding Africa. For example, the percentage of isolates from South America that were resistant to aminoglycosides, quinolones, sulphonamides and tetracyclines were 80.3% (98/122), 78.7% (96/122), 77% (94/122) and 79.5% (97/122) respectively. MDR varied substantially by continent of isolation, ranging from 18.9% (56/297) in African eBG31 isolates to 81.1% (99/122) in South American eBG31 isolates.

AMR was only observed in eBG297 isolates that had been found in Africa, with isolates from all other continents being pan-susceptible. 97.4% (151/155) of African eBG297 isolates also had no AMR determinants. 1.9% (3/155) of the African eBG297 were MDR; these 3 isolates contained genes encoding resistance to aminoglycosides, extended-spectrum beta-lactams, chloramphenicols, macrolides, sulphonamides, tetracyclines and trimethoprims.

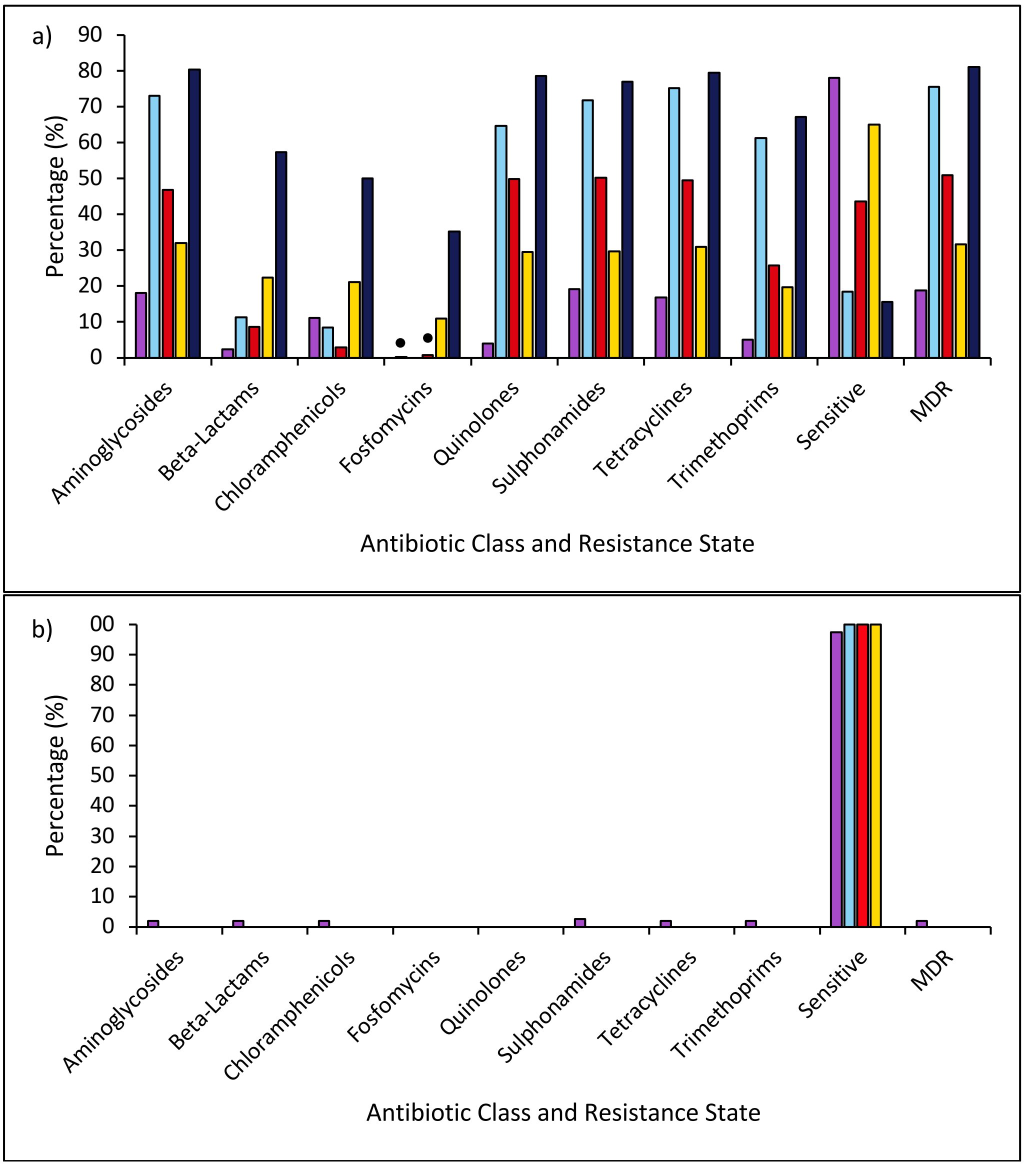


Figure 4.5 Difference in AMR distribution by continent

Shown as a percentage of isolates from each continent. • values are less than 1.5% but greater than 0.

a) eBG31: Africa (n=297); Asia (n=238); Europe (n=959); North America (n=2793) and South America (n=122)

b) eBG297: Africa (n=155); Asia (n=3); Europe (n=20); North America (n=2) and South America (n=0)

Africa 🟪 Asia 🟦 Europe 🟥 North America 🟨 South America ⬛

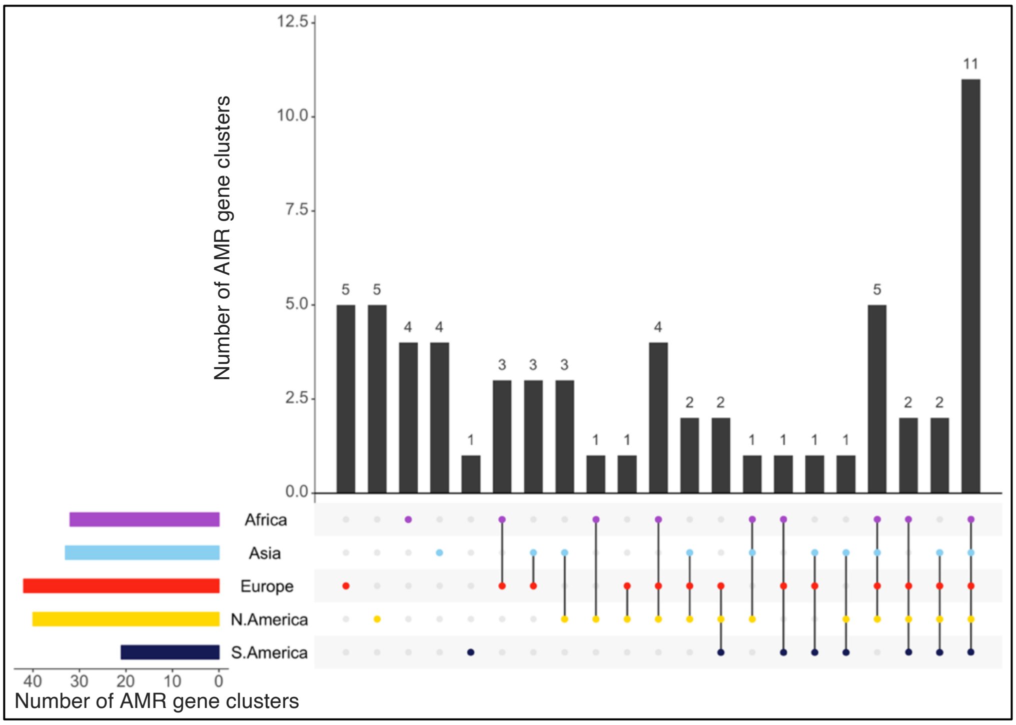


Figure 4.6 Distribution of AMR gene clusters across eBG31 by continent
UpSet plot showing the number of AMR gene clusters unique to and shared between eBG31 isolates from each continent.
Africa 🟪 Asia 🟦 Europe 🟥 North America 🟨 South America ⬛

Despite eBG31 isolates from Europe and North America having lower percentages of isolates with AMR, they had the highest number of AMR gene clusters unique to each continent (Figure 4.6). AMR genes unique to African, Asian and South American eBG31 isolates were also identified. Only 17.7% (11/62) of the AMR genes were found in isolates from all the continents.

4.3.1.5 *AMR Variation by Year*

The occurrence of AMR in *S.* Infantis in each isolation year group was compared to identify whether there were trends by time (Figure 4.7). Resistance to all of the included classes of antibiotic was seen for every year group in eBG31, with the exception of fosfomycins which was identified in isolates from 2006-2010 onwards. The highest frequency of MDR in eBG31 isolates was seen between 2017-2018, at 52.8% (615/1164), with significantly higher levels seen than in any other year group ($p<0.05$ for 1989-2005 vs. 2017-2018, $p<0.00001$ for all other comparisons). The percentage of isolates with resistance to each included antibiotic class was highest for this most recent year group with the exception of trimethoprims. The percentage of eBG31 isolates with resistance to beta-lactams, fosfomycins and quinolones increased consistently throughout the time periods in this project. Also, the levels of resistance to aminoglycosides, sulphonamides, tetracyclines and overall levels of MDR appear to be associated, with similar levels seen in each year group.

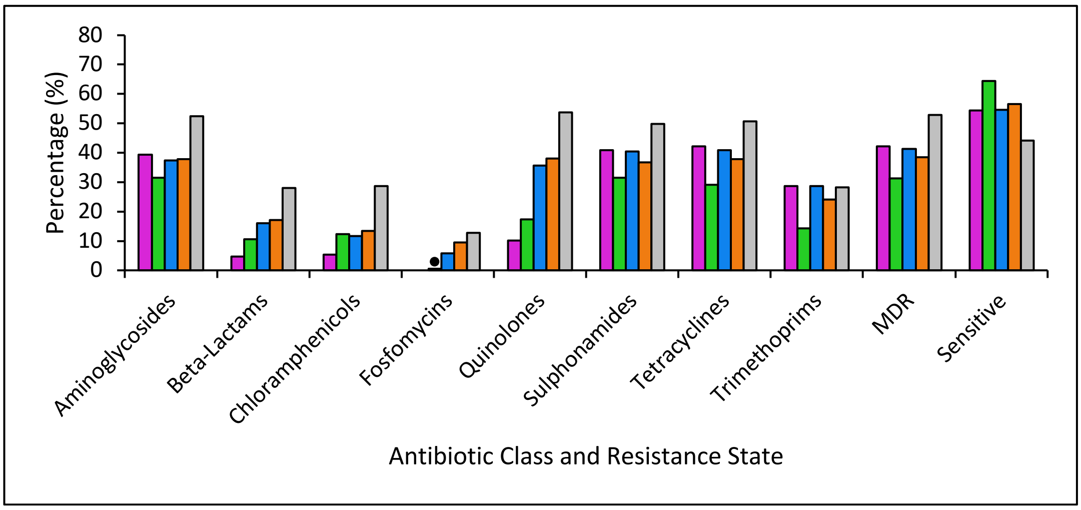


Figure 4.7 Difference in AMR in eBG31 by year group
Shown as a percentage of isolates from each continent. • values are less than 1% but greater than 0.
1989-2005 (n=147) 🟪 2006-2010 (n=404) 🟩 2011-2014 (n=759) 🟦 2015-2016 (n=1203) 🟧
2017-2018 (n=1164) ⬜

No eBG297 isolates with AMR were identified prior to 2011. Seven gene clusters were identified solely in isolates from 2011-2014; the other two resistance gene clusters were found between both 2011-2014 and 2015-2016. The 3 MDR eBG297 strains were isolated from South Africa between July and October 2012, with the remaining isolate with resistance to macrolides and sulphonamides being isolated from South Africa in 2015.

The distribution of eBG31 AMR gene clusters across the different time periods is shown in Figure 4.8. 25.8% (16/62) of the AMR gene clusters were found in isolates in every time period, including genes which encoded resistance to aminoglycosides, chloramphenicols, sulphonamides, tetracyclines, trimethoprims and beta-lactams, including extended-spectrum beta-lactams. The number of AMR gene clusters peaked in 2015-2016, where isolates contained resistance genes for all of the 11 antibiotic classes identified. 18 AMR gene clusters identified in 2015-2016 were no longer present in eBG31 isolates in 2017-2018, these genes encoded resistance to the following antimicrobial classes: aminoglycosides, beta-lactams, chloramphenicols, sulphonamides, trimethoprims, lincosamides, macrolides, polymyxins, quinolones and tetracyclines. The largest number of genes unique to a time period was also seen between 2015-2016, including gene clusters such as *blaSHV*, *catA1*, *lnu(F)*, *mcr1* and *qnrD*. Interestingly, whilst isolates from 2017-2019 had the highest levels of MDR (52.8%, 615/1164), there were no AMR gene clusters that were found exclusively in this time period.

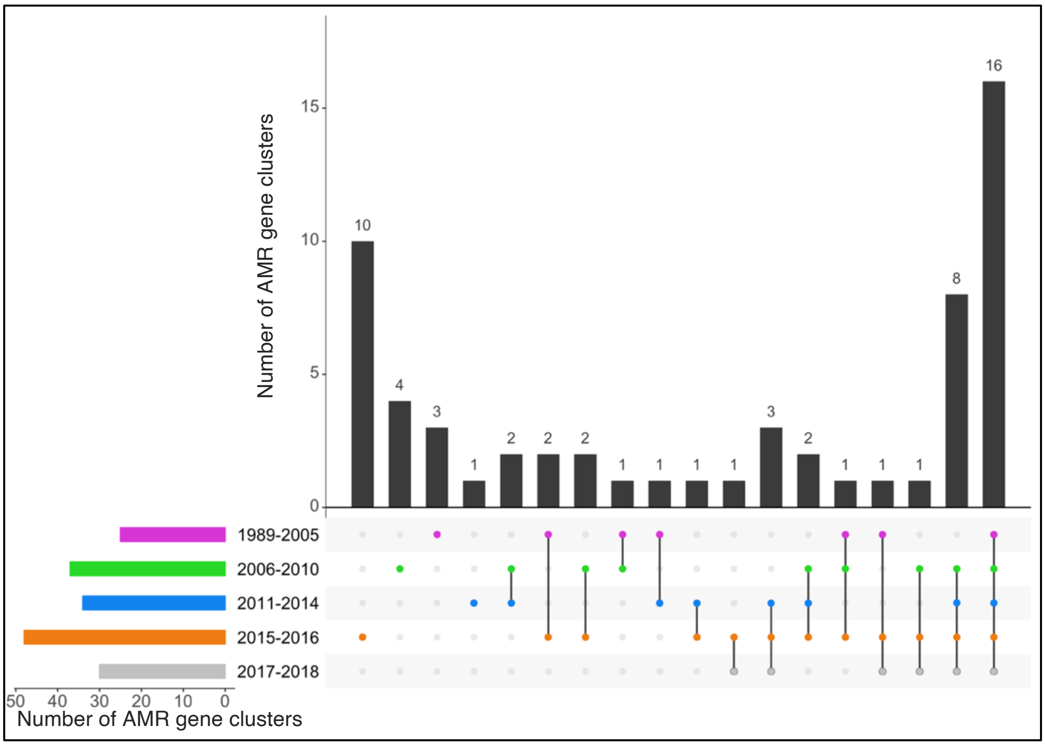


Figure 4.8 Distribution of AMR gene clusters across eBG31 by year group
UpSet plot showing the number of AMR gene clusters unique to and shared between eBG31 isolates from each year group.
1989-2005 🟪 2006-2010 🟩 2011-2014 🟦 2015-2016 🟧 2017-2018 ⬜

4.3.2 Mobile Genetic Elements

To establish the distribution of plasmids across the *S.* Infantis population, the *S.* Infantis collection was screened for the presence of plasmids. *ARIBA* was run with the *PlasmidFinder* database and the genomes were screened separately for the presence of pESI (Appendix VI Table VI.4).

46 different plasmid types were matched from the *PlasmidFinder* database. These were grouped by Inc type, resulting in 24 plasmid groups. While the number of plasmids found in each isolate varied between 0 and 5, 86% (4013/4670) lacked any plasmids from the *PlasmidFinder* database. Of the 24 plasmid groups, the following were identified in fewer than 1% of eBG31 or eBG297 isolates: Col156, Col3M, Col440I, Col8282, Col(RNAI), Col(BS512), Col(IMGS31), Col(MG828), Col(pVC), IncB/O/K/Z, IncH, IncL/M, IncN, IncQ, IncR, IncU, p0111 and repA. These plasmids groups were therefore not included in graphs

discussing results grouped by Inc type, their results can be found in Appendix IV Table IV.2, IV.3, IV.4, IV.5.

pESI was found in 33% (1541/4670) of the *S.* Infantis genomes (Appendix VI Table VI.5). Therefore, 55.6% (2597/4670) of the *S.* Infantis genomes did not contain plasmids from either the *PlasmidFinder* database or pESI search. Running *Integron Finder* on the Israeli pESI plasmid identified two class 1 integrons, the first 1,863 bases in length and containing the *dfrA14* trimethoprim resistance gene (Integron A); the second 2,465 bases long and containing *aadA1,* the streptomycin and spectinomycin resistance gene (Integron B). Heatmaps of pESI showed that these integrons were absent from some of the pESI positive isolates (Appendix IV Figure IV.2). 40.2% (620/1541) of the pESI positive isolates were missing Integron A and 6.1% (94/1541) were missing Integron B. The location of *bla*$_{\text{CTX-M-65}}$ in pESI positive isolates was identified to determine whether the gene was carried on the plasmid. 38.3% (590/1541) of the pESI positive isolates had *bla*$_{\text{CTX-M-65}}$ on the plasmid.

86.7% (9177/10588) of the AMR genes that were identified by *ARIBA* in the eBG31 isolates were in isolates that contained pESI. Furthermore, of the 1549 isolates that had quinolone resistance due to mutations in *gyrA*, *gyrB*, *parC* or *parE*; 95.3% (1476/1549) contained pESI. Also, of the 1737 of the eBG31 isolates that had MDR; 87.1% (1513/1737) were positive for pESI.

4.3.2.1 *Plasmid Variation by eBG*

To determine the difference in plasmid prevalence by eBG, the occurrence of each plasmid group and variant of pESI in *S.* Infantis was plotted (Figure 4.9).

Whilst 54.1% (2428/4486) of eBG31 were without a plasmid, 91.8% (169/184) of eBG297 isolates didn't contain one. A higher proportion of eBG31 isolates were positive for each included plasmid group than seen in the eBG297 isolates, with the exception of IncF and pSL483.

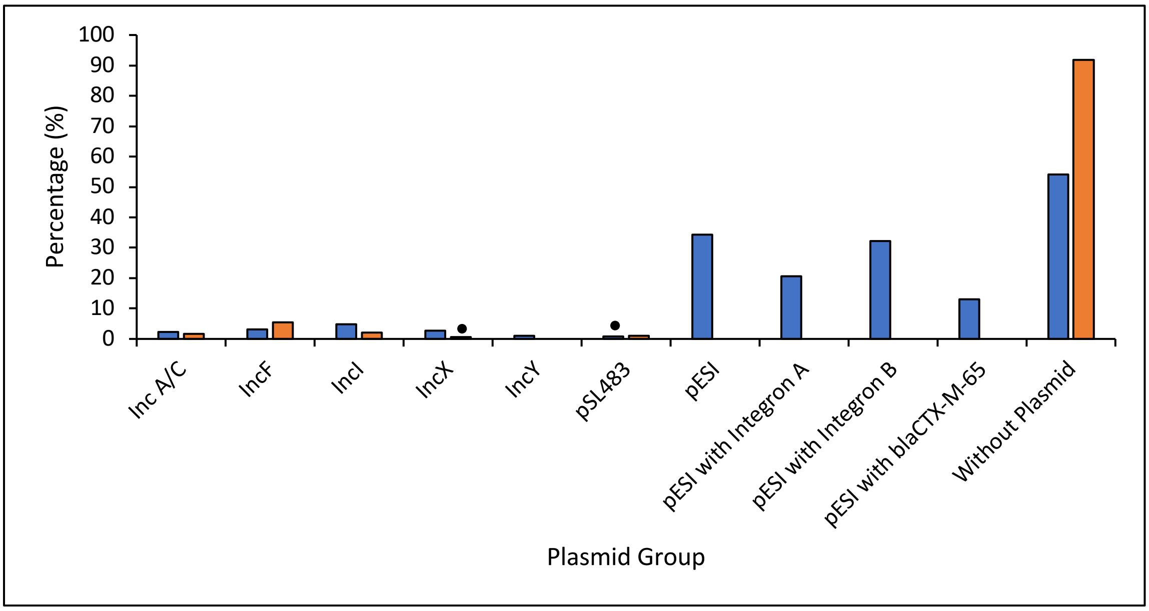


Figure 4.9 Difference in plasmids between eBG31 and eBG297
eBG31 (n=4486) and eBG297 (n=184). • values are less than 1% but greater than 0.
eBG31 🟦 eBG297 🟧

The variation in the types of plasmid found in eBG31 and eBG297 is shown in Figure 4.10. Unsurprisingly, considering the higher number of plasmids in eBG31 isolates, a large number of plasmids unique to that eBG were identified. The only plasmid group that was found solely in eBG297 was Col(IMGS31), however only one of the eBG297 isolates contained this, a strain which was isolated from South Africa in 2004.

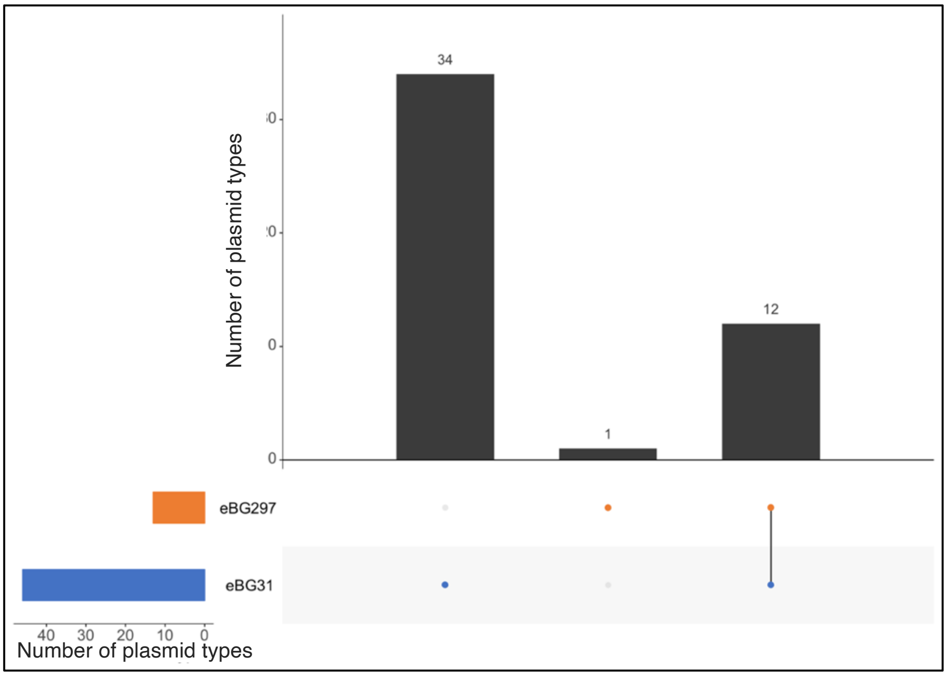


Figure 4.10 Distribution of plasmid types shared by eBG31 and eBG297
UpSet plot showing the number of plasmid types unique to and shared between eBG31 and eBG297 isolates.
eBG31 🟦 eBG297 🟧

4.3.2.2 *Plasmid Variation by Isolation Source*

The distribution of plasmid groups from each isolation source of *S.* Infantis was compared to quantify how the frequency of plasmid presence and type varies by source (Figure 4.11).

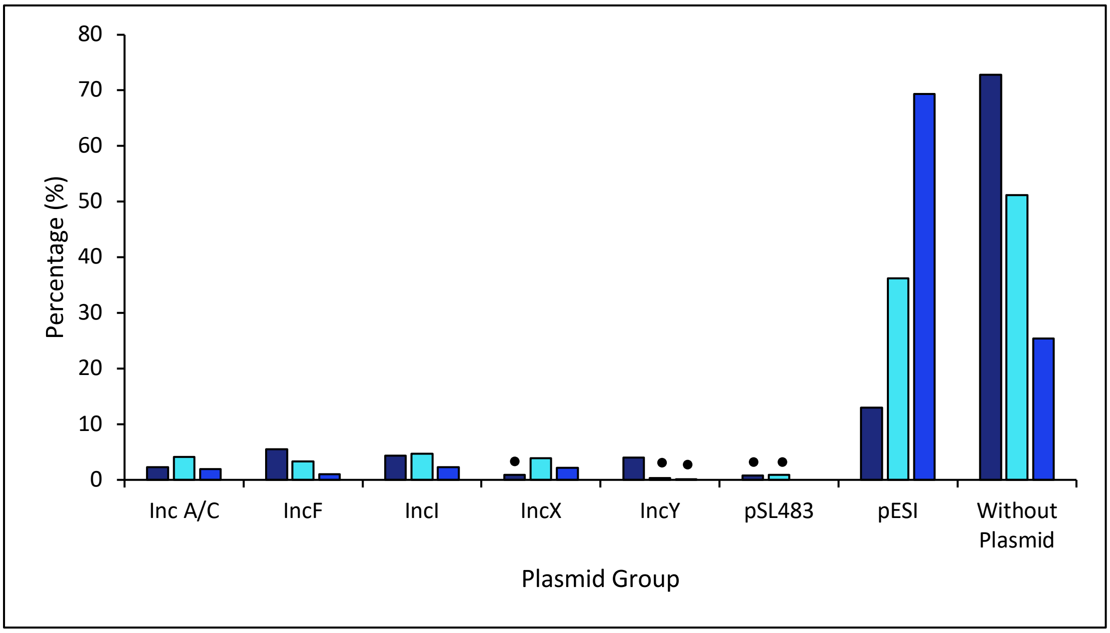


Figure 4.11 Difference in plasmids between eBG31 isolation sources
Shown as a percentage of isolates from each source. • values are less than 1% but greater than 0.
Environmental (n=947) ■ Human (n=1519) ■ Poultry (n=947) ■

Although eBG31 isolates from environmental sources had the highest percentage of isolates with IncF and IncY plasmids, 5.5% (52/947) and 4.0% (38/947) respectively, it was also the isolation source with the largest percentage of isolates without pESI and without any plasmid at all. For all plasmid groups except for pESI, the proportion of human eBG31 isolates with the plasmid was greater than seen in eBG31 isolates from poultry sources. However, for pESI the opposite was true, significantly more of the poultry eBG31 isolates (657) were positive for the plasmid than human isolates (551) (p-value < 0.00001).

Correlating with the lack of AMR genes in eBG297 isolates from environmental and unknown sources, no plasmids were identified in eBG297 isolates from these sources. The majority of the plasmid groups identified from human eBG297 isolates were only identified in one isolate.

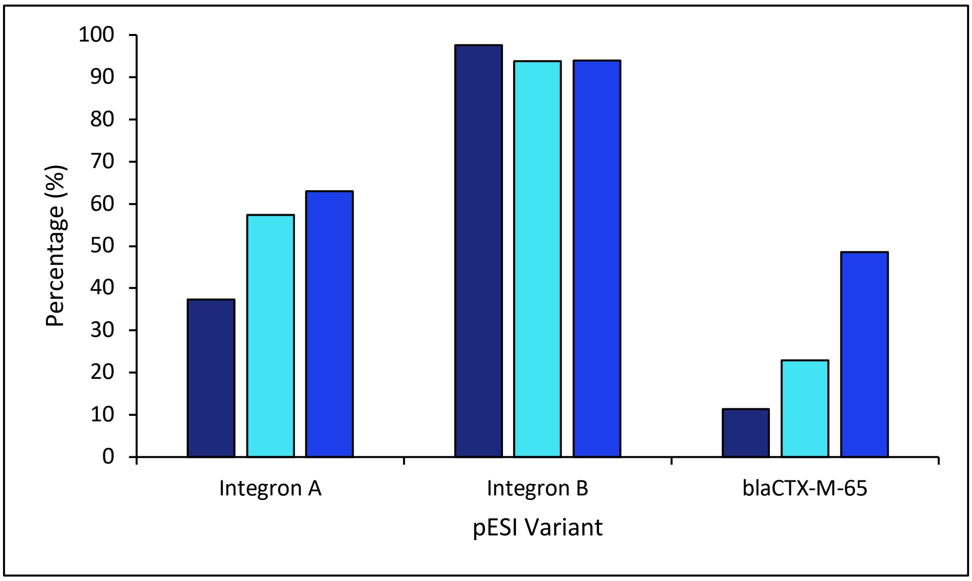


Figure 4.12 Difference in pESI variant between eBG31 isolation sources
Shown as a percentage of pESI positive isolates from each source: environmental (n=123); human (n=551) and poultry (n=657). pESI was only identified in eBG31.
Environmental ■ Human ■ Poultry ■

The frequency of each variant of pESI occurring in each of the isolation sources was calculated (Figure 4.12). Greater variation was seen in the frequency of presence of Integron A than Integron B, which appeared to be more conserved across pESI. Whilst similar frequencies of Integron A were seen in eBG31 isolates from human and poultry sources, it was less commonly found in isolates from environmental sources. Of the eBG31 human pESI positive isolates, 5.4% (30/551) were isolated from urine. There was no association with isolation from urine and presence of the trimethoprim resistance encoding Integron A; 43.3% (13/30) did not contain the integron and 56.7% (17/30) did. Concerningly, 48.6% of the 657 eBG31 pESI positive isolates from poultry sources were found to be carrying the *bla*$_{\text{CTX-M-65}}$ gene on the plasmid.

The distribution of plasmid types across eBG31 isolation sources is plotted in Figure 4.13. Despite poultry isolates having the smallest percentage of isolates without a plasmid, none of the plasmid groups were found uniquely in isolates sourced from poultry. Also, more plasmid types were identified in eBG31 isolates from humans than any other source, with that source having more plasmids unique to it and having more shared plasmids than poultry and environmental isolates.

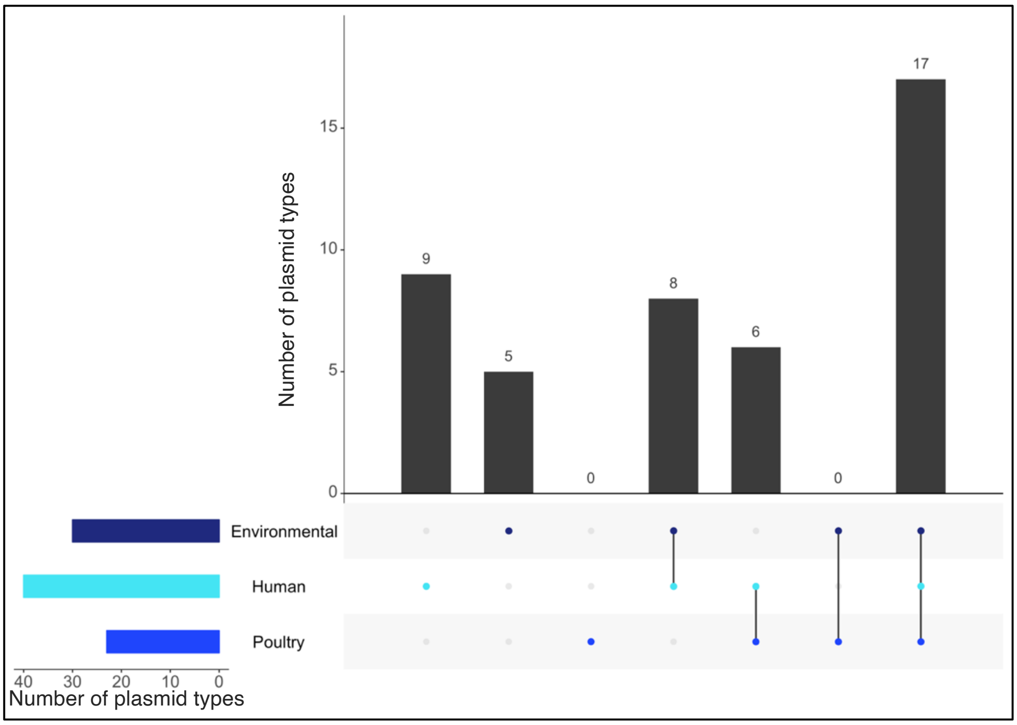


Figure 4.13 Distribution of plasmid types between eBG31 isolation sources
UpSet plot showing the number of plasmid types unique to and shared between eBG31 isolation sources.
Environmental ■ Human ■ Poultry ■

4.3.2.3 *Plasmid Variation by Origin*

With a view to identify the geographical variation in plasmid presence across *S.* Infantis isolates, the occurrence of each included plasmid group from each continent was plotted in Figure 4.14.

Given the lack of AMR in the eBG31 African sequences, it was also the continent with the largest percentage of isolates without a plasmid at 73.7% (219/297). The plasmid group most frequently present in the African eBG31 sequences was IncA/C at 13.8% (41/297). One of African IncA/C positive sequences was from PHE where the patient had travel history to Africa, the remaining 40 were all isolated from South Africa from human or environmental sources between 2005 and 2015.

For all other continents the plasmid type most frequently identified was pESI; 799 isolates with pESI were identified from North America, 456 from Europe, 162 from Asia, 94 from South America and 5 from Africa.

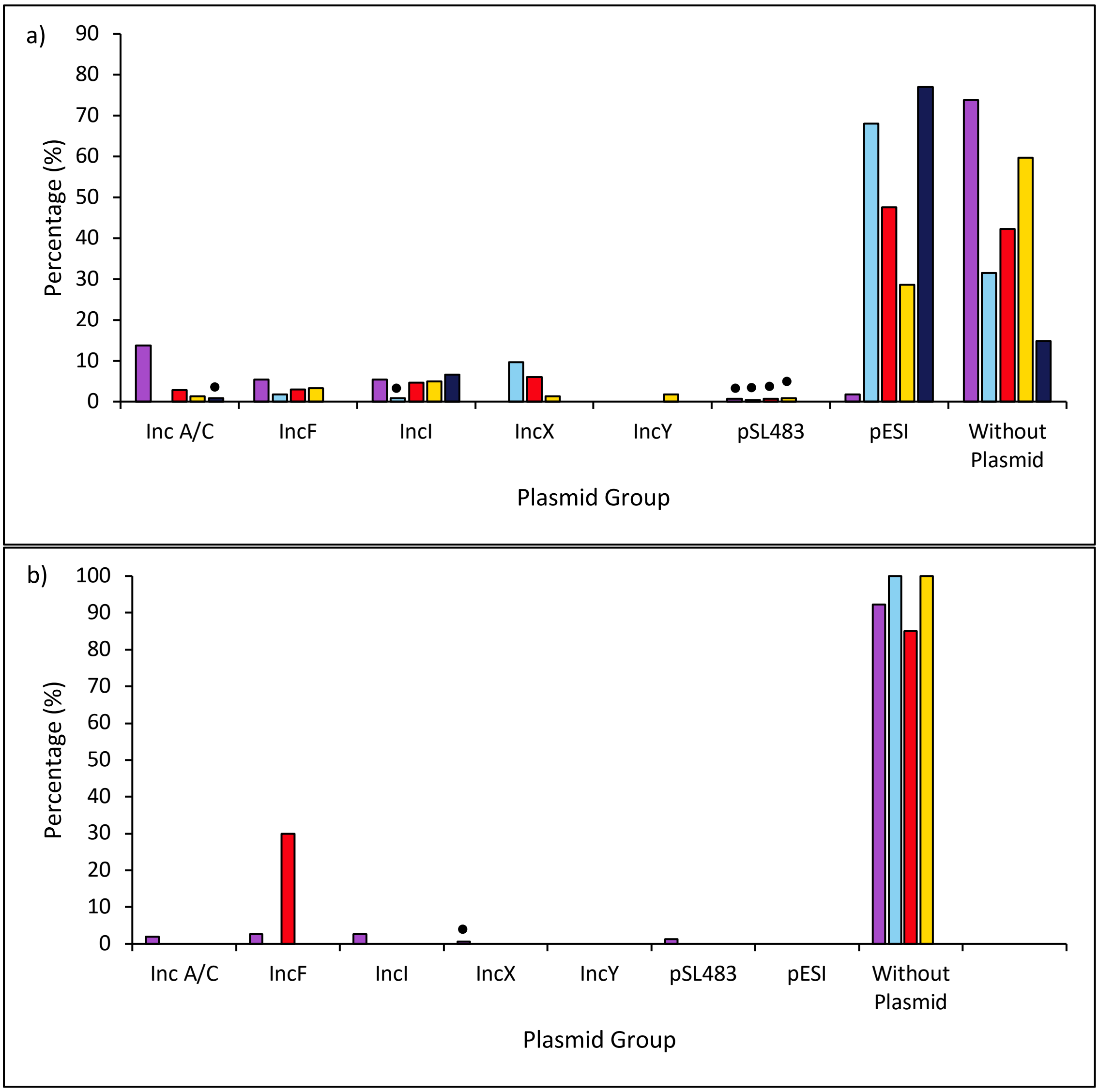


Figure 4.14 Difference in plasmid type between *S.* Infantis by continent

Shown as a percentage of isolates from each continent. • values are less than 1% but greater than 0.

a) eBG31: Africa (n=297); Asia (n=238); Europe (n=959); North America (n=2793) and South America (n=122)

b) eBG297: Africa (n=155); Asia (n=3); Europe (n=20); North America (n=2) and South America (n=0)

Africa 🟪 Asia 🟦 Europe 🟥 North America 🟨 South America ■

In the eBG297 population, the isolates from Asia and North America lacked any plasmids. The only plasmid group found in the European sequences was IncF which was present in 6 of the 20 samples. Within the African eBG297 samples plasmid groups, IncF, IncI and IncA/C were the most frequently identified, these were also most frequently found in the eBG31 African isolates.

The distribution of pESI variants in each of the continents is shown in Figure 4.15. As seen with isolation source, the percentage of isolates with Integron B does not vary greatly between continents. However much greater variation is seen in the number of isolates with Integron A. A low percentage was seen in the African pESI sequences but as only 5 African sequences contained pESI this was of little significance. 41.9% (191/456) of the European pESI positive isolates contained Integron A, whilst 86.2% (81/94) of the South American sequences contained it.

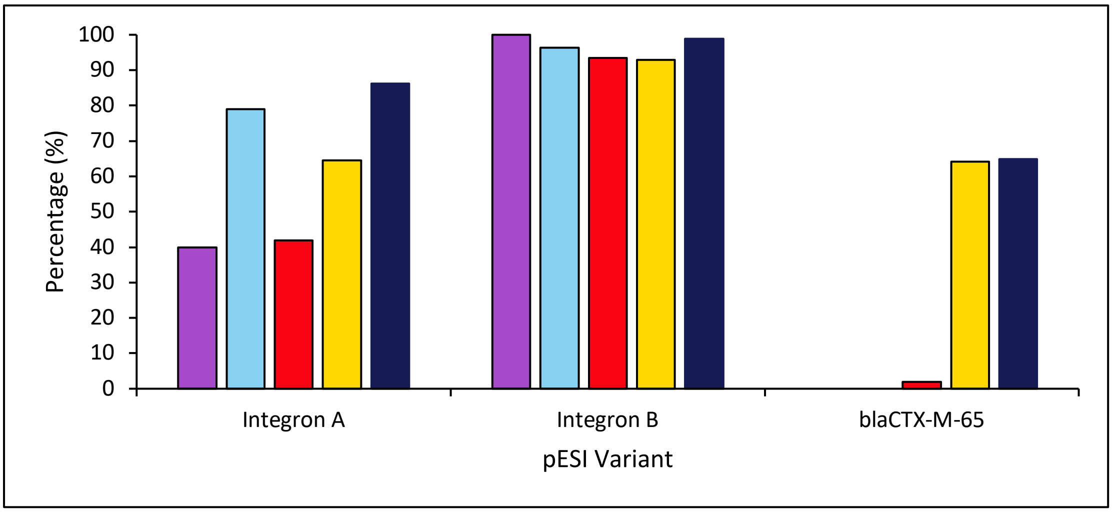


Figure 4.15 Difference in pESI variant across eBG31 by continent
Shown as a percentage of pESI positive isolates from each continent: Africa (n=5); Asia (n=162); Europe (n=456); North America (n=799) and South America (n=94).
Africa 🟪 Asia 🟦 Europe 🟥 North America 🟨 South America ⬛

Variation was also seen in the percentage of pESI positive isolates with the *bla*$_{\text{CTX-M-65}}$ gene on the plasmid from each continent. Whilst none of the Africa and Asian sequences and very few of the European sequences had the gene on pESI, 64.1 % (512/799) of the North American and 67% (63/94) of the South American isolates did.

The distribution of types of plasmid from eBG31 isolates in each continent is plotted in Figure 4.16. No plasmid groups were found exclusively in the South American, Asian or African eBG31 isolates. The only plasmid groups that were found in every continent were IncI plasmids and pESI. The largest group with shared plasmids was North America and Europe, which shared 21.7% (10/46) of the plasmid groups. In eBG297, 11 plasmid groups were found in isolates from Africa and 2 in isolates from Europe. No plasmid groups were shared between the eBG297 isolates from different continents.

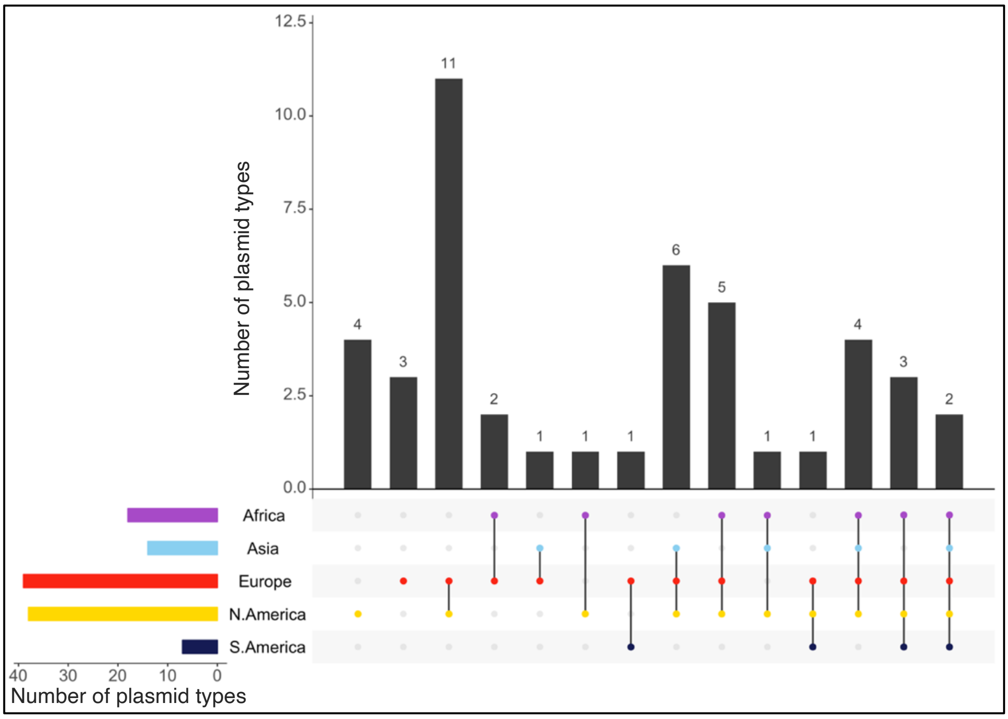


Figure 4.16 Distribution of plasmid types across eBG31 by continent
UpSet plot showing the number of plasmid types unique to and shared between eBG31 continents of isolation.
Africa 🟪 Asia 🟦 Europe 🟥 North America 🟨 South America ⬛

A summary of the AMR and pESI results for *S.* Infantis isolates from different sources and continents is shown in Table 4.3.

		eBG31				eBG297			
		Sensitive	MDR	pESI	n	Sensitive	MDR	pESI	n
eBG		57.5	38.7	34.4	4486	97.8	1.6	0	184
Source	Environmental	78.2	18.7	13.0	947	100	0	0	9
	Human	52.1	42.8	36.3	1519	97.6	1.8	0	168
	Poultry	26.0	72.2	69.4	947	0	0	0	0
Origin	Africa	78.1	18.9	1.7	297	97.4	1.9	0	155
	Asia	18.5	75.6	68.1	238	100	0	0	3
	Europe	43.7	51.0	47.5	959	100	0	0	20
	North America	65.1	31.7	28.6	2793	100	0	0	2
	South America	15.6	81.1	77.0	122	0	0	0	0

Table 4.3 AMR and pESI levels in *S.* Infantis
Percentage of isolates from each eBG, isolation source and continent that were susceptible to all known antimicrobials, had MDR and contained pESI. n = total number of isolates from each eBG, source group and continent.

4.3.2.4 *Plasmid Variation by Year*

To investigate trends in plasmid presence by time in *S.* Infantis, the occurrence of plasmids in *S.* Infantis in each year group is plotted in Figure 4.17.

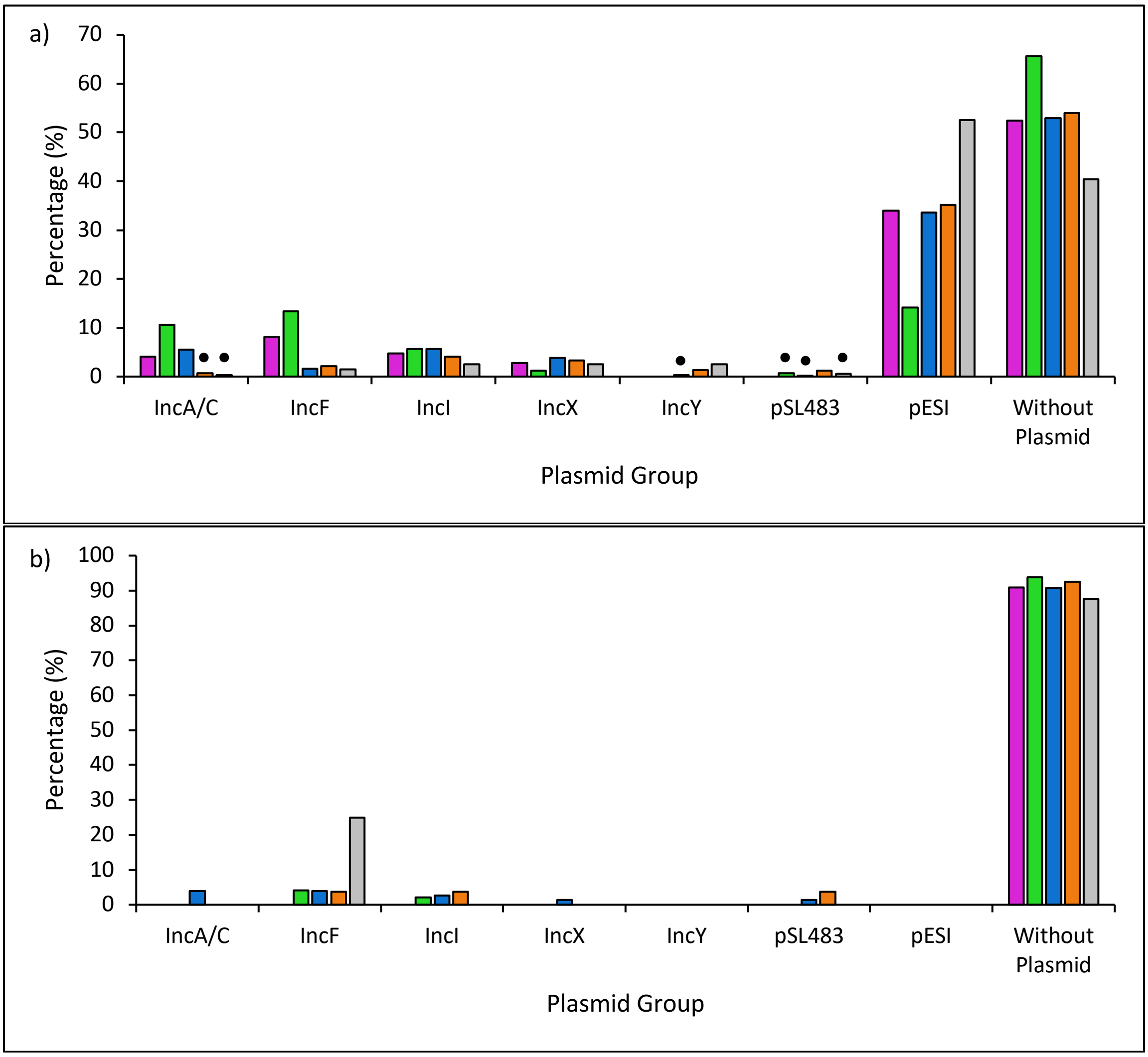


Figure 4.17 Difference in plasmids between *S.* Infantis by year group

Shown as a percentage of isolates from each year group. • values are less than 1% but greater than 0.

a) eBG31: 1989-2005 (n=147), 2006-2010 (n=404), 2011-2014 (n=759), 2015-2016 (n=1203), 2017-2018 (n=1164)

b) eBG297: 2003-2005 (n=11), 2006-2010 (n=49), 2011-2014 (n=75), 2015-2016 (n=27), 2017-2019 (n=16)

1989-2005 🟪 2006-2010 🟩 2011-2014 🟦 2015-2016 🟧 2017-2019 ⬜

Four of the plasmid groups were identified in eBG31 isolates in every year group; IncA/C, IncF, IncI and IncX were present in less than 10% of isolates from every year group, with the exception of IncA/C and IncF in 2006-2010. pESI was also identified in isolates from every year group, increasing in frequency in the past 10 years. The earliest isolation date of an isolate with pESI was 1999, four pESI positive isolates were found in this year, all from humans in Japan.

In eBG297 isolates none of the included plasmid groups were identified prior to 2006; of the rare plasmids, Col(IMGS31) was identified from South Africa in 2004. Whilst IncA/C and IncX were found exclusively in one year group, IncI and IncF were both maintained in the eBG297 population, with the former present in isolates between 2006-2016 and the latter 2006-2019. However, these were both at low numbers, with 1 or 2 isolates with the plasmid found each year.

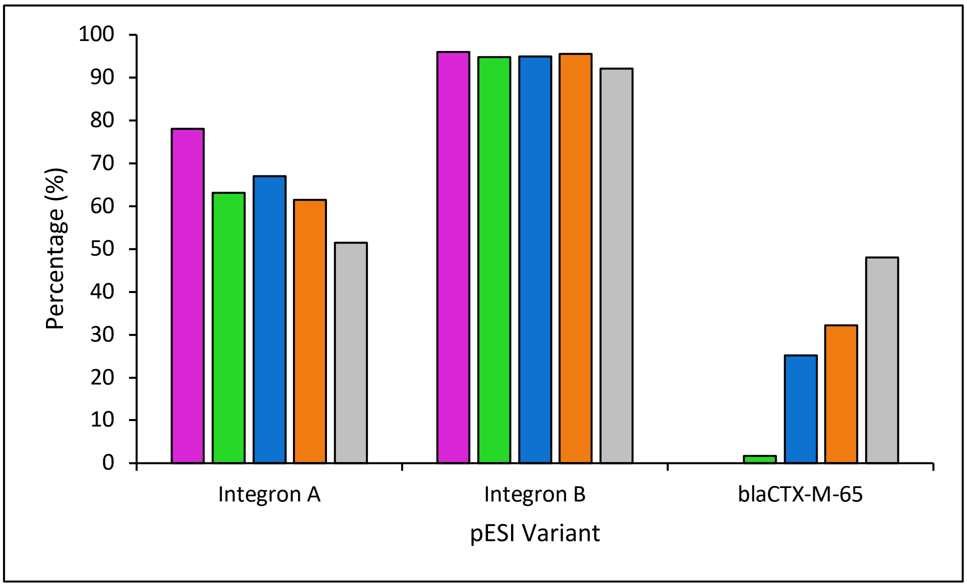


Figure 4.18 Difference in pESI variant across eBG31 by year group
Shown as a percentage of pESI positive isolates from each year group: 1989-2005 (n=50), 2006-2010 (n=57), 2011-2014 (n=255), 2015-2016 (n=423), 2017-2018 (n=612).
1989-2005 🟪 2006-2010 🟩 2011-2014 🟦 2015-2016 🟧 2017-2018 ⬜

Figure 4.18 shows the frequency of variants of pESI identified in each year group. As seen with isolation source and origin, there is little change in the presence of Integron B by year group. Integron A does fluctuate in frequency by year with a slight downward trend visible since 2011. A significant increase in the presence of the *bla*$_{\text{CTX-M-65}}$ gene on pESI is visible, with a complete absence of the gene on the plasmid on strains isolated prior to 2006 and 48.2% (295/612) of the pESI positive isolates from 2017-2019 containing *bla*$_{\text{CTX-M-65}}$ on pESI ($p<0.00001$).

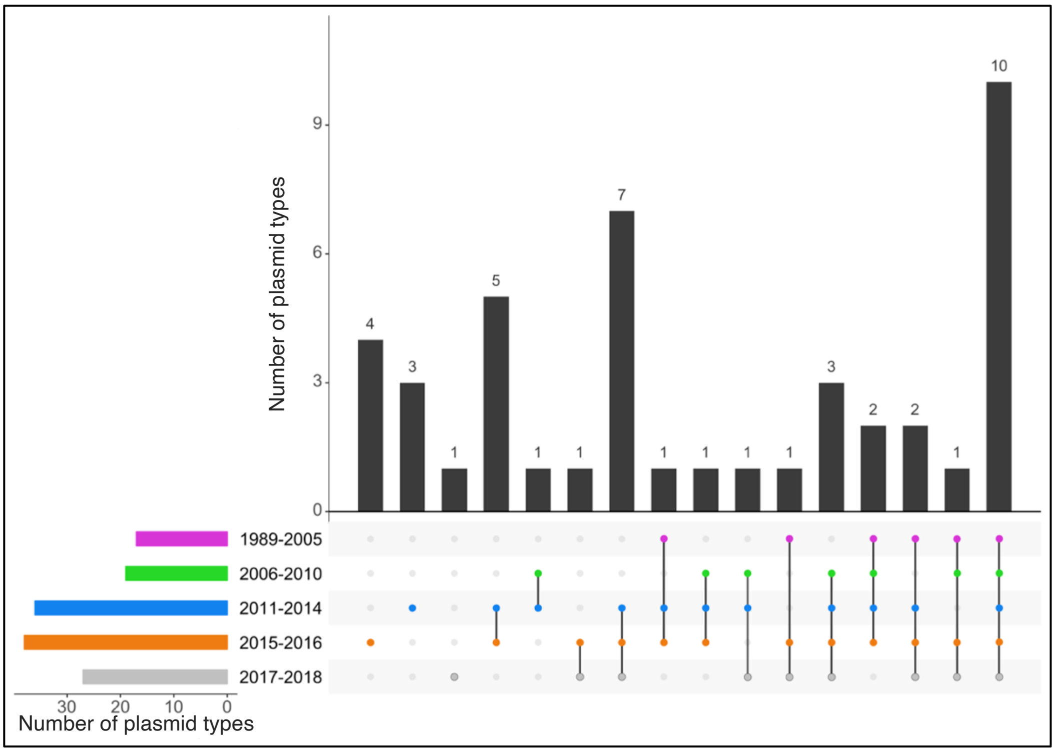


Figure 4.19 Distribution of plasmid types across eBG31 by year group
UpSet plot showing the number of plasmid types unique to and shared between eBG31 isolates from each year group.
1989-2005 🟪 2006-2010 🟩 2011-2014 🟦 2015-2016 🟧 2017-2018 ⬜

In the eBG297 isolates, 61.5% (8/13) of all the plasmids types were found exclusively in one year group. None of the plasmids were identified in every year group. The distribution of plasmid types in eBG31 across each year group is plotted in Figure 4.19. Unlike eBG297, a low proportion of plasmid types were found in a single year group (16.3%, 8/49). 69.4% (34/49) of the plasmid types were found in more than 2 year groups. Correlating with the decrease in diversity of AMR gene clusters between 2015-2016 and 2017-2018, fewer plasmid groups were present in the 2017-2018 isolates; this year group also had the lowest number of plasmids unique to it.

4.3.2.5 *Phylogenetic Structure of pESI*

To ascertain the genetic diversity amongst the pESI sequences a SnapperDB was made, using the eBG31 nanopore reference and the Israeli pESI reference as a combined reference. 1539 of the pESI positive isolates passed into the database. 1013 5SNP clusters were present in the pESI SnapperDB, the largest of which containing 97 isolates; 83.6% (847/1013) contained only a single isolate.

Representatives of each 5SNP cluster within the database were selected and a soft-core SNP phylogeny was made, masking prophages, recombination and the eBG31 reference (Figure 4.20). The total number of SNPs represented in the phylogeny is 555.

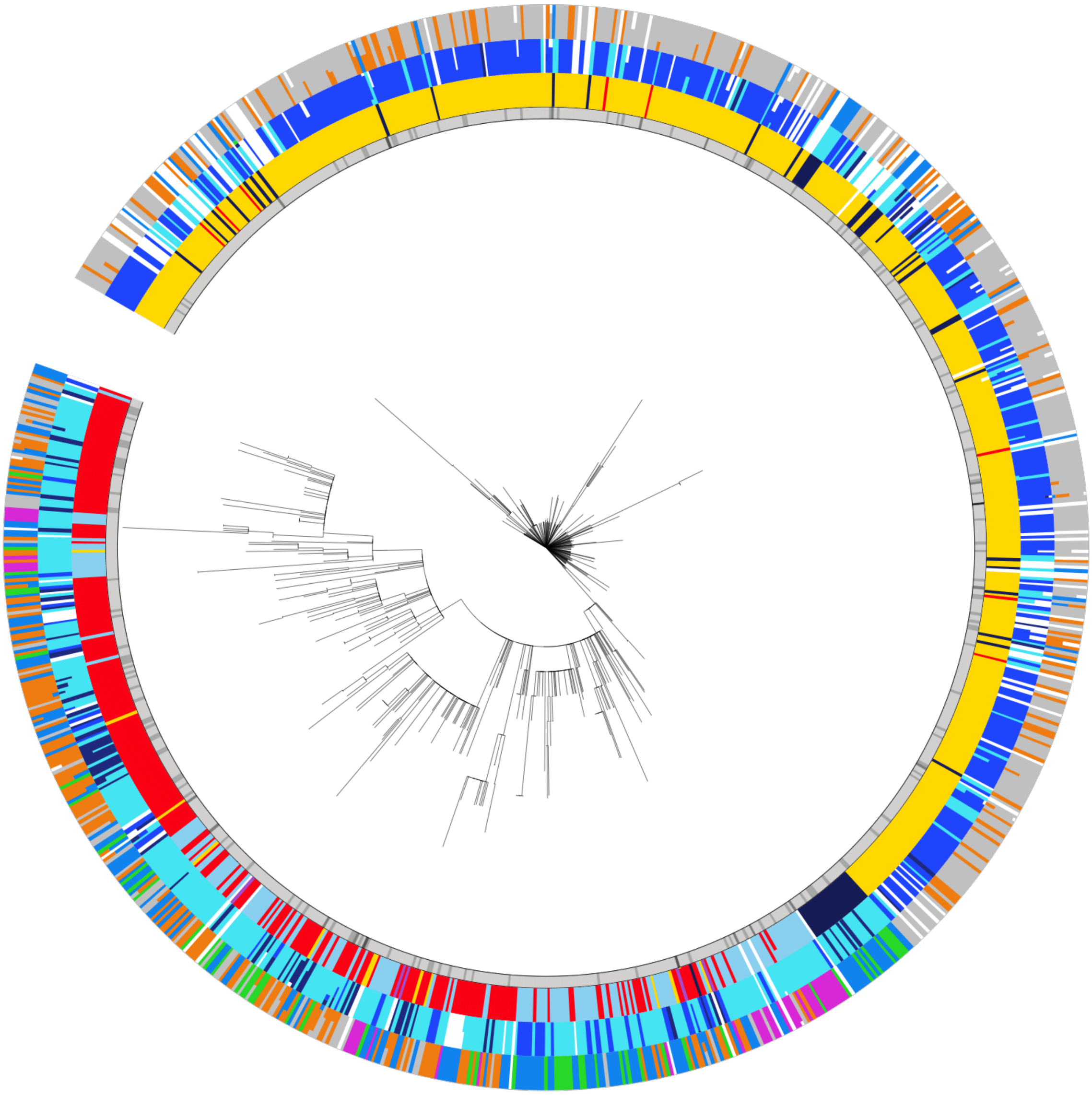


Figure 4.20 Phylogenetic structure of pESI
Soft-core SNP Maximum Likelihood Phylogeny of 1013 5SNP cluster representatives of pESI. The rings are annotated with the percentage of isolates in each 5SNP cluster that were isolated from each metadata group.

Inner ring, number of sequences in 5SNP cluster: ⬜ 1 ▨ 2-5 ▨ 6-20 ⬛ 21-50 ⬛ >50

Second ring, percentage of sequences in each cluster from each continent:
Africa 🟪 Asia 🟦 Europe 🟥 North America 🟨 South America ⬛ Unknown □

Third ring, percentage of sequences in each cluster from each isolation source:
Environmental ⬛ Human 🟦 Poultry 🟦 Unknown □

Outer ring, percentage of sequences in each cluster from each year group:
1989-2005 🟪 2006-2010 🟩 2011-2014 🟦 2015-2016 🟧 2017-2018 ▨ Unknown □

The phylogeny, which was rooted to its most ancestral node, contains several clades with small branch lengths close to the root. The majority of the sequences (440) belong to one of these clades.

The second ring in Figure 4.20 shows the percentage of isolates from each 5SNP cluster that were isolated from each continent. Clustering by origin of isolation is apparent, with clades comprised mainly of North American sequences visible as well as clades comprised mainly of European or Asian sequences. The geographical distribution is very similar to that seen in the eBG31 phylogeny (Figure 3.13).

The third ring in Figure 4.20 shows the percentage of isolates from each 5SNP cluster that were isolated from each isolation source. Clustering is also visible by source, with the clades that contain North American sequences comprised mainly of isolates from poultry. Conversely, the European and Asian clades contained large clusters of isolates from humans.

The outer ring of Figure 4.20 shows the percentage of isolates from each 5SNP cluster that were isolated in each year group. Some clustering is visible, the North American clades are mainly comprised of sequences from 2017-2018. However, little clustering by year group is seen in the remainder of the tree.

The pESI phylogeny was also annotated with pESI variant (Appendix IV Figure IV.3). Clustering of isolates lacking Integron A was observed with the most noticeable example present in one of the Eurasian clades, where most of the European sequences did not have the integron and the majority of the Asian sequences did. Clustering of pESI sequences with *bla*$_{\text{CTX-M-65}}$ was also visible, with the North American clades containing almost all of the occurrences of the gene on the plasmid.

To determine the distribution of pESI and pESI variants across the eBG31 soft-core SNP phylogeny (Figure 3.9), it was annotated with the percentage of isolates in each cluster that contained pESI (Appendix IV Figure IV.4). 89.9% (1384/1541) of the pESI positive sequences clustered in two of the fastbaps clusters, 1 and 5. One 25SNP representative group in fastbaps cluster 4 also contained pESI; this cluster was comprised of sequences from every continent and isolation source. Isolates containing Integron A were found in fastbaps clusters 4 and 5 but not in cluster 1. Despite the number of pESI plasmids carrying *bla*$_{\text{CTX-M-65}}$, only 12/831 25SNP representative groups contained these sequences, one of which containing 96.1% (567/590) of them. The pESI positive isolates with *bla*$_{\text{CTX-M-65}}$ in this group originated from Europe, North America, South America or had an unknown origin. They also came from all the source groups and were isolated from every year between 2010 and 2018.

4.3.2.6 *Integron Occurrence*

The number of integrons in each of the sequences was quantified to identify the variation in integron frequency across the *S.* Infantis population. Integron Finder was run on the scaffolded whole genome assemblies, including plasmids, of all *S.* Infantis isolates whose assemblies passed the assembly quality checks. This resulted in 4390 eBG31 isolates and 180 eBG297 sequences being included.

66.1% of the *S.* Infantis isolates did not contain any complete integrons, 18.9% contained 1 and 15% contained 2 integrons. All the integrons were found to contain the *int1* gene and therefore be Class 1 integrons. Figure 4.21 shows the percentage of eBG31 and eBG297 isolates that contained integrons.

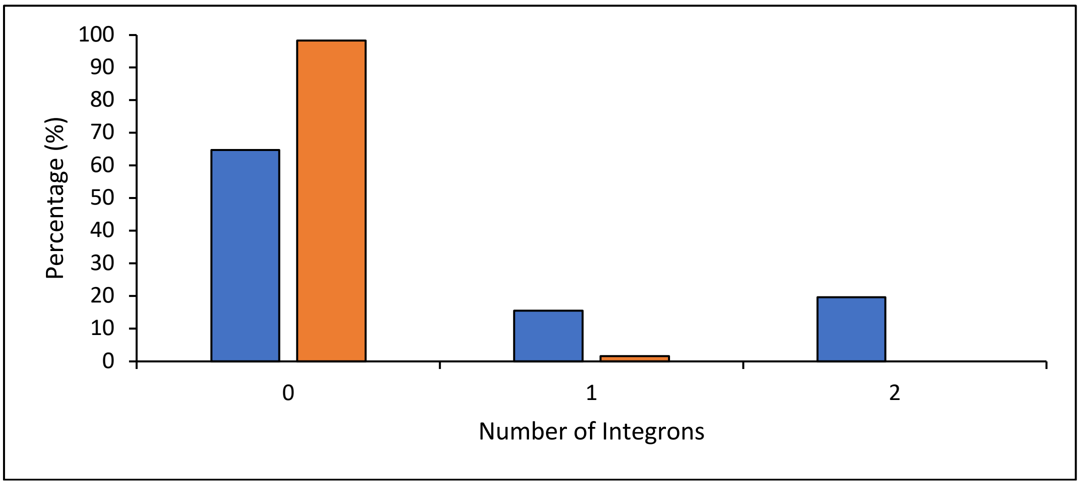


Figure 4.21 Difference in integron frequency between eBG31 and eBG297
eBG31 (n=4390), eBG297 (n=180).
eBG31 🟦 eBG297 🟧

A significant difference was seen between the percentage of isolates that had no integrons in eBG31 vs. eBG297 ($p<0.00001$). Only 1.7% (3/180) of eBG297 isolates had any integrons; conversely 35.2% (1544/4390) of eBG31 had integrons, with a slightly higher proportion of these containing 2 integrons rather than 1. Of the 1544 eBG31 isolates that had integrons, 92.9% (1435/1544) were positive for pESI. In contrast, of the 1,490 pESI positive isolates that were checked for integron presence, only 2.3% did not contain an integron.

The difference in proportion of isolates from each isolation source that had integrons is displayed in Figure 4.22. Unsurprisingly considering AMR and plasmids were found exclusively in eBG297 isolates from humans, integrons were also found only in these samples, present in 1.8% (3/164) of the human samples. The percentage of isolates containing different numbers of integrons varied by isolation source in eBG31. The source with the lowest percentage of isolates with integrons was environmental, with only 15.2% (142/931) containing any integrons. This increased significantly with 70.7% (668/945) of poultry isolates containing 1 or 2 integrons (p<0.00001). A significantly larger proportion of poultry eBG31 isolates contained integrons than human eBG31 isolates (p< 0.00001).

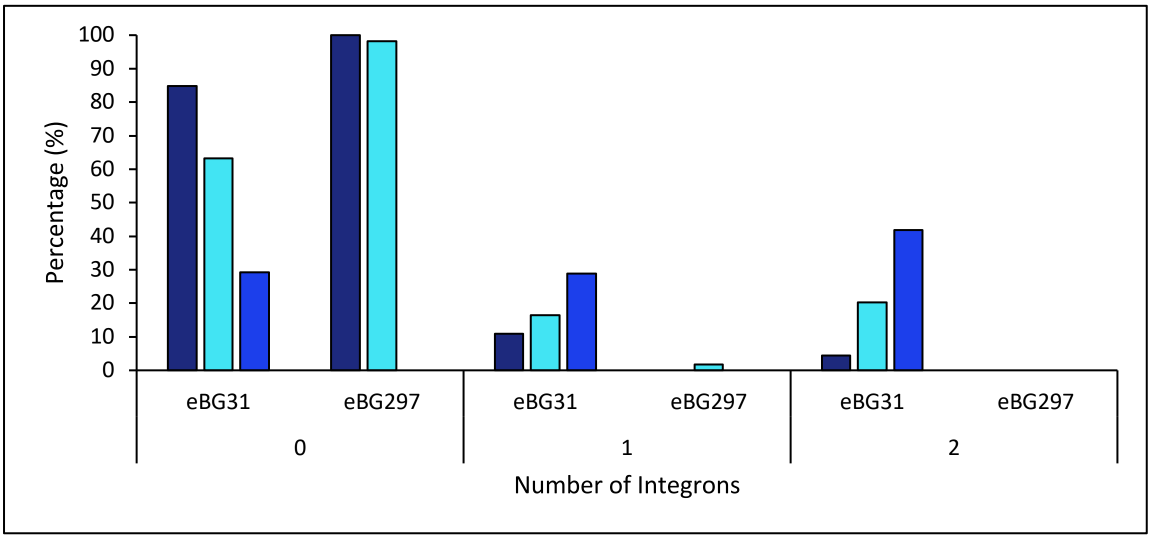


Figure 4.22 Difference in integron frequency between *S.* Infantis isolation sources
Shown as a percentage of isolates from each source.
eBG31: environmental (n=931); human (n=1482); and poultry (n=945)
eBG297: environmental (n=9); human (n=164) and poultry (n=0)
Environmental ■ Human ■ Poultry ■

The frequency of integron isolation from each continent is shown in Figure 4.23. The number of integrons in each isolate varied depending on which continent it was isolated from, ranging from 2.4% (7/291) of African eBG31 sequences containing integrons to 78.5% (95/121) of South American eBG31 sequences. In the eBG297 sequences, the only continent that contained isolates with any integrons was Africa, with just 2% (3/151) containing 1 integron.

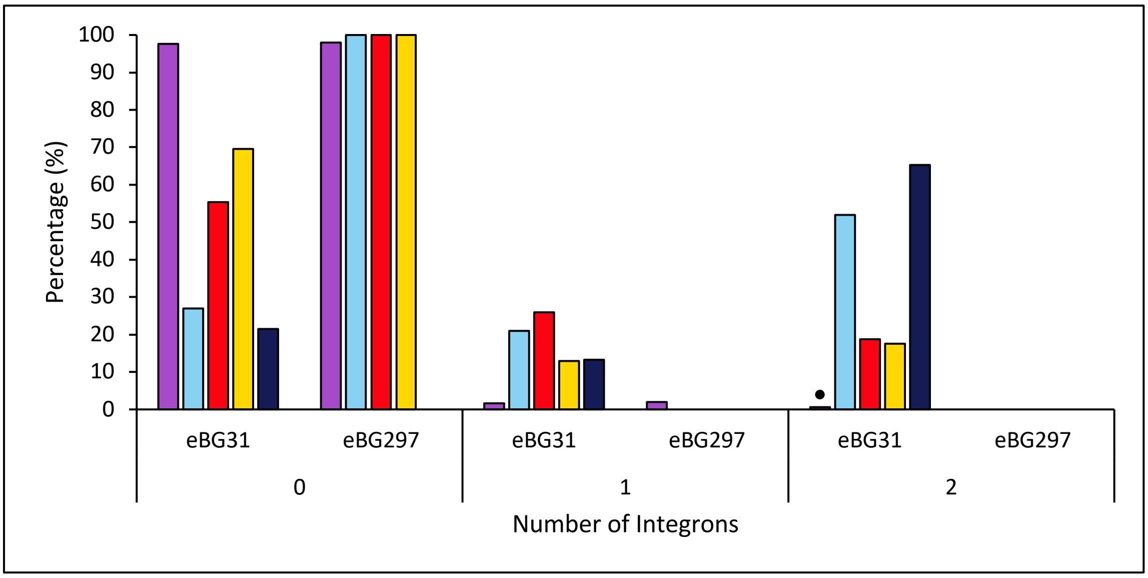


Figure 4.23 Difference in integron frequency across *S.* Infantis by continent
Shown as a percentage of isolates from each continent. ● values are less than 1% but greater than 0.
eBG31: Africa (n=291); Asia (n=233); Europe (n=900); North America (n=2774) and South America (n=121)
eBG297: Africa (n=151); Asia (n=3); Europe (n=20); North America (n=2) and South America (n=0)
Africa 🟪 Asia 🟦 Europe 🟥 North America 🟨 South America ⬛

The frequency of integron occurrence from each year group was also compared (Appendix IV Figure IV.1). Integrons were only seen in eBG297 strains isolated between 2011 and 2014, which correlates with the only year group to have MDR eBG297 isolates. The percentage of eBG31 isolates containing integrons has been increasing since 2006; 51.5% (596/1158) of isolates from 2017-2019 contained at least one integron.

4.4 Discussion

A significant difference was found in the occurrence of AMR between the two eBGs in *S.* Infantis. This cannot be attributed to temporal differences in when the strains were isolated, as the eBG31 strains were isolated between 1989-2018 and the eBG297 strains between 2003-2019. It could be due to the much larger size of the eBG31 population when compared to eBG297, or may be because of the differences in geographical distribution identified in Chapter 3.3.1.6; the majority (84%) of the eBG297 sequences were isolated from Africa compared to only 7% of eBG31 sequences.

Low levels of AMR were observed in isolates from both eBGs isolated from Africa. As the majority of the sequences representing this continent were from South Africa, this finding is surprising due to the proportion of immuno-compromised people in this country. With a high proportion of the South African population being positive for AIDS, the need for frequent antibiotic consumption is increased, which would be expected to be associated with increased levels of resistance to these antibiotics (Essack *et al.*, 2017; Statistics South Africa, 2018). However, this is not seen within the *S.* Infantis population.

There was a large amount of variation in the AMR gene clusters that were found in eBG31 isolates from each continent, with only 11 of the 62 genes being found in isolates from all of the continents. This indicates that movement of *S.* Infantis between continents, or at least of isolates with AMR, has not been frequently occurring. Furthermore, AMR genes were found in eBG31 isolates that were unique to each of the continents, suggesting that there are continent-specific AMR profiles within the eBG31 population.

As broiler chickens are most frequently associated with *S.* Infantis, chicken meat is predicted to be the biggest source of zoonotic transmission of *S.* Infantis to humans (EFSA and ECDC, 2019a). However, the percentage of MDR S. Infantis isolates from poultry was significantly higher than those isolated from humans. This could indicate that poultry is not such a large contributor to *S.* Infantis in humans; as human and environmental eBG31 isolates shared the most AMR gene clusters, possibly sources from this group are contributing more to human infection. Alternatively, it is plausible that a subgroup of *S.* Infantis from poultry with lower rates of MDR is the source of the human infection.

The high levels of AMR in eBG31 poultry isolates could explain the difference in AMR between the eBG31 and eBG297 isolates. Isolates from poultry are a large contributor to AMR in the eBG31 population; it is conceivable that the lack of any eBG297 isolates from poultry sources explains the low levels of AMR seen in this eBG.

As seen with other *Salmonella* serovars (FDA, 2019b), the frequency of identification of AMR in eBG31 is increasing, with the highest percentage of isolates containing resistance genes for 7 of the 11 antibiotic classes identified in 2017-2019. This concurs with other research that has shown that AMR is an increasing problem within *S.* Infantis (Hindermann *et al.*, 2017; Szmolka *et al.*, 2018). None of the identified AMR genes were unique to 2017-2019, suggesting that this increase in resistance is due to resistance profiles already present in the population rather than the introduction of new resistance profiles. Over a quarter of the AMR gene clusters identified were present in eBG31 samples isolated from every year group; genes encoding to resistance to aminoglycosides, chloramphenicols, sulphonamides, tetracyclines, trimethoprims and beta-lactams, including extended-spectrum beta-lactams, are therefore being maintained within the *S.* Infantis population. The eBG297 population had consistently low levels of antibiotic resistance, with MDR only present in isolates between July and October 2012, and therefore plausibly part of an outbreak.

The small number of shared plasmids across eBG31 suggests that, other than between specific places such as North America and Europe, there is not a large amount of transmission of *S.* Infantis between continents. This supports the finding of a geographical signal across the eBG31 population in Chapter 3.3.2.5. Whilst in the eBG31 phylogeny the North American and European sequences were found to cluster separately, the European cluster did contain a large group of North American sequences, potentially explaining why these continents in particular share several plasmids.

Concerningly the number of *S.* Infantis isolates with plasmids was at its highest between 2017 and 2019; associated with the highest levels of AMR. Whilst several different groups of plasmids were maintained in the eBG31 population, the same was not seen for eBG297 isolates where those that were seen in multiple year groups, occurred at low frequencies. Significantly fewer integrons were also identified in the eBG297 isolates than in eBG31. This suggests that the presence of these mobile genetic elements was not advantageous enough to be maintained; or that due to the small size of the eBG297 population, transmission was lower than in eBG31.

The global presence of pESI in *S.* Infantis was 33%, therefore it would be expected that pESI would have been present in eBG297, this was not the case. This could indicate that pESI is found specifically in eBG31 isolates, supporting the findings of Chapter 3.3.6.3

that eBG31 and eBG297 are distinct from one another and eBG297 should not be classified as *S.* Infantis. However, this absence of pESI in eBG297 could also be attributed to geography. Only 1.7% of the eBG31 African isolates were positive for pESI and as 84.2% of eBG297 sequences were found in Africa this could explain the lack of the plasmid in eBG297.

pESI, previously found to be associated with broilers (Hindermann *et al.*, 2017), was found more frequently in eBG31 isolates from poultry sources than any other source, with 69.7% of the poultry isolates in the project being positive for the plasmid. This suggests pESI is spreading throughout the poultry industry. Whilst at lower levels than seen in poultry isolates, pESI was also present in 36.3% of eBG31 isolates from humans. Due to the presence of several AMR genes on the plasmid this is a public health concern, particularly if the plasmid becomes more prevalent in areas with a larger proportion of immuno-compromised people, such as Africa. Interestingly, pESI was only identified in 13% of the eBG31 environmental isolates. This suggests that whilst pESI is present in livestock and other environmental sources, it is not common in this source of *S.* Infantis. The decreased frequency of Integron A in the pESI positive environmental eBG31 isolates indicates that adaptation to niche has occurred and that the environmental isolates are distinct from the human and poultry isolates.

Also of concern is the presence of the *bla*$_{\text{CTX-M-65}}$ gene on pESI amongst the poultry eBG31 isolates, which increased significantly in occurrence and was found to be present on 48.6% of pESI from poultry. As *S.* Infantis is found at high levels in poultry meat from EU member states, this could result in an increase in human cases with resistance to extended-spectrum beta-lactams (EFSA and ECDC, 2019a). The fact that no plasmid groups were found exclusively in poultry isolates suggests that this could be the source of infection for the other source groups; any plasmids these isolates acquire have been transmitted to the other sources. This further increases the risk of the *bla*$_{\text{CTX-M-65}}$ gene spreading across the *S.* Infantis population.

This project has identified pESI in *S.* Infantis from Japan in 1999; pESI has therefore been maintained in the *S.* Infantis population for the last 2 decades. Whilst it is plausible that pESI originated in Asia, more isolates from that time period would be needed to support this. The presence of Integron A on pESI varies by continent, source and year of isolation. This could indicate that in some continents more than one population of *S.* Infantis are coexisting, carrying different variants of pESI. It could also be due to the selective pressure for the integron, perhaps the need for resistance to trimethoprim, varies with continent, source or time. There was no association with Integron A presence and isolation from urine, however data on trimethoprim use was not available; with this data the cause of the variation in the presence of Integron A may become more apparent.

This project has identified, for the first time, the severity of the pESI problem within the *S.* Infantis population. Previous research has identified the presence of pESI in isolates from select countries, but the overall numbers of isolates with pESI was before now unknown. pESI was found to be associated with a large proportion of the AMR in the *S.* Infantis population. Whilst it has not been determined that pESI carried these AMR determinants, there is a clear increase in particular in AMR and integrons in isolates with the plasmid.

4.4.1 Conclusions

To conclude this work has shown, for the first time, the global and source distribution of AMR and pESI in *S.* Infantis. The levels of AMR vary drastically in the *S.* Infantis population, with high levels seen in eBG31 and low levels in eBG297. Of particular concern is the high levels of resistance seen in isolates sourced from poultry. pESI is a large contributor to the high levels of AMR and therefore the identification of its presence in samples should be made a high priority by public health laboratories. As high-quality genomes are required for accurate determining of AMR gene and pESI presence, it is possible that low coverage in areas of the genome containing these elements could result in an underestimation. Therefore, the true levels of AMR and pESI seen in *S.* Infantis may be even higher.

5. Chapter 5. Genome-Wide Association Study of *S.* Infantis Isolated from Humans and Poultry

5.1 Introduction

To reduce cases of salmonellosis, informing prevention measures with source attribution is needed (Pires *et al.*, 2014). Genome wide association studies (GWAS) can be used to identify if genetic differences between isolates are associated with a phenotype. For example, when comparing 440 *S. enterica* isolates from 15 serovars and different isolation sources, researchers were able to identify genes significantly associated with avian, bovine, swine and fish sources (Vila Nova *et al.*, 2019). The absence of genes was not found to be associated with isolation source.

The sizes of the pan and core genomes of *Salmonella* varies between studies due to the inclusion of different serovars or the numbers of genomes being compared. An assessment of 4,893 *S. enterica* isolates, including strains from each subspecies, identified a pan genome estimated at 25,300 genes, with 1,500 belonging to the core genome (Laing, Whiteside and Gannon, 2017). A smaller study, including 440 *S. enterica* isolates, found the pan genome consisted of 21,835 genes, with 2,705 being core (Vila Nova *et al.*, 2019).

While several studies have been published looking at the pan genome structure of multiple *Salmonella* serovars, fewer have looked within serovars. A comparison of 37 *S.* Typhimurium isolates had a pan genome of 6,433 genes, with 4,003 core; 21 *S.* Dublin isolates had a pan genome of 5,066 genes, with 4,326 core and 32 *S.* Newport isolates had a pan genome of 5,351 genes, with 4,290 core (Liao *et al.*, 2019). Another study looking at 115 strains of *S.* Weltevreden found a pan genome of 11,969 coding sequences and a core of 4,046 (Makendi *et al.*, 2016).

Many virulence factors have been identified in *Salmonella*, with virulence genes varying between and within serovars (Cheng *et al.*, 2015). IGRs have been identified as virulence factors in *Salmonella*; however, little research has been performed looking at the number of pan and core IGRs in *Salmonella*. When evaluating 68 *S.* Typhimurium strains, the total number of IGRs identified was 1857; the majority (1576) of these were core IGRs (Fu *et al.*, 2015).

Previous research has found that *S.* Infantis is less invasive than *S.* Typhimurium and that the virulence factors present vary depending on country of isolation; with *sopE* absent in *S.* Infantis from Israel but the most common virulence factor identified in

Pakistan (Aviv *et al.*, 2019; Wajid *et al.*, 2019). Differences in virulence gene levels have also been seen between *S.* Infantis isolates from different sources; virulence gene patterns were observed in isolates from broilers but missing in isolates from breeders and layers (Sever and Akan, 2019).

5.1.1 Aims and Objectives

S. Infantis is an important serovar in poultry, being the serovar most frequently identified from broilers and broiler meat in EU member states (EFSA and ECDC, 2019a). However, despite being seen at such high levels in broilers, it is seen relatively less frequently in humans, being the fourth most common serovar in EU member states. Identifying why this is the case is important for public health as it could be used to predict and reduce the risk of the numbers of human cases increasing.

Previous chapters in this thesis have identified that, on the eBG31 phylogeny, several clusters of human sequences were visible, indicating the potential for a sub-group of S. Infantis that is more virulent in humans. The human eBG31 sequences were also found to be more diverse than poultry isolates. Conversely, the poultry eBG31 isolates had significantly higher levels of AMR than the isolates from humans ($p < 0.00001$); furthermore, pESI was significantly more commonly identified in isolates from poultry than humans ($p < 0.00001$).

Little is known about the pan genome or IGRs in *S.* Infantis and if they vary between isolates from different sources. As it is possible that genetic differences in *S.* Infantis isolates could explain the differences in the numbers of cases seen in poultry and humans, the aims of this chapter were to:

- Identify the variation in genes between *S.* Infantis isolated from humans and poultry
- Identify the variation in IGRs between *S.* Infantis isolated from humans and poultry
- Determine associations with other components of the genome and isolation source

5.2 Specific Methods for Feature Identification and Comparison

5.2.1 Strain Selection

The eBG31 collection was analysed for this chapter, with the Illumina short read sequence data for the eBG31 reference strain being used (Ashton *et al.*, 2017). The strains were filtered by isolation source to include strains isolated from humans or chickens; the metadata phrases that were used to include isolates as from humans or poultry are listed in Appendix II Table II.1.

5.2.2 Virulence Factors

ARIBA was run using the parameters described in Chapter 2.7.4 and the *vfdb_full* database (Chen *et al.*, 2016; Hunt *et al.*, 2017). *ARIBA* summary was run with the cluster_cols match option to output the match column.

5.2.3 Genes and Intergenic Regions

The pan genome and IGR analyses were both carried out using scaffolded assemblies that were of sufficient quality (Chapter 2.6.7). *Roary* (version 3.12.0) was run on the gff outputs of *Prokka* (Chapter 2.6.8) using 16 threads, making a fast core gene alignment with -e and -n and not splitting paralogs with -s (a requirement for *Piggy* input) (Seemann, 2014; Page *et al.*, 2015). *Artemis* (version 17.01.1) was used with the gff output of *Prokka* to identify the amino acids present in hypothetical proteins of interest (Rutherford *et al.*, 2000). This was then run through *Protein BLAST* with default settings to identify homologous proteins (McGinnis and Madden, 2004).

Piggy (accessed 05.06.19) was run with 16 threads using the gff outputs of *Prokka* (Chapter 2.6.8) and the output folder produced by *Roary* with default parameters (Thorpe *et al.*, 2018). The core gene and core IGR alignments produced by *Roary* and *Piggy* respectively were used as input for *RAxML*, which was used with the settings described in Chapter 2.7.2 to generate core gene and core IGR phylogenies. The phylogenies were rooted to their most ancestral node and visualised and annotated using *iToL* and the *iToL* colored strip file (Letunic and Bork, 2016).

TreeBreaker (accessed 10.06.19) was run on the core gene and core IGR phylogenies with default settings (Ansari and Didelot, 2016; Azim and Didelot, 2018). The

TreeBreaker R script (accessed 22.10.19) was then used with *R* (version 3.5.1) and *RStudio* (version 1.2.1335) to generate the phylogenies annotated with posterior probability; I amended the script to alter the display of the phylogenies to circular and reduce the size of the tip labels (R Core Team, 2018; RStudio, 2018).

5.2.4 Statistical Analyses

Scoary (version 1.6.16) was used with the *ARIBA*, *Roary* and *Piggy* results on *CLIMB* to identify virulence factors, genes and IGRs that were associated with isolation source (Chapter 2 2.7.5) (Brynildsrud *et al.*, 2016; Connor *et al.*, 2016). Unlike in Chapter 4, the pairwise comparisons option, which is default, was used for all the analyses, enabling correction based on population structure. The option --p 1.0 was also used to not exclude by p-value. A Benjamini-Hochberg adjusted p-value of less than 0.05 was used to define significance, correcting for multiple comparisons.

The p-values within box plots were determined using the Mann Whitney U test using *RStudio* and *R*. The two population proportions z test calculator was used to determine whether proportions were significantly different (Stangroom, 2019).

Figures 5.5 and 5.9 show the distribution of the differences in percentage between poultry and human associated genes/IGRs. To create these figures the percentage of isolates with each gene/IGR was calculated for both sources. The smaller percentage was then subtracted from the larger percentage; for example, if a gene was present in 80% of poultry isolates and 20% of human isolates, it would be poultry associated, with a difference in percentage of 60%. The frequency of these differences in percentage was then plotted.

5.2.5 PySeer – Unitig Association with Source

For this analysis, the contigs.fasta outputs of *SPAdes* (Chapter 2.6.6) was used as input. The human and chicken *S.* Infantis isolate assemblies that were good quality (Chapter 2.6.7) were copied to *CLIMB* where all of the software was run (Bankevich *et al.*, 2012). *Unitig-counter* (version 1.0.5) was run on all the assemblies using four cores; this involves the generation of a De-Bruijn graph of all the assemblies, nodes on the graph are shared by genomes and are called unitigs (Jaillard *et al.*, 2018; Lees, 2019). *MASH* (version 2.1.1) was used to generate a distance matrix for all the assemblies with 4 threads and a sketch size of 10,000, with the square_mash script provided with *PySeer* (version 1.1.2) used to convert the output (Ondov *et al.*, 2016; Lees *et al.*, 2018).

The scree_plot_pyseer script that was provided with *PySeer* was used to generate a scree plot of the *MASH* output to determine the *PySeer* setting for multidimensional scaling; this shows the results of a principal component analysis where the data was split into components and the variance in each measured in eigenvalues. The 'knee' of the plot (Figure 5.1) was recommended as the value to be used, so 2 was selected for this option (Lees and Galardini, 2018). *Pyseer* was then run with the uncompressed option, using 2 as the value for max-dimensions, the output of *unitig-counter* as k-mers and the distance matrix generated by *MASH* to control for population structure. The names of the isolates containing each unitig was added using the print-samples flag and the output_patterns option was used to create a patterns file to determine the significance threshold.

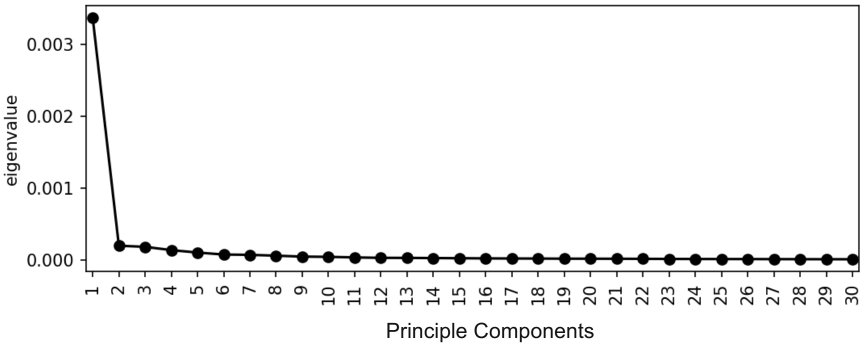


Figure 5.1 Scree plot of the variance across the distance matrix.

The variance across the first 30 principle components in the distance matrix calculated using eBG31 isolates from humans (n=1432) and poultry (n=945).

The *Roary* core gene phylogeny was also used with *PySeer* instead of the *MASH* output to adjust for population structure. The qq_plot script included with *PySeer* was used to produce plots comparing the p-values. Using *MASH* to control for population structure gave a better result and was therefore selected over the use of the phylogeny.

The count_patterns.py script provided by *PySeer* was used with the patterns file, an alpha value of 0.05 and 4 cores to determine the Bonferroni p-value threshold for defining significance, corrected for multiple comparisons, which was 1.02x10^{-6}. The lrt p-value, the p-value adjusted for population structure, was then used to identify the significant unitigs. Sed commands were used to replace each isolates name with either 'Human' or 'Poultry'; the number of occurrences of each of these for each unitig was then counted.

The phandango_mapper command with *PySeer* was used to create a Manhattan plot of the significant unitigs associated with either isolation source, mapped against the eBG31 nanopore reference. This was visualised using *phandango* and the *Prokka* annotation file of the reference (Chapter 2.6.8) (Hadfield *et al.*, 2018). The coordinates of areas of unitig coverage of the reference were used to identify which coding sequences were present in the *Prokka* file.

5.3 Results

To further investigate whether there were any genetic differences between *S.* Infantis isolates that infect poultry and humans, the variation in the virulence factors, pan genome, IGRs, SNPs and unitigs between isolates from these source groups were compared.

As no eBG297 strains were isolated from poultry, the eBG297 human isolates were excluded from this chapter's analyses; as eBG31 represented 96.1% of the *S.* Infantis collection this is a good representation of the population.

5.3.1 Virulence Factors

The virulence factors in all eBG31 human (n=1,519) and poultry (n=947) isolates were identified using *ARIBA* and the full *VFDB* database. 698 virulence factors were identified, 36 of which were conserved across all of the isolates; the remaining 662 virulence factors varied in frequency of identification between the two sources. The previously mentioned genes *sopE* and *sopE2* were identified in 0.2% (6/2466) and 99.4% (2451/2466) of the isolates respectively.

Scoary was used to determine whether any of the virulence factors were significantly associated with either source group, with a Benjamini-Hochberg corrected p-value of less than 0.05. 30 were identified as being significantly associated to a source group, 21 of these were associated with poultry isolates, although none were found exclusively from that source group (Appendix VI.5). The factor most associated with poultry isolates was *irp1*, the yersiniabactin biosynthetic protein, which was present in 67.5% (639/947) of the poultry isolates and 33.0% (502/1519) of the human isolates.

9 of the virulence factors were significantly associated with being isolated from humans, 7 of which were identified exclusively from humans. However, these 7 factors were only present in 17 of the human isolates and were variants of the *pil* gene. Another *pil* gene, *pilT* was significantly associated with the human eBG31 isolates although it was only present in 25 human and 3 poultry isolates. The only other gene that was significantly more frequently identified in the human isolates was *sspH2;* present in 11.9% (181/1519) of the human isolates versus 7.1% (67/947) of the poultry isolates.

5.3.2 Variation in Pan Genome Structure

In order to determine whether the pan genome of human eBG31 isolates and poultry eBG31 isolates varied, 2427 good quality scaffolded assemblies, 1482 from humans and 945 from poultry, were annotated using *Prokka* and the genes present and shared across the isolates identified using *Roary*.

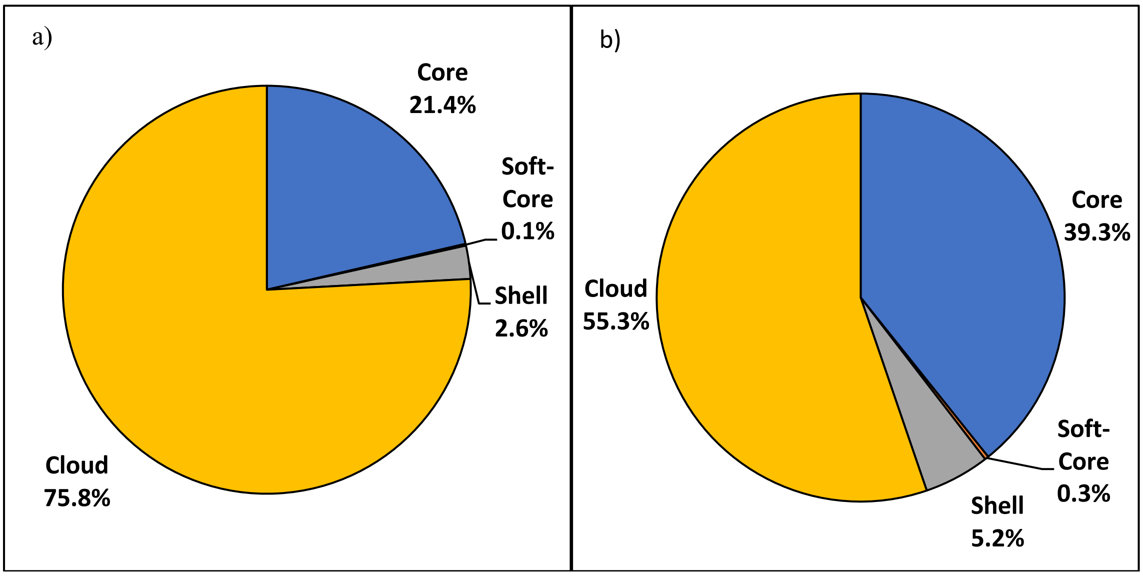


Figure 5.2 Pan genome composition of eBG31 human and eBG31 poultry isolates

a) Pan genome structure of eBG31 human isolates (n=1482)
b) Pan genome structure of eBG31 poultry isolates (n=945)

Core: 99% ⩽ Strains ⩽100%, Soft-core: 95% ⩽ Strains < 99%, Shell: 15% ⩽ Strains < 95%, Cloud: Strains < 15%

A total of 21,976 genes were identified across the 2,427 isolates, with the lowest number of genes present in a single isolate being 4,173 and the highest 5,065. The core genome, defined as gene presence in greater than or equal to 99% of strains, consisted of 4,132 genes. The accessory genome, all genes that were not core, was split into: soft-core which encompassed genes that were present in between 95% and 99% of strains; shell which was genes present in between 15% and 95% of strains and cloud which was genes seen in less than 15% of isolates. 0.2% (34/21976) of the genes identified were soft-core; 2.6% (561/21976) were shell and 78.5% (17248/21976) were cloud genes.

The pan genome structure within each isolation source was compared (Figure 5.2). Although the structure by source appears to vary considerably, a similar number of genes were core; 4,137 genes in the human eBG31 isolates were core, versus 4,141 in the poultry isolates. A larger number of genes, 19,343, were identified in eBG31 isolates from humans when compared to the number seen in poultry isolates, 10,550. The 83.3%

increase in the number of genes in the human isolates is due to the cloud genes, which increased in number by 151.7% from that seen in the poultry isolates.

This difference in the number of genes identified cannot be attributed to isolates from humans having more genes as the opposite was found to be true; a significant difference was found between the distribution of the number of genes in eBG31 isolates from either source, with over 50% of the poultry isolates having more genes than seen in 75% of the human isolates (p-value $< 2.2\text{x}10^{-16}$) (Figure 5.3).

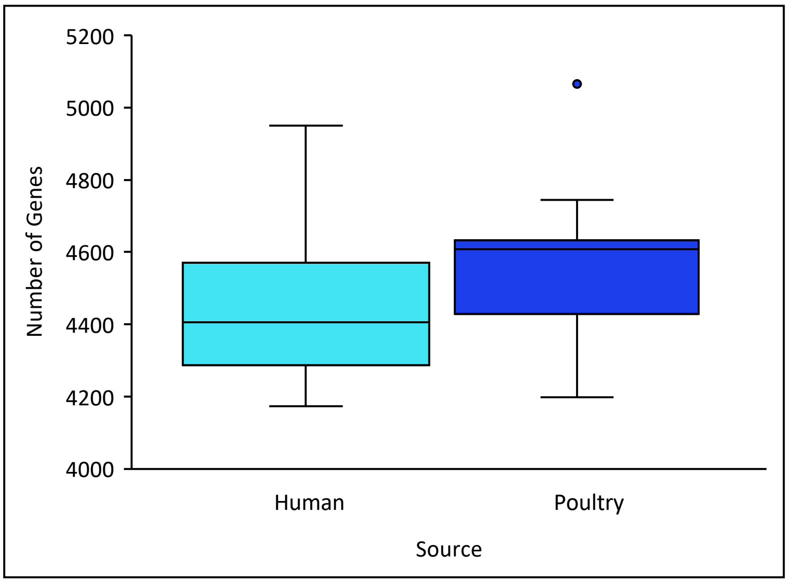


Figure 5.3 Difference in the distribution in the number of genes in eBG31 human and poultry isolates
Box plot showing the variation in the number of genes in each eBG31 isolate from human sources (n=1,482) and poultry sources (n=945).
Human 🟦 Poultry 🟦

To identify whether differences seen between the pan genome structure of eBG31 human and poultry isolates was due to a small number of extremely divergent sequences or that the discrepancy was seen uniformly throughout the groups, the number of genes in each of the isolates that belonged to each pan genome section was calculated (Figure 5.4).

A significant difference was seen between the number of core genes in each human and poultry isolated eBG31 isolate (p-value $< 2.2\text{x}10^{-16}$). The number of core genes was higher in the poultry isolates with 97.4% (920/945) of the poultry isolates containing more core genes than the largest number from a human isolate. A larger range was seen in human eBG31 isolates, with the lowest number of core genes seen in two historical PHE isolates at 4,044; both isolates were associated with foreign travel to Asia, one isolated in 2007 and the other in 2011.

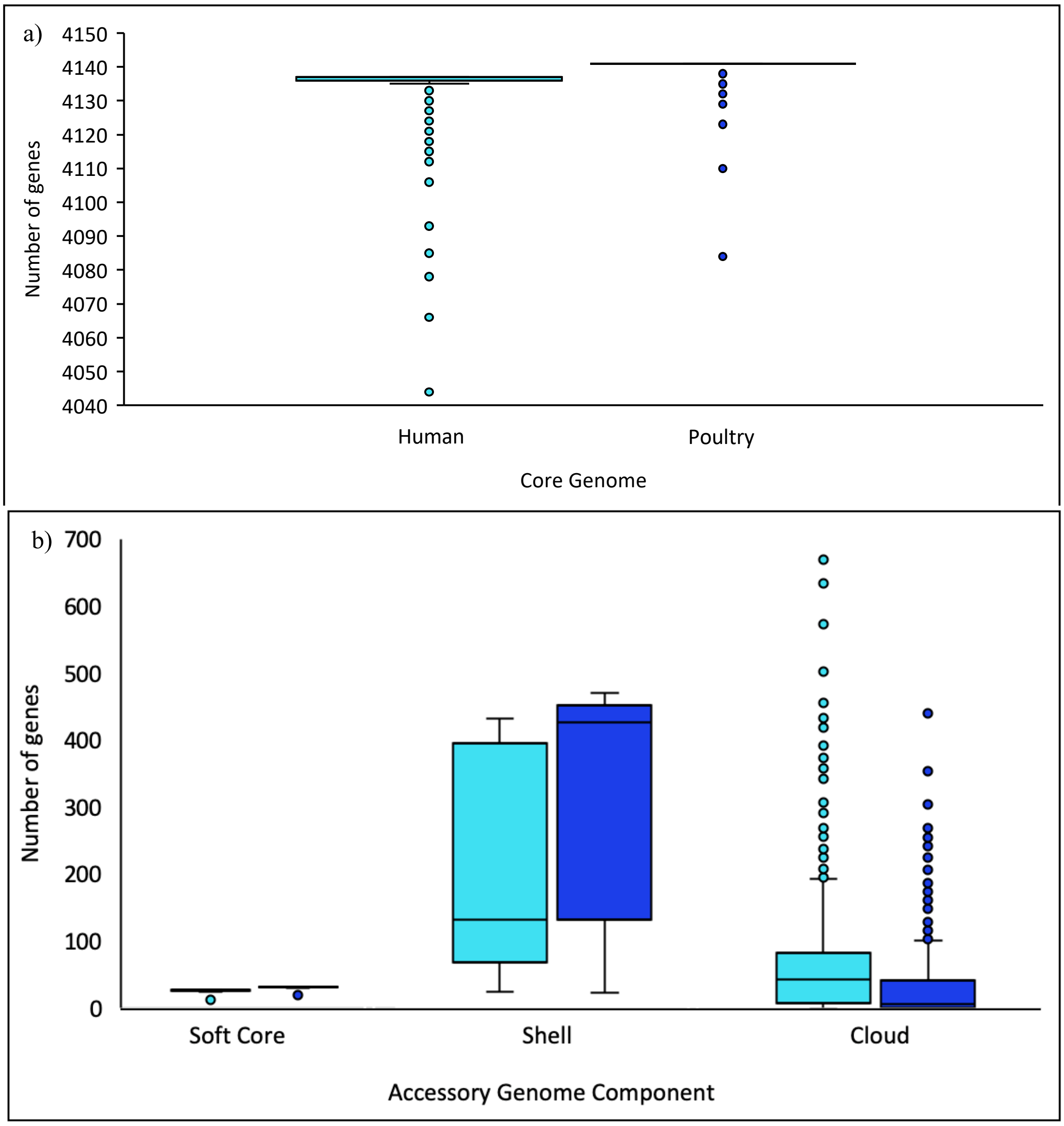


Figure 5.4 Distribution of the number of genes belonging to each pan genome component from each eBG31 human and poultry isolate

a) Number of core genes in each human (n=1482) and poultry (n=945) eBG31 isolate

b) Number of soft-core, shell and cloud genes in each human (n=1482) and poultry (n=945) eBG31 isolate

Human □ Poultry ■

A significant difference was also seen between human and poultry eBG31 isolates from each of the accessory genome components (p-value < 2.2×10^{-16}). A significantly greater number of both soft-core genes and shell genes were seen in poultry isolates (p-value < 2.2×10^{-16}). Conversely, significantly more cloud genes were identified in the human isolates; whilst isolates with 0 cloud genes were present, 50% had greater than or equal to 44 cloud genes (p-value < 2.2×10^{-16}). 99 human isolates had more than 194 cloud genes and are plotted as outliers on the box plot; these were isolated from Africa, Europe, North America and South America from every year between 2005-2018.

5.3.3 Genes Associated with Isolation Source

To establish whether any genes were significantly associated with isolation from human or poultry sources, *Scoary* was run on the *Roary* output. 1,342 genes were significantly associated with isolation source; 716 with human isolation and 626 with poultry isolation (p-value < 0.05). The difference in the percentage of isolates with associated genes from either source varied (Figure 5.5).

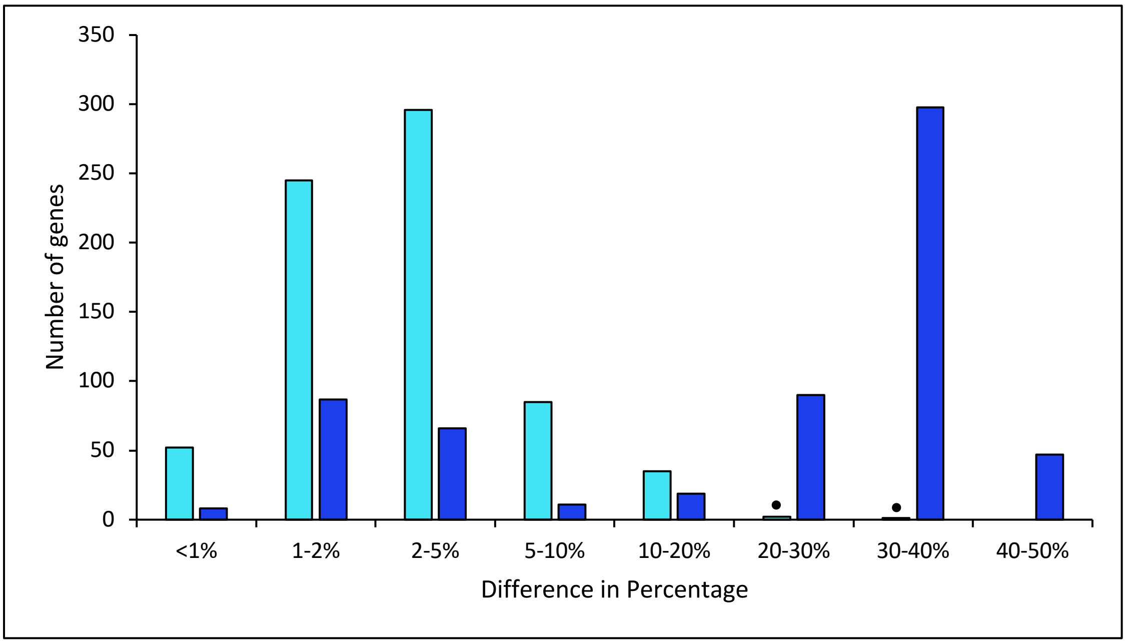


Figure 5.5 Distribution in the difference in the percentage of associated genes between sources
The percentage of isolates with a lower proportion of each significantly associated gene was subtracted from the percentage of isolates with each gene from the other source ● values are less than 5 but greater than 0.
Human (n=1482) □ Poultry (n=945) ■

Whilst more genes were significantly associated with human sources, many were identified in low numbers of human sequences; the majority of these genes (82.8%, 593/716) were only present in up to 5% more of the human isolates than in poultry isolates. Conversely, the majority (55.1%, 345/626) of the poultry associated genes were identified in over 30% more of the poultry isolates than in human isolates.

Focussing on genes that were identified in large numbers of sequences, 47 poultry associated genes were identified with a difference in percentage of 40% to that seen in human isolates, 40 of which were hypothetical proteins. The other seven genes were: *parM*, *srp54*, *pifC*, *vapC_2*, *xerC_2*, *pndA* and *tufA_1*. The gene most associated to poultry eBG31 isolates, *parM*, which is plasmid associated, was present in 56.9% (538/945) of isolates from poultry and 10.3% (152/1482) of isolates from humans. Another variant of the elongation factor gene, *tufA_2* was the gene most associated with eBG31 human isolation, present in 86.9% (1288/1482) of human eBG31 isolates and 50.2% (474/945) of eBG31 poultry isolates.

The human and poultry isolates in the eBG31 soft-core SNP phylogeny (Figure 3.9) were annotated with the *tufA* variant they contained (Appendix V Figure V.1). Clustering is visible, with several large groups of human eBG31 sequences containing exclusively *tufA*_2. In areas where the human eBG31 sequences clustered with poultry isolates there is more disparity in the variant of *tufA*, with both human and poultry isolates containing either variant or both.

5.3.4 Genes Unique to an Isolation Source

Many genes were also identified as being found exclusively in isolates from either source group. 11,426 of the genes were unique to the human isolates, 342 of which were significantly associated with that source ($p < 0.05$). A smaller number, 2632, were identified exclusively in eBG31 poultry isolates, 23 of these were significantly associated ($p < 0.05$). The distribution of the number of unique genes in each of the isolates from either source is plotted in Figure 5.6.

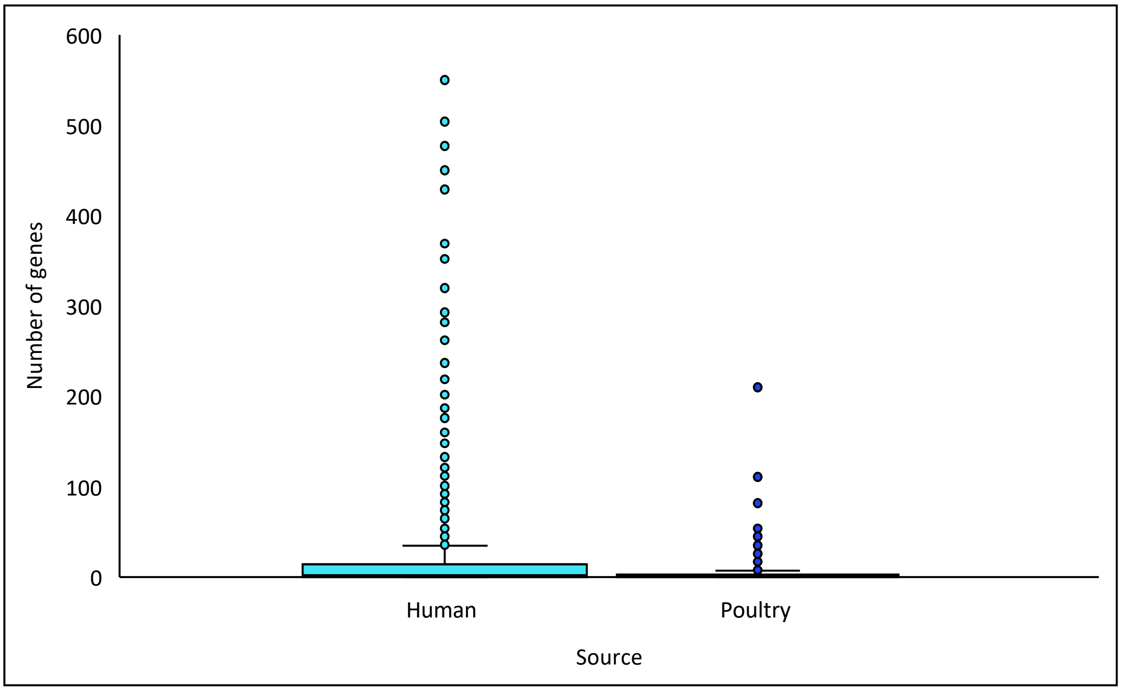


Figure 5.6 Distribution of the number of unique genes in eBG31 human and eBG31 poultry isolates
Box plot showing the variation in the number of genes found exclusively in each source group in each eBG31 isolate from human sources (n=1482) and poultry sources (n=945).
Human 🟦 Poultry 🟦

eBG31 isolates from human sources had a significantly different distribution of unique gene identification per isolate compared to the eBG31 poultry isolates (p-value < $2.2x10^{-16}$). Whilst the median number of unique genes per isolate from each source was comparable, the range varied considerably. Furthermore, significantly more human isolates than poultry isolates contained at least one unique gene (1066 human isolates, 551 poultry isolates, p-value < 0.00001).

Hypothetical proteins were encoded by 78.7% (8988/11426) and 77.5% (2041/2632) of the genes unique to human and poultry sources respectively. In both sources the majority of these were found in only one isolate; the largest number of isolates that shared a hypothetical protein was 47, which was observed twice in human isolates.

The distribution of the number of isolates from either source that contained unique genes encoding known proteins was calculated (Appendix V Figure V.2). Whilst significantly more of the human eBG31 isolates had unique genes, 68.9% (1682/2440) of the genes that encoded a defined protein were only identified in a single isolate. 2.7% (66/2440) of the genes were identified in more than 10 isolates, with the highest occurrence being variants of the *dam*, *ftsH*, *hin* genes which were seen in 34 human eBG31 isolates.

As with the human eBG31 unique genes, the majority (90.9%, 539/593) of the defined protein-coding genes unique to the eBG31 poultry isolates were only identified in 1 isolate. Unlike the human sourced isolates, only 2 of the genes were identified in over 10 samples, *ycbX_2* and a gene encoding the IS3 family transposase IS1353.

5.3.5 Variation in Intergenic Region Composition

To establish whether the IGRs of *S.* Infantis isolates varied between isolates sourced from humans and from poultry, *Piggy* was run on the gff output of *Prokka* and the *Roary* output. 1482 eBG31 isolates from humans and 945 eBG31 isolates from poultry sources were compared.

12,458 different IGRs were identified across the *S.* Infantis human and poultry isolates; 20% (2464/12458) were core, 1% (149/12458) soft-core, 4% (454/12458) shell and 75% (9391/12458) cloud. The proportion of IGRs from human and poultry eBG31 isolates that were core and accessory is shown in Figure 5.7. While 6,139 IGRs were identified in the eBG31 poultry isolates, 11,286 IGRs were present in the eBG31 human isolates. This 83.8% increase, as seen with the structure of the pan genome, contributes to the varying structure of the IGR population by source.

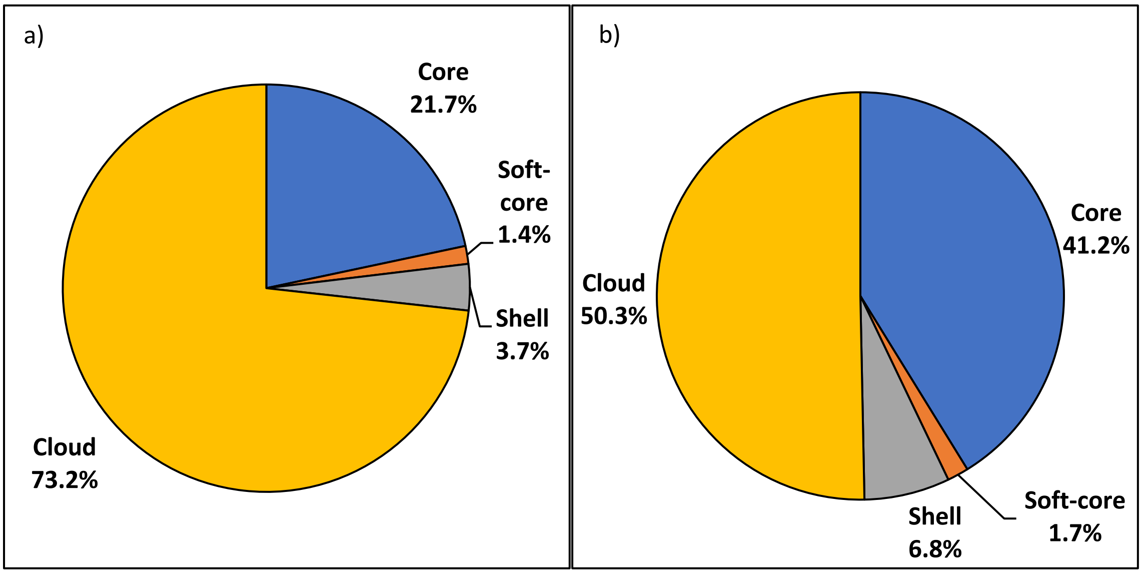


Figure 5.7 IGR composition of eBG31 human and eBG31 poultry isolates

a) IGR structure of eBG31 human isolates (n=1482)

b) IGR structure of eBG31 poultry isolates (n=945)

Core: 99% ⩽ Strains ⩽100%, Soft-core: 95% ⩽ Strains < 99%, Shell: 15% ⩽ Strains < 95%, Cloud: Strains < 15%

The distribution of the number of IGRs in each eBG31 human and poultry isolate was calculated (Appendix V Figure V.3). Despite a higher number of IGRs being identified from the human eBG31 isolates, the distribution of the number of IGRs in isolates from each source was significantly different (p-value < 2.2x10^{-16}). As seen with the distribution of genes, over 50% of the poultry eBG31 isolates contained more IGRs than seen in 75% of the eBG31 human isolates.

The number of IGRs in each isolate belonging to each component of the IGR is plotted in Figure 5.8. The eBG31 poultry isolates contained significantly more core IGRs than the eBG31 human isolates (p-value < 2.2x10^{-16}); only one eBG31 poultry strain contained fewer than the maximum number of core IGRs in the eBG31 human isolates, DRR002281, which was isolated from Asia in 2010 and contained 2,413 core IGRs.

A significant difference was also seen between the number of IGRs within each accessory IGR component from either isolation source (p-value < 2.2x10^{-16}). 99.6% (1476/1482) of the eBG31 human isolates had greater than 104 soft-core IGRs in their genome, the maximum that was seen in the poultry isolates. While significantly more shell IGRs were seen in the eBG31 poultry isolates, a significantly higher number of cloud IGRs were present in the eBG31 human isolates; 11 had greater than 188 IGRs, the maximum number in any eBG31 poultry isolate (p-value < $2.2\text{x}10^{-16}$). These isolates were from Europe, North America and Africa; isolated between 2008 and 2017.

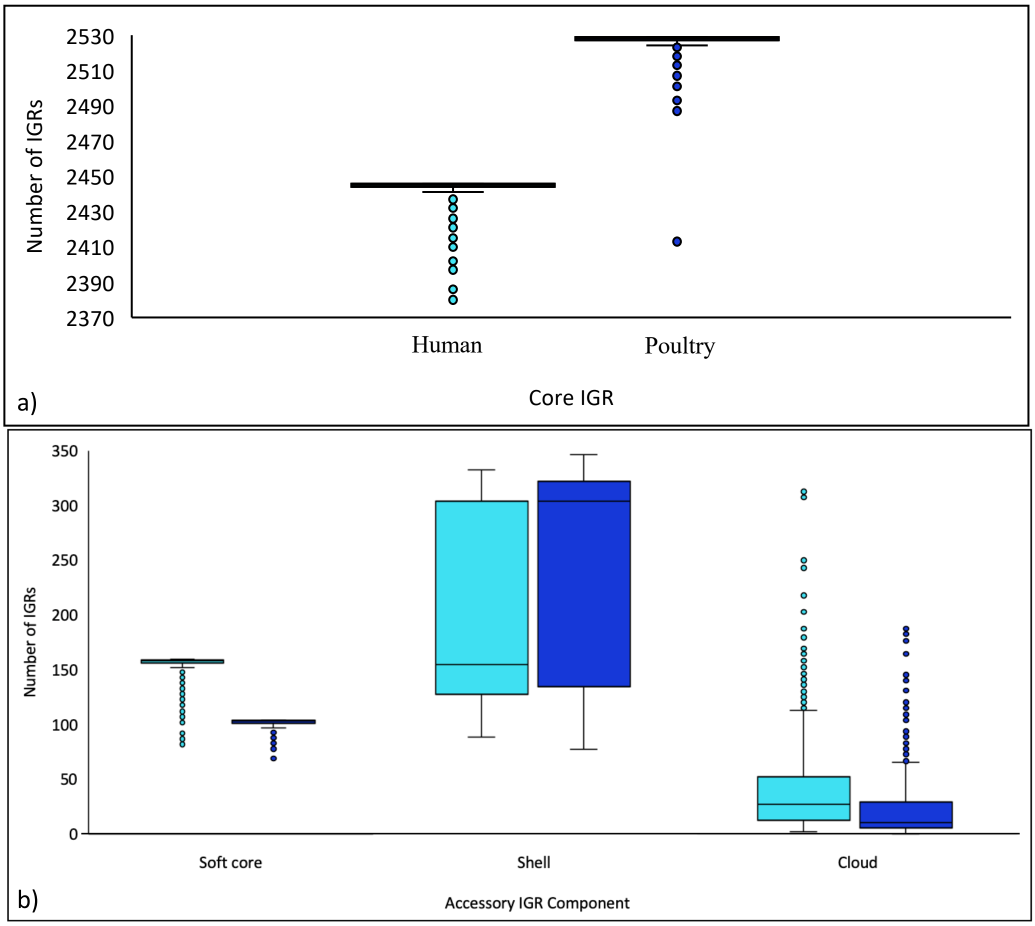


Figure 5.8 Distribution of the number of IGRs belonging to each IGR component from each eBG31 human and poultry isolate

a) Number of core IGRs in each human (n=1482) and poultry (n=945) eBG31 isolate

b) Number of soft-core, shell and cloud IGRs in each human (n=1482) and poultry (n=945) eBG31 isolate

Human ■ Poultry ■

5.3.6 IGRs Associated with Isolation Source

The association between each IGR and isolation from human or poultry sources was determined using *Scoary*. 963 IGRs were significantly associated with isolation source; 445 from human sources and 518 from poultry sources (p<0.05).

The difference in the percentage of isolates containing associated IGRs between each source was determined (Figure 5.9). Despite fewer IGRs being associated with isolation from humans than genes, the distribution in the difference between the occurrence of these IGRs in human and poultry eBG31 isolates is comparable to what was seen in the source associated genes (Appendix V Figure V.4). 74.4% (331/445) of the human associated IGRs were only present in up to 5% more human eBG31 isolates than poultry eBG31 isolates. However, the majority of the poultry associated IGRs were seen with a larger difference in percentage between the sources, with 50.2% (260/518) present in over 20% more poultry isolates.

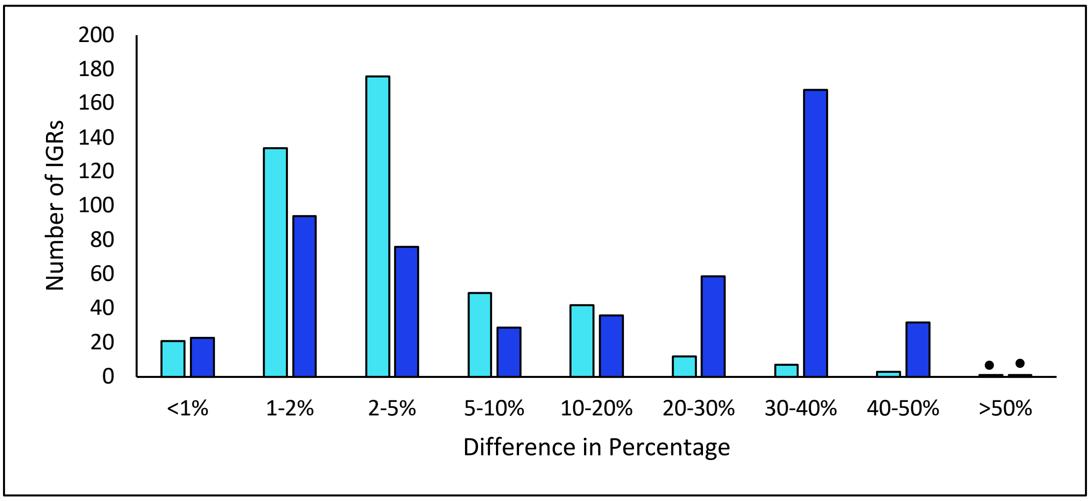


Figure 5.9 Distribution in the difference in the percentage of associated IGRs between sources
The percentage of isolates with a lower proportion of each significantly associated IGR was subtracted from the percentage of isolates with each IGR from the other source. • values are less than 2 but greater than 0.
Human (n=1482) 🟦 Poultry (n=945) 🟦

Two IGRs which were present in over 50% more isolates in its associated source group were identified, one human associated and one poultry associated. The poultry associated IGR was present in 55.3% (523/945) of the poultry isolates and 4.9% (73/1482) human isolates; the human associated IGR present in 95.0% (1408/1482) of human isolates and 44.7% (422/945) of poultry isolates. Both of these IGRs contained double promoters resulting in divergent transcription of the genes either side of it. They were also between the same genes on all the genomes, *yfkN* and a gene encoding a hypothetical protein similar to cytoplasmic proteins. Every isolate had only one of these two IGRs with the exception of one human isolate which had neither; the human associated one was 136 bases long and the poultry associated one, 634 bases. Neither of the IGRs contained ncRNAs previously identified in other *Salmonella* serovars; as the ncRNA suite in *S*. Infantis has not yet been solved it is possible that ncRNAs unique to *S*. Infantis were present.

The eBG31 soft-core SNP phylogeny (Figure 3.9) was annotated with the variant of the IGR that the human and poultry isolates contained (Appendix V Figure V.6). 99% (592/596) of the isolates with the poultry associated IGR variant belonged to the same

25SNP cluster, which could indicate that they were part of an outbreak. 89% (520/596) of the isolates in that cluster with that variant were from poultry sources, isolated between 2014 and 2018 and from 30 different states across the USA.

5.3.7 IGRs Unique to an Isolation Source

Of the 12,458 IGRs that were identified in the *S.* Infantis isolates, 1,172 were found exclusively in poultry isolates, 45 of which were significantly associated and 6,319 in human isolates, 180 that were significantly associated ($p < 0.05$). The distribution of the number of unique IGRs in each of the isolates is shown in Figure 5.10.

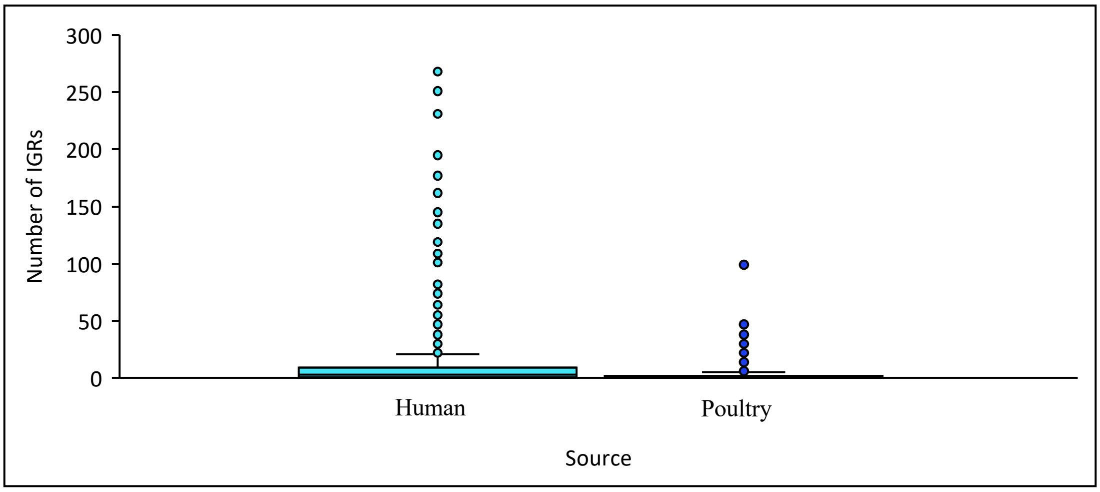


Figure 5.10 Distribution of the number of IGRs unique to eBG31 human and eBG31 poultry isolates
Box plot showing the variation in the number of unique IGRs in each eBG31 isolate from human sources (n=1482) and poultry sources (n=945).
Human 🟦 Poultry 🟦

The distributions of unique IGRs from human and poultry sources were significantly different (p-value < $2.2x10^{-16}$). Whilst 78.5% (742/945) of strains had fewer than 3 IGRs unique to poultry isolation, 55.6% (796/1432) of the human isolates contained 3 or more IGRs unique to human isolation. Additionally, significantly more poultry isolates contained 0 unique IGRs than human isolates (human n=286, poultry n=380, p-value < 0.00001).

The distribution of isolation frequency of IGRs unique to isolation source was calculated (Appendix V Figure V.5). 72.3% (847/1172) of IGRS unique to poultry and 62.3% (3935/6319) of IGRs unique to humans were only identified in 1 isolate. However, several were present in larger number of isolates, with a higher number of IGRs unique to isolation from humans seen in multiple isolates. 8 unique IGRs were identified in over 40 eBG31 human isolates, 6 of these were present in between 40 and 60 isolates. An IGR between *cysG* and a gene encoding a hypothetical protein, promoting transcription in this order, was present in 87 isolates. The most common IGR unique to human sources was present in 105 isolates and was predicted to divergently transcribe two genes encoding hypothetical proteins. 1 IGR unique to poultry sourced isolates was seen in 50 isolates and was between tyrosine recombinase producing gene *xerC* and a gene encoding an IS6 family transposase which were transcribed in the order written.

5.3.8 SNPs Associated with Isolation Source

In order to identify other components of the genome that may be associated with isolation source, variation within the core genome, core IGRs and unitigs were investigated.

5.3.8.1 *Differences Within Core Genetic Elements*

A maximum likelihood phylogeny was generated using the core gene alignment of 1,482 human eBG31 isolates and 945 eBG31 poultry isolates produced by *Roary* (Figure 5.11).

Clustering by isolation source is clear within the phylogeny. The majority of the isolates from poultry sources cluster in one section of the phylogeny and although interspersed with isolates from humans, several clades are present containing only poultry isolates. Clustering of isolates from human sources is also evident with several large clades entirely comprised of these isolates.

A maximum likelihood phylogeny of the core IGRs from the same isolates was also generated (Appendix V Figure V.7). Similar patterns of clustering by isolation source were seen in this phylogeny with the majority of the eBG31 poultry isolates clustering in one section and several large clades containing isolates from humans also present.

To identify whether any of the clades within the core gene and IGR phylogenies had distinct phenotype distributions, the software *TreeBreaker* was used (Appendix V Figure V.8, V.9). Several of the branches within both phylogenies were significantly associated with a change of isolation source, having a posterior probability of greater than 0.5.

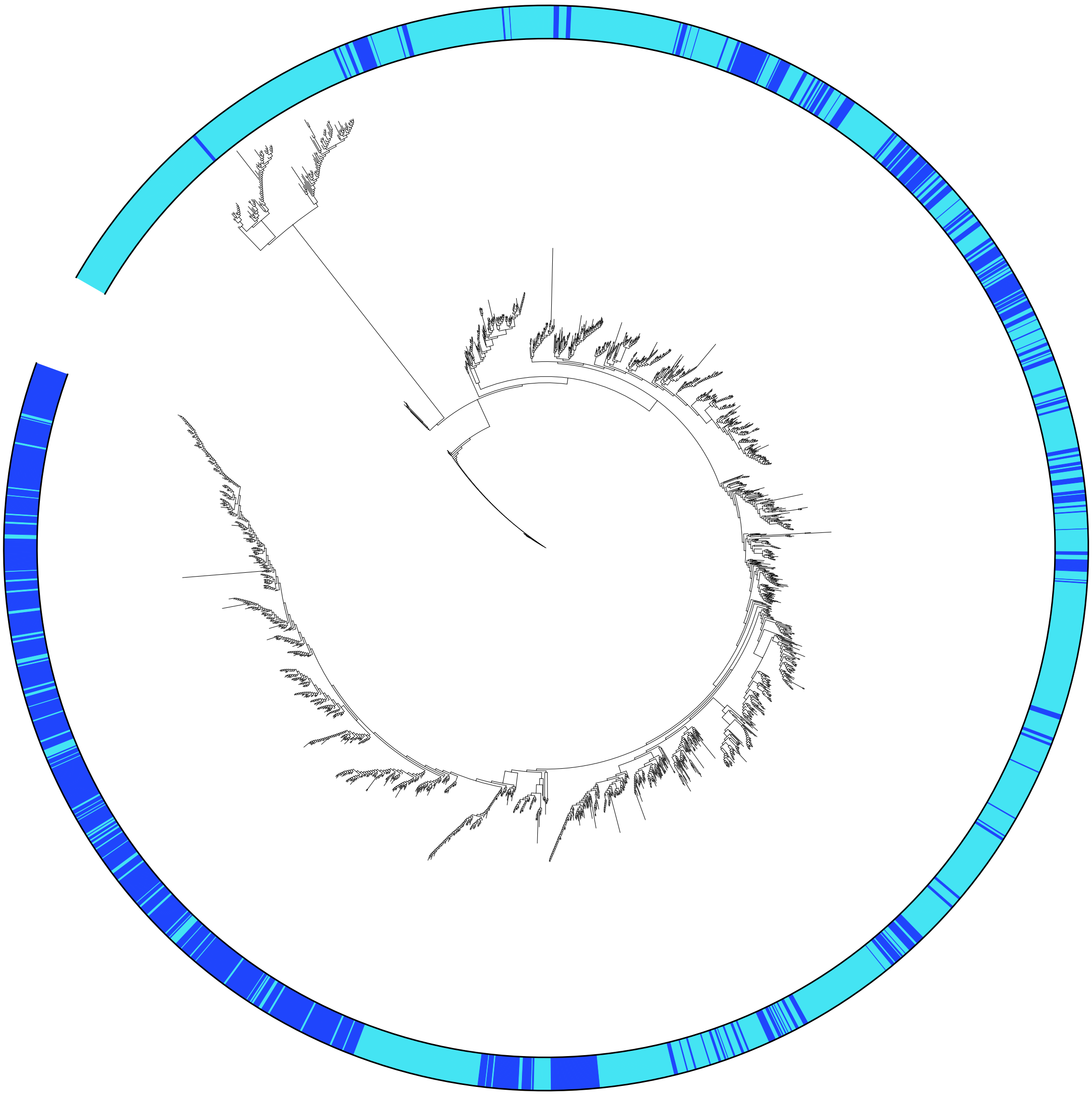


Figure 5.11 Maximum likelihood phylogeny of core genes

Phylogeny of 4132 core genes from 1482 eBG31 human isolates and 945 eBG31 poultry isolates.

Human □ Poultry ■

5.3.8.2 *Unitigs Associated with Isolation Source*

To identify smaller areas of the genome that were associated with isolation source, *PySeer* was used with unitigs as the input. The contigs assembly format was used with poor-quality assemblies excluded; resulting in the inclusion of 1483 eBG31 isolates from human sources and 944 from poultry sources. 1 isolate from human sources and one from poultry sources that had been included in the pan genome analyses were excluded and two new human samples were included.

70,216 unitigs were identified which were associated with isolation from either human or poultry sources. Those with a population structure adjusted and Bonferroni corrected p-value of less than $1.02x10^{-6}$ were classed as significantly associated, resulting in 17,052 significant unitigs, 6,539 that were associated with humans and 10,508 with poultry. The length of the significant unitigs varied, the smallest unitig was 31 nucleotides long, found from both sources; the largest 1,411 nucleotides in the human isolates and 6,522 in the poultry isolates.

Figure 5.12 shows the inverse p-values for the significant unitigs and their position across the genome. Significant unitigs associated with both isolation sources are present throughout the genome. Higher levels of significance were observed in the poultry associated unitigs, with many of the unitigs having higher inverse p-values than the human associated unitigs.

3 clusters of significant unitigs were identified in Figure 5.12, regions 1, 2 and 3. While significant unitigs were seen in both source groups, regions 1 and 3 were more densely populated in the poultry eBG31 isolates and region 2 more so in the human eBG31 isolates. The genes present in these regions were identified from the eBG31 reference annotation (Appendix VI.6). Region 1 contained 72 genes, 55 of which encoded hypothetical proteins and the remainder encoded proteins involved in, for example, lipid biosynthesis and transport, DNA replication and translation and formate transmembrane activity. Region 2 contained 34 genes, 29 encoding hypothetical proteins and 5 genes encoding known proteins whose functions were: translation regulation, cell membrane component, rRNA processing, heat response and copper ion binding.

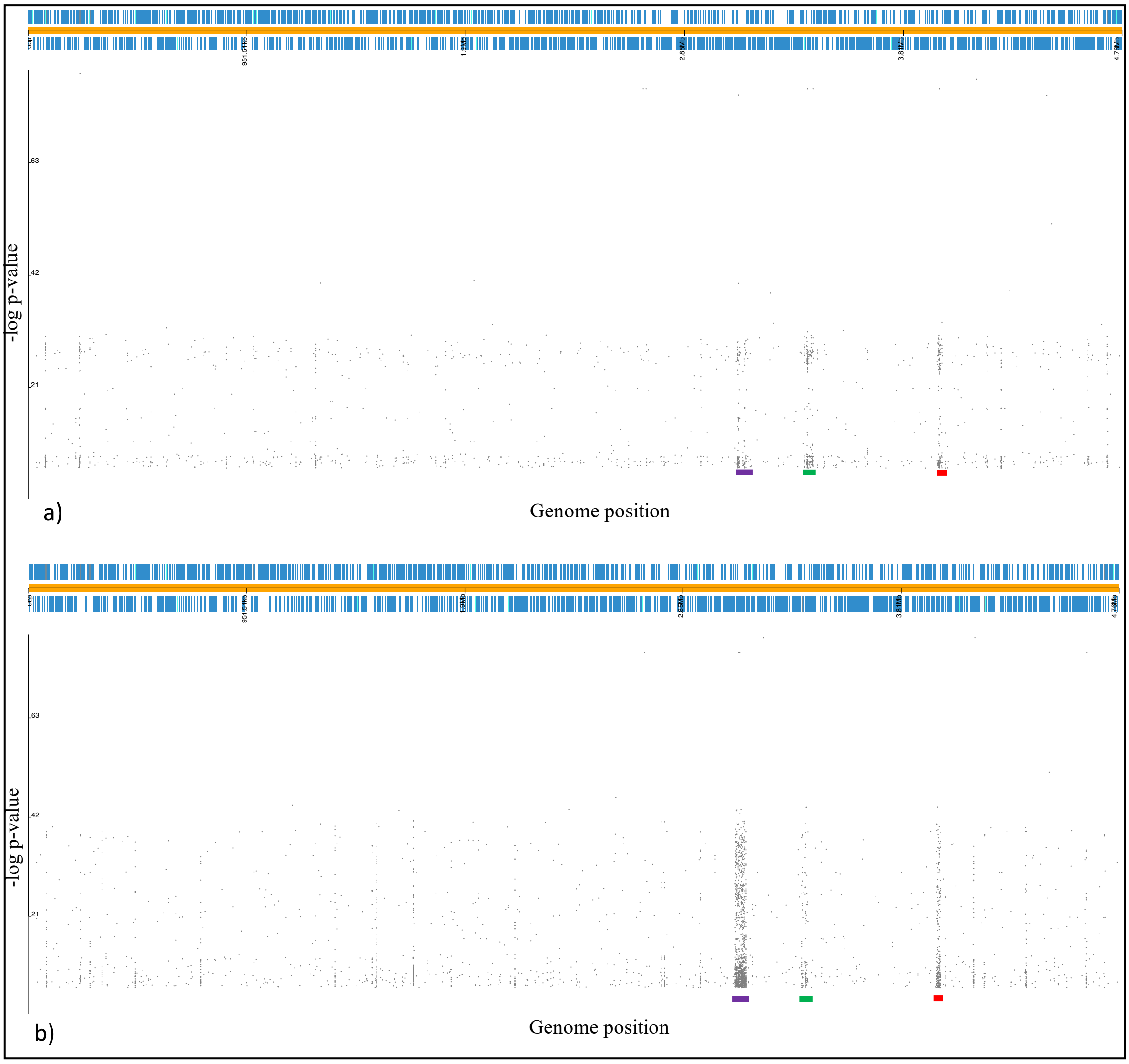


Figure 5.12 Manhattan plots of significant unitigs
17,052 unitigs identified in 1483 eBG31 human isolates and 944 eBG31 poultry isolates that were significantly associated with isolation source plotted against the eBG31 reference.

a) Unitigs significantly associated with isolation from humans (n=6,539)
b) Unitigs significantly associated with isolation from poultry (n=10,508)

🟪 region 1 🟩 region 2 🟥 region 3

Region 3 contained 53 genes, 39 of which encoded hypothetical proteins. The other 14 encoded proteins involved in, for example, nucleic acid repair, transcription regulation, translation, metal ion binding and also genes associated with recombination and prophages. Also present was the *uppP* gene.

5.4 Discussion

The aim of this chapter was to identify any genetic basis for *S.* Infantis being the serovar most frequently identified from domestic fowl in Europe but only the fourth most common serovar causing infection in humans (EFSA and ECDC, 2018b). To the authors knowledge, this GWAS is the largest performed on isolates within a *Salmonella* serovar.

Although 9 virulence factors were identified which were significantly associated with isolation from humans, none of these were seen at a high frequency across the human isolates and therefore cannot explain the difference in prevalence of *S.* Infantis in poultry and humans. A higher number, 21 virulence factors, were significantly associated with isolation from poultry, including *irp1*, the yersiniabactin biosynthetic protein which has previously been identified on pESI-like plasmids (Franco *et al.*, 2015). Whilst the increased association with virulence factor presence in poultry isolates could explain the high incidence of *S.* Infantis amongst poultry isolates, it does raise the question, is poultry a major source of human *S.* Infantis infection? This could indicate either that whilst the poultry associated virulence factors are needed for colonisation of poultry, they are not needed for human colonisation; or that the main contributor to human *S.* Infantis infection is not, as is suspected, poultry meat.

Whilst 4,132 genes were seen in all but 1% of isolates, only 34 genes were seen in the soft-core genome, suggesting that aside from the genes conserved in both human and poultry eBG31 isolates, the genes were seen a lot more sporadically. The vast majority of the genes were seen in less than 15% of the isolates, indicating the genetic diversity within the eBG31 population. There also appears to be greater genetic diversity in eBG31 than in some *Salmonella* serovars; a comparison of 1,326 genomes from an eBG containing monophasic *S.* Typhimurium sequences identified fewer core genes than in eBG31 (4,050), a smaller pan genome (13,135) and fewer cloud genes (8,588) (Palma *et al.*, 2018). However, a study of 622 *S.* Enteritidis sequences found a pan-genome of 18,091 genes; comparable to what was observed in eBG31 (Feasey *et al.*, 2016).

The structure of the pan genome varied by isolation source, most noticeably in the overall number of genes identified across the isolates, with 83.3% more genes being identified from the human isolates. Conversely, the distribution of the number of genes in each isolate were significantly different, with poultry isolates having more than those from humans. A larger proportion of the eBG31 poultry isolates pan genome was comprised of core genes, suggesting that there is less variation within the eBG31 poultry population when compared to the eBG31 human population. The variation within the eBG31 human population was not due to a small number of divergent sequences, as 50% of the human sequences contained over 43 cloud genes; significantly fewer cloud genes were identified from each of the eBG31 poultry isolates.

Unsurprisingly, considering the variation in the structure of the pan genome, there were several thousand genes that were significantly associated with isolation from human or poultry sources. Whilst there was a greater number of genes which were significantly associated with isolation from human sources, the majority of these genes were seen in low numbers of isolates, again highlighting the diversity within the human eBG31 population. Even though a smaller number of genes were significantly associated with isolation from poultry, they were seen in higher frequencies, suggesting either several instances of emergences of these genes or small outbreaks of isolates sharing them. Whilst this again highlights the difference between *S.* Infantis isolates from these sources, this data can currently only be used to speculate that perhaps some of these genes, or even a combination of them, need to be present in order to cause human infection. The increased number of genes associated with and unique to eBG31 human isolates does support this hypothesis, also emphasising the greater genetic diversity amongst the eBG31 human isolates. This could suggest that the source of infection for these human cases was not poultry sources, supporting the conclusions drawn from Chapter 4 that despite the prevalence of *S.* Infantis in poultry, it is not the main contributor to infections in humans. As the human isolates had similar AMR levels to those seen in the environmental isolates it could be that this source group causes a large number of human *S.* Infantis infections.

Two variants of *tufA*, the translation elongation factor (UniProt, 2019b), were identified in the *S.* Infantis human and poultry isolates; one of the variants significantly associated with human eBG31 isolates and the other with isolates sourced from poultry. The fact that 86.9% of human eBG31 isolates, collected over across three decades, have maintained *tufA*_2 indicates that there is a selective advantage for the presence of this gene in colonising the human host. As previous research has identified that synonymous codon changes in the *tufA* and *tufB* genes in *S.* Typhimurium altered levels of protein expression, it is plausible that the variants of *tufA* associated with human eBG31 isolates are associated with different levels of protein expression and that this is advantageous in causing infection in humans (Brandis and Hughes, 2016). It could also be argued that the increased prevalence of the *tufA* variant significantly associated with poultry eBG31 isolates is a factor in the success of *S.* Infantis in poultry. If the *S.* Infantis poultry population is split into two groups, with either variant of *tufA*, but isolates containing only one of the variants are more virulent to humans, then this immediately reduces the likelihood of a human coming in contact with an *S.* Infantis isolate likely to cause infection. It would be pertinent to identify the difference between the variants of *tufA*, to determine if it is responsible for the different frequencies in *S.* Infantis isolation from humans and poultry.

The composition of the IGRs in eBG31 was comparable to the pan genome structure. The number of core IGRs in eBG31 was larger than the number found in a comparison of 68 S. Typhimurium (1,576), suggesting greater IGR diversity in eBG31 (Fu *et al.*, 2015). Significantly more IGRs were identified in the eBG31 human isolates that were present in fewer than 15% of the samples. This supports the findings that the eBG31 human isolates are more diverse than the isolates from poultry. Despite the overall increased number of IGRs identified from eBG31 human isolates, significantly more IGRs were present in each of the eBG31 poultry isolates. The increased number of genes and IGRs present in the poultry eBG31 isolates adds further evidence that the two source groups are two separate niches within the *S.* Infantis population; if the source of human infection was mainly poultry products then the frequency of genes and IGRs identified would be similar.

As the majority of the IGRs unique to a source were identified only in 1 isolate this suggests that they are not maintained in the *S.* Infantis population and that they were acquired sporadically. The presence of unique IGRs in multiple isolates suggests either a possible outbreak or that the altered gene expression caused by the IGR is beneficial to the bacteria.

IGRs were identified that were significantly associated with a source group. The IGR that was most frequently identified exclusively from poultry was between recombinase and transposase encoding genes; this could be attributed to the increased percentage of eBG31 poultry isolates with plasmids, in particular pESI, when compared to isolates from humans (Chapter 4.3.2.2).

Two IGRs were identified in all but 1 isolate, between a gene encoding a hypothetical protein and *yfkN* which encodes a trifunctional nucleotide phosphoesterase protein involved in DNA translation (UniProt, 2019a). One of the variants was significantly associated with isolation from humans, with 95% of isolates from humans containing this IGR and 44.7% of the poultry isolates containing it. However, upon annotating the eBG31 phylogeny with these variants it was discovered that 99% of the poultry associated variant of the IGR were from the same 25SNP cluster in the phylogeny, indicating that they could be part of an outbreak. However, as these strains were isolated over several years and from 30 states in the USA, this could suggest that this strain of *S.* Infantis is endemic in poultry in the USA and that the human-associated variant is what is normally seen in *S.* Infantis. It is possible that this IGR variant could be attributed to the success of this endemic strain. It would therefore be beneficial to identify what this IGR does to determine the risk it presents.

Clustering by isolation source was identified within both the core gene and core IGR phylogenies with large clades containing exclusively eBG31 human isolates evident. There were also large clusters of poultry isolates that were interspersed by human isolates, supporting the hypothesis that a subset of poultry *S*. Infantis are associated with human infection.

A SNP in an IGR has been identified in *S*. Typhimurium that differentiates between ST19 and ST313, causing increased virulence gene transcription in ST313 which is associated with causing large numbers of iNTS infections (Hammarlöf *et al*., 2018). It is therefore possible that SNPs in the *S*. Infantis IGRs could also affect transcription of a virulence factor, resulting in organisms that are more virulent to a particular host. Whilst the genes and IGRs included were core to isolates from both sources, there were SNPs in these genes that were maintained within clusters of organisms in the phylogeny. Many of these clusters contained isolates from both sources, but there were several clades containing isolates from one source. This supports the hypothesis that some of the *S.* Infantis isolates that infect humans and poultry originate from two genetically distinct populations, but that there are also cases where poultry products have caused human infection.

Several thousand unitigs were identified throughout the genome that were significantly associated with either human or poultry sources. This is considerably larger than seen in a comparison of 440 *S. enterica* genomes, where only 52 areas of the genome were found to be significantly associated with isolation source (Vila Nova *et al.*, 2019). The fact that so many significantly source associated unitigs have been acquired and been conserved in the population suggests that the organisms causing infections in humans and poultry are two distinct populations. Three regions were identified containing a high density of significant unitigs in the genomes, two of which were associated with poultry and the third with humans.

For such a high density of unitigs to be present in a region could be explained by that region not being conserved and that SNPs have accumulated over time; however, that would not explain why these unitigs were all significantly more present in isolates from one of the sources. These regions therefore may be evidence of adaptation to niche. The majority of the genes in these regions encoded hypothetical proteins or housekeeping genes. In one of the more poultry associated regions was the gene *uppP* which is predicted to confer resistance to the antibiotic bacitracin (UniProt, 2019c). It is possible that the poultry associated eBG31 isolates have acquired mutations in this gene or in the other genes that gives them a selective advantage when infecting poultry; increasing the occurrence of *S.* Infantis in this source group.

5.4.1 Conclusion

To conclude, despite the high numbers of *S.* Infantis cases reported in domestic fowl, it appears that either only a subset of the isolates from poultry are capable of causing infection in humans or that other sources are responsible for causing a large proportion of human infections. While a genetic element was not identified that was present in all human eBG31 isolates and none of the isolates from poultry, several candidate genes, IGRs and unitigs were identified which were significantly associated with isolation from humans. Although further research is needed to identify how the presence of these genetic elements affects the colonisation of humans, this information is of use to public health teams as isolates positive for these elements have a higher risk of causing infection in humans; this therefore allows prioritisation of *S.* Infantis case control in poultry. Identifying the association with other sources and human infection would also be beneficial.

6. Chapter 6. Overall Discussion

In this project I amassed a collection of 4,670 *S*. Infantis genomes, determining the global population structure, AMR and plasmids levels and performing a GWAS comparing human and poultry eBG31 isolates. This work represents the largest comparison of *S*. Infantis genomes, with the largest published work currently containing just 264 genomes (Acar *et al.*, 2019).

Whilst eBG31 was the dominant eBG seen in *S.* Infantis across the globe, the proportion of isolates belonging to eBG297 varied by continent; globally 3.9% were eBG297 but this increased to 34.1% in Africa. The genetic distance within and between the eBGs was identified; despite the number of isolates included in the eBG31 collection, the distance between eBG297 and eBG31 was 4.1 times greater than the distance within the eBG31 collection. This suggests that the two eBGs are too genetically distinct to belong to the same serovar, raising the question, should eBG297 be called *S*. Infantis? In order to answer this further research should be performed, determining whether they are biologically different and comparing the distance between and within other *S. enterica* serovars.

A strong geographical signal was identified in the eBG31 phylogeny, with the six clusters in the phylogeny each dominated by isolates from a single continent. The African eBG31 isolates were found to be the most distant from other continents, suggesting African associated lineages within the eBG31 population. Clustering by isolation source was also present in the phylogeny; this information on the geographic and isolation source signals within the eBG31 population will be beneficial to public health teams when investigating *S.* Infantis outbreaks. As the methods used in database generation mirrored those used by PHE, the databases created can be imported into their system; the lineage that new isolates, when added to the databases, belong to could then indicate their source and continent of origin.

A difference in AMR levels was observed between the two eBGs, with just 1.6% of eBG297 having MDR versus 37.9% of eBG31. The levels of AMR in the eBG31 isolates were found to be increasing, with numerous AMR profiles maintained in the population over several years. The geographical signal seen in eBG31 was also observed in the accessory genome, with several AMR gene clusters and plasmids found exclusively in a continent.

This work has also shown the importance of pESI in the eBG31 population, identifying that whilst not present in eBG297, 33% of the eBG31 isolates contained the plasmid. An association between pESI presence in poultry isolates was identified, with 69.7% of poultry isolates containing pESI. Also of concern is the significant increase with time in occurrence of bla$_{CTX-M-65}$ on pESI positive poultry isolates, present in 48.6% of poultry isolates with the plasmid. The presence of this gene, as well as *aadA1*, *sul1*, *tetA* and *dfrA14* on pESI, is a public health concern due to the obvious risk to human and chicken health it presents if the number of isolates with the plasmid increases. Future work could include the identification of other ESBLs or other AMR genes present on pESI to further identify the risk it presents. Low levels of pESI were identified in isolates from Africa, an increase in incidence here would also be concerning due to the increased number of immuno-compromised people in this continent. As a result of these findings, testing for pESI presence in any *S.* Infantis isolates identified by public health teams should be implemented as the spread of the plasmid needs to be monitored.

Differences in the source distribution were seen between the eBGs, with no eBG297 isolates being isolated from poultry. This could suggest that eBG297 cannot colonise domestic fowl, however, it is more likely that due to the low numbers of isolates identified in that eBG, eBG297 isolates have not yet been found in poultry. Sequencing of *S.* Infantis isolated from poultry in South Africa, where eBG297 was frequently identified, would be beneficial in identifying whether eBG297 is also present in poultry.

Large clusters of human isolates were visible in the eBG31 phylogeny which led to the hypothesis that some lineages of *S*. Infantis have adapted to become more virulent to humans. Analysis of the pan and core genome found key differences between eBG31 isolates from human and poultry sources. Greater diversity was seen amongst the human isolates, with a pan genome that was 83.3% larger than seen in the poultry isolates. Furthermore, several thousand genes and kmers were found to be significantly associated with one of the sources. An example is *tufA*, two variants of this gene were identified in the isolates, *tufA*_2 which was significantly associated with humans and the other, *tufA*_1 with poultry. Despite the three-decade time period that the human isolates were collected over, 86.9% of the isolates contained *tufA*_2, suggesting it confers a selective advantage for infection in humans.

Due to the findings of this research, I have two hypotheses to explain the low numbers of *S.* Infantis infection seen in humans. The first being that the *S.* Infantis poultry population is split into two groups, with one containing a genetic element, such as *tufA*_2, which increases the pathogen's ability to colonise humans and the other group less capable of infecting humans. The alternative hypothesis is that poultry products aren't in fact the major source of human infection and that other environmental sources are associated with more infections. This is supported by the human and environmental isolates having the largest number of shared AMR gene clusters and a smaller genetic distance than seen between human and poultry isolates.

In order to determine whether either of these hypotheses are correct, the next step would be to identify whether any of the genes, IGRs and kmers that have been identified as being associated with human infection are advantageous for *S.* Infantis survival within the human host. Further experiments, such as the use of a chicken caecum model and human colon model, could be used to identify whether these candidate genetic elements are advantageous for *S.* Infantis survival in humans or poultry. A genome-wide screening method such as TraDIS could be used to quantify how knocking out genetic elements identified as being significantly associated with either source group affects *S*. Infantis survival. The use of TraDIS would also be beneficial as it could identify genes essential for survival in both the hosts.

Additionally, another GWAS could be performed, comparing the eBG31 human and environmental samples to see if they are genetically more similar than when compared to the poultry isolates. It may be beneficial to break down the environmental sources into smaller groups such as cattle and swine. An alternative method that could be implemented is machine learning. Previous researchers have used a Random Forest classifier to perform source attribution of a global collection of *S.* Typhimurium isolates; identifying 50 genetic regions that could be used to predict whether the isolate was from livestock (Zhang *et al.*, 2019). Support Vector Machine classifiers have been built that discriminate between bovine and human *E. coli* isolates with an accuracy of 83% (Lupolova *et al.,* 2017). This approach has also been used to identify how similar genomic features are between *E. coli* isolates from different hosts, predicting the zoonotic potential of isolates from cattle (Lupolova *et al.*, 2016). Another approach that machine learning has been used for is the identification of strains associated with iNTS infection, with a Random Forest classifier developed that discriminated between invasive and gastrointestinal *Salmonella* serovars (Wheeler, Gardner and Barquist 2018). It would be beneficial to apply these approaches to the *S.* Infantis collection to identify sub-groups of strains that pose a greater risk of causing invasive infections; to further explore the differences between the strains isolated from human, poultry and environmental sources and identify which source the human isolates resembled more closely.

One limitation in this project was the restrictions the use of SnapperDB put on the analyses that could be done. Whilst using the software was beneficial as it allowed calculations of clustering present with eBG31 and eBG297 and enables integration of the results directly with PHE's data analysis pipeline; it also limited the number of sequences that could be included in the project due to the time the software took to run once the databases were populated. This resulted in the requirement of a cut-off date for when sequences could no longer be included from PHE or Enterobase. Also, as each database could only contain isolates belonging to one eBG, this limited the phylogenetic comparisons that could have been performed between eBG31 and eBG297 isolates.

Previous studies have identified that long term storage of *Salmonella* strains in agar is associated with genome rearrangement (Edwards *et al*., 2001; Porwollik *et al.,* 2004; Matthews, Rabsch and Maloy, 2011) and plasmid loss when compared to isolates that are frozen (Olsen *et al.,* 1994). As the hPHE isolates were stored on agar it is possible that the numbers of plasmids identified from these isolates is an underrepresentation of what was originally present.

Another limitation to this project was the potential of sampling bias. Due to the inclusion of samples from online databases, often lacking metadata, it is possible that smaller studies selected isolates for sequencing for reasons such as increased virulence. Historically there was also a sampling bias for the collection of *Salmonella* isolates by the NICD from blood in South Africa. For these reasons, calculations such as the invasive index of *S.* Infantis could not be performed.

6.1 Conclusion

In conclusion, *S.* Infantis is a polyphyletic serovar comprised of eBG31, whose population splits into lineages associated with geography, and eBG297 which is strongly associated with isolation from Africa. High levels of AMR were identified in the eBG31 population, associated with pESI, which was especially common in isolates from poultry. The eBG31 human population had a greater genetic diversity than seen in the eBG31 poultry population with several thousand genes significantly associated with either source. This could explain the difference in the incidence of *S*. Infantis in poultry and humans, it is possible that only a sub-group of *S*. Infantis is capable of causing human infection; or that sources other than poultry are causing the human infections seen. The increased understanding this work provides on this emerging pathogen will be beneficial for public health teams globally.

I Appendix I. Scripts

I.I Input filtering for MOST

This script pulls out the required lines from the MOST output, sorts them, removes unwanted characters and strings; resulting in a summary file for each xml input (Tewolde *et al.*, 2016).

```
#!/bin/bash

sample=$1

find $1 | xargs egrep -n '<result type="MLST" value="' > result$1.xml
find $1 | xargs egrep -n 'QC_max_percentage_non_consensus_base_for_all_loci' > consensus$1.xml
find $1 | xargs egrep -n '<ngs_sample id' > id$1.xml
find $1 | xargs egrep -n ' <result_data type="QC_traffic_light" value=' > traffic$1.xml
sort -g traffic$1.xml result$1.xml consensus$1.xml id$1.xml > summary$1
awk ' />/ {print}' ORS=',' summary$1 > line$1
tr -d " \t" < line$1 > nospaces$1
sed 's/\"//g' nospaces$1 > noquotes$1

sed -e "s/1:<ngs_sampleid=//g" noquotes$1 | sed -e "s/_1.fastq//g" |sed -e "s/_R1.fastq//g" | sed -e
"s/4:<resulttype="MLST"value=//g" | sed -e
"s/12:<result_datatype=QC_max_percentage_non_consensus_base_for_all_locivalue=//g" | sed -e
"s/14:<result_datatype=QC_traffic_lightvalue=//g"  > sed$1
sed -e "s/>//g" sed$1 > sed2$1
sed 's/\///g' sed2$1 > filtered_$1

rm result$1.xml
rm consensus$1.xml
rm id$1.xml
rm traffic$1.xml
rm summary$1
rm line$1
rm nospaces$1
rm noquotes$1
rm sed$1
rm sed2$1
```

I.II Concatenating filtered output for MOST

This script concatenates all summary xml outputs of 'Input filtering for MOST' in a folder into one xml, inserting a line break between each entry and adding titles to each column.

```
#!/bin/bash

awk 1 ./*filtered* > tmp_awk.xml

echo -e "ID, Sequence Type, Max  Percentage Non Consensus Base for all loci, , , , , , , Traffic Light," | cat -
tmp_awk.xml > filtered_cat_results.xml

rm tmp_awk.xml
```

I.III Annotating a phylogeny containing cluster representatives

This script requires a metadata file containing sequence ID, the metadata of interest and the SNP address of each isolate at the cluster level used to generate the phylogeny. It is run using a while loop with a list of the representative sequences and their SNP addresses. It collates all isolates in each cluster into a file with the metadata and then counts the occurrences of each metadata type in each cluster.

```
#!/bin/bash

address=$1
representative=$2

echo -e "with open('"All_eBG31_Metadata.txt"', '"r"') as input_file, \\" >>$representitive.py
echo -e " \t open('$representitive.txt', 'w') as output_file:" >>$representitive.py
echo "" >>$representitive.py
echo -e " \t for line in input_file:" >>$representitive.py
echo -e " \t\t if '   $address' in line:" >>$representitive.py
echo -e " \t\t\t output_file.write(line)" >>$representitive.py
python $representitive.py
rm $representitive.py

address=$1 >>meta_$representitive.py
representative=$2 >>meta_$representitive.py

echo "import re" >>meta_$representitive.py
echo "from collections import Counter" >>meta_$representitive.py
echo "" >>meta_$representitive.py
echo "with open('"$representitive.txt"', '"r"') as input_file, \\" >>meta_$representitive.py
echo "    open('meta_"$representitive.txt"', 'w') as output_file:" >>meta_$representitive.py
echo "" >>meta_$representitive.py
echo "    for i in input_file:" >>meta_$representitive.py
echo "            i = i.rstrip()" >>meta_$representitive.py
echo "" >>meta_$representitive.py
echo "            array = re.split(r'\t+', i)" >>meta_$representitive.py
echo "            id = array[0]; origin = array[1]; address = array[5]" >>meta_$representitive.py
echo "            #print id, origin, year, travel, mlst; raw_input()" >>meta_$representitive.py
echo "" >>meta_$representitive.py
echo "            match = re.search('"Africa"', origin)" >>meta_$representitive.py
echo "            if match:" >>meta_$representitive.py
echo "                    output_file.write('Africa, ')" >>meta_$representitive.py
echo "" >>meta_$representitive.py
echo "            match = re.search('"NAmerica"', origin)" >>meta_$representitive.py
echo "            if match:" >>meta_$representitive.py
echo "                    output_file.write('NAmerica, ')" >>meta_$representitive.py
echo "" >>meta_$representitive.py
echo "            match = re.search('"Asia"', origin)" >>meta_$representitive.py
echo "            if match:" >>meta_$representitive.py
echo "                    output_file.write('Asia, ')" >>meta_$representitive.py
echo "" >>meta_$representitive.py
echo "            match = re.search('"Europe"', origin)" >>meta_$representitive.py
echo "            if match:" >>meta_$representitive.py
echo "                    output_file.write('Europe, ')" >>meta_$representitive.py
echo "" >>meta_$representitive.py
echo "            match = re.search('"SAmerica"', origin)" >>meta_$representitive.py
echo "            if match:" >>meta_$representitive.py
echo "                              output_file.write('SAmerica, ')" >>meta_$representitive.py
echo "" >>meta_$representitive.py
echo "                match = re.search('"Unknown"', origin)" >>meta_$representitive.py
echo "                if match:" >>meta_$representitive.py
echo "                              output_file.write('Unknown, ')" >>meta_$representitive.py
python meta_$representitive.py
rm meta_$representitive.py

address=$1 >>count_$representitive.py
representative=$2 >>count_$representitive.py

echo "import re" >>count_$representitive.py
echo "import collections" >>count_$representitive.py
echo "" >> count_$representitive.py
echo "wordcount = collections.Counter()" >> count_$representitive.py
echo "with open('"meta_$representitive.txt"') as file:" >>count_$representitive.py
echo "  for line in file:" >>count_$representitive.py
echo "          wordcount.update(line.split())" >>count_$representitive.py
echo "" >> count_$representitive.py
echo "for k,v in wordcount.iteritems(): print '"$representitive"', k, v," >>count_$representitive.py

python count_$representitive.py
rm count_$representitive.py
```

I.IV Calculating the number of zero's in the pESI coverage matrix

Using the pESI coverage as input, this script calculates the number of bases with zero reads coverage.

```
#!/bin/bash

location=$1
sample=${location#*\/}

awk '{print $2}' /Results/"$location"_pesi_coverage.txt > col_$sample
grep -c 0 col_$sample > n_$sample
printf '%s' "$sample:" | cat - n_$sample
rm col_$sample
rm n_$sample
```

I.V Generating a coverage matrix for pESI

This script produces a bsub file which is submitted to a high-performance cluster system where the job is then run. *SMALT* is used to map the sequence to the pESI pseudomolecule (Ponstingl and Ning, 2014). *SAMtools* then sorts the output into a BAM file and calculates the depths of coverage for every base, including those with no coverage (Li *et al.*, 2009). The pESI coverage results are then pulled out and given a heading.

```
#!/bin/bash

location=$1
sample=${location#*\/}


echo "CREATE bsub scripts"
echo $sample

echo "#!/bin/sh" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "#BSUB -q short-eth" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "#BSUB -R "rusage[mem=1600]"" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "#BSUB -M 1600" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "#BSUB -J mapping" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "#BSUB -oo mapping_"$sample".out" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "#BSUB -eo mapping_"$sample".err" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo ". /etc/profile" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "module add smalt/0.7.6" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "module add samtools/1.5/gcc" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "mkdir /Mapping_pESI/"$sample"" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "mkdir /Mapping_pESI/"$sample"" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "mkdir /Coverage/"$sample"" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "cp /Fastqs/"$location"/"$sample"*1_paired.fastq.gz /Mapping_pESI/"$sample"/"
>>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "cp /Fastqs/"$location"/"$sample"*2_paired.fastq.gz /Mapping_pESI/"$sample"/"
>>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "chmod 755 /Mapping_pESI/"$sample"/*fastq.gz" >>/pESI_with_SMALT/mapping_bsubs/$sample-
mapping.bsub
echo "smalt map -r 5 -o /Mapping_pESI/"$sample"/"$sample".sam
pesi_reference/Mapping_pESI/"$sample"/"$sample"*1_paired.fastq.gz
/Mapping_pESI/"$sample"/"$sample"*2_paired.fastq.gz" >>/pESI_with_SMALT/mapping_bsubs/$sample-
mapping.bsub
echo "samtools sort /Mapping_pESI/"$sample"/"$sample".sam -o /Coverage/"$sample"/"$sample".bam"
>>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "samtools depth -aa /Coverage/"$sample"/"$sample".bam >
/Coverage/"$sample"/"$sample"_raw.txt" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "grep "ASRF" /Coverage/"$sample"/"$sample"_raw.txt > /Coverage/"$sample"/"$sample"_asrf.txt"
>>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "awk '{print \$1\":\"\$2\"\t\"\$3}' /Coverage/"$sample"/"$sample"_asrf.txt >
/Coverage/"$sample"/"$sample"_awk.txt" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "echo -e '"Base "\t" "$sample""' | cat - /Coverage/"$sample"/"$sample"_awk.txt >
/Results/"$location"_pesi_coverage.txt" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "rm -r -f /Mapping_pESI/"$sample"" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "rm -r -f /Mapping_pESI/"$sample"" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
echo "rm -f -r /Coverage/"$sample"" >>/pESI_with_SMALT/mapping_bsubs/$sample-mapping.bsub
```

I.VI Transposing the coverage matrices

The following script to transpose a large text file was found on a coding forum (Hægland, 2013).

```
{
   for (i=1; i<=NF; i++) a[NR,i]=$i
}
END {
   for (i=1; i<=NF; i++) {
      for (j=1; j<=NR; j++) {
         printf "%s", a[j,i]
         if (j<NR) printf "%s", OFS
      }
      printf "%s",ORS
   }
}
```

I.VII Generating R scripts to make the heatmaps

This script generates R scripts for each coverage matrix (R Core Team, 2018). A list of the isolates in the matrix is inputted with R to generate a phylogeny. The coverage matrix is then imported into R using data.table (Dowle and Srinivasan, 2018). Every value in the matrix above 20 is converted to 20 and below 20 to 0. The heatmap is then generated using phylotools, exported as a tiff file (Revell, 2012).

```
#!/bin/bash

sample=$1

echo "library(ape, lib.loc=\”/R/x86_64-redhat-linux-gnu-library/3.5\")" >>/R_workspace/scripts/$sample.R
echo "library(phytools, lib.loc=\”/R/x86_64-redhat-linux-gnu-library/3.5\")"
>>/R_workspace/scripts/$sample.R
echo "library(data.table, lib.loc=\”/R/x86_64-redhat-linux-gnu-library/3.5\")"
>>/R_workspace/scripts/$sample.R
echo "" >>/R_workspace/scripts/$sample.R
echo "tree_"$sample" <- read.tree(file = \"pESI_tree_lists/"$sample"_tree\")"
>>/R_workspace/scripts/$sample.R
echo "data_"$sample" <- fread(\"pESI_heatmap_lists/transposed/"$sample"_transposed\", header = TRUE,
check.names = FALSE)" >>/R_workspace/scripts/$sample.R
echo "matrix_"$sample" <- as.matrix(data_"$sample", rownames = 1)" >>/R_workspace/scripts/$sample.R
echo "matrix_"$sample"[matrix_"$sample"<20] = 0" >>/R_workspace/scripts/$sample.R
echo "matrix_"$sample"[matrix_"$sample">20] = 20" >>/R_workspace/scripts/$sample.R
echo "colours<-colorRampPalette(colors=c(\"blue\",\"red\"))(100)" >>/R_workspace/scripts/$sample.R
echo "tiff("\"$sample"_heatmap.tiff\", height=8.27, width=11.69, units='in', res=600)"
>>/R_workspace/scripts/$sample.R
echo "phylo.heatmap(tree_"$sample", matrix_"$sample",scale=0.2, fsize=0.1, pts = FALSE, colors=colours,
legend = FALSE, split = c(0.25, 0.75))" >>/R_workspace/scripts/$sample.R
echo "dev.off()" >>/R_workspace/scripts/$sample.R
```

I.VIII Identifying whether *bla*$_{\text{CTX-M-65}}$ is on pESI

This script creates bsub files for each sequence to then be submitted to a high performance cluster. Each assembly is nucleotide blasted against CP016407 with the subject and query ID's and start and ends outputted (McGinnis and Madden, 2004). The blast results are then filtered, excluding hits that don't cover *bla*$_{\text{CTX-M-65}}$.

```
#!/bin/bash

location=$1
sample=${location#*\/}


echo "CREATE bsub scripts"
echo $sample

echo "#!/bin/sh" >>$sample-esbl.bsub
echo "#BSUB -q short-eth" >>$sample-esbl.bsub
echo "#BSUB -R "rusage[mem=1000]"" >>$sample-esbl.bsub
echo "#BSUB -M 1000" >>$sample-esbl.bsub
echo "#BSUB -J esbl" >>$sample-esbl.bsub
echo "#BSUB -oo esbl"$sample".out" >>$sample-esbl.bsub
echo "#BSUB -eo esbl"$sample".err" >>$sample-esbl.bsub
echo "" >>$sample-esbl.bsub
echo ". /etc/profile" >>$sample-esbl.bsub
echo "module add ncbi-blast/2.9.0+/gcc" >>$sample-esbl.bsub
echo "" >>$sample-esbl.bsub
echo "mkdir /Mapping_pESI/"$sample"/" >>$sample-esbl.bsub
echo "cp /Assemblies/"$location"/"$sample"_scaffolds.fasta /Mapping_pESI/"$sample"/" >>$sample-
esbl.bsub
echo "chmod 755 /Mapping_pESI/"$sample"/*" >>$sample-esbl.bsub
echo "blastn -query /Mapping_pESI/"$sample"/"$sample"_scaffolds.fasta -db eBG31_Tate_pESI.fa -task
megablast -outfmt '10 sseqid qseqid length qstart qend sstart send' | grep 'CP016407.1' >
/Mapping_pESI/"$sample"/"$sample"_blast.txt" >>$sample-esbl.bsub
echo "awk -F "," '(NR>1) && ($6 < 279352 ) ' /Mapping_pESI/"$sample"/"$sample"_blast.txt >
/Mapping_pESI/"$sample"/"$sample"_l6.txt" >>$sample-esbl.bsub
echo "awk -F "," '(NR>1) && ($7 > 280227 ) ' /Mapping_pESI/"$sample"/"$sample"_l6.txt >
/Mapping_pESI/"$sample"/"$sample"_h7.txt" >>$sample-esbl.bsub
echo "awk -F "," '(NR>1) && ($7 < 279352 ) ' /Mapping_pESI/"$sample"/"$sample"_blast.txt >
/Mapping_pESI/"$sample"/"$sample"_l7.txt" >>$sample-esbl.bsub
echo "awk -F "," '(NR>1) && ($6 > 280227 ) ' /Mapping_pESI/"$sample"/"$sample"_l7.txt >
/Mapping_pESI/"$sample"/"$sample"_h6.txt" >>$sample-esbl.bsub
echo "cat /Mapping_pESI/"$sample"/"$sample"_h7.txt /Mapping_pESI/"$sample"/"$sample"_h6.txt >
/ESBL_Containing/Results/"$sample"_results.txt" >>$sample-esbl.bsub
echo "rm -rf /Mapping_pESI/"$sample"/" >>$sample-esbl.bsub
```

I.IX Extracting the integron finder results

This script creates a bash script for each of the sequences which extracts the column with the title ‘Complete’ from the Integron Finder summary output (Cury *et al.*, 2016a). It then counts the occurrences of 1’s (Complete integron) and 0’s (Incomplete integron) and output these into a results file.

```
#!/bin/bash

location=$1
sample=${location#*\/}

echo "awk -v header="\$\{1:-complete\}" '" >> /Integron_Finder/extraction_scripts/"$sample".sh
echo "BEGIN { FS=\" \"; c=0 }" >> /Integron_Finder/extraction_scripts/"$sample".sh
echo "NR == 1 { for (i=1;i<=NF;i++) { if (\$i==header) { c=i }} }" >>
/Integron_Finder/extraction_scripts/"$sample".sh
echo "NR > 1 && c>0 { print \$c }" >> /Integron_Finder/extraction_scripts/"$sample".sh
echo "' /Integron_Finder/"$location"/"$sample"_scaffolds.summary > "$sample"_tmp1" >>
/Integron_Finder/extraction_scripts/"$sample".sh
echo "grep -c 0 "$sample"_tmp1 > "$sample"_Incomplete" >>
/Integron_Finder/extraction_scripts/"$sample".sh
echo "grep -c 1 "$sample"_tmp1 > "$sample"_Complete" >>
/Integron_Finder/extraction_scripts/"$sample".sh
echo "sed -i -e 's/^/"$sample": /' "$sample"_Incomplete" >>
/Integron_Finder/extraction_scripts/"$sample".sh
echo "sed -i -e 's/^/"$sample": /' "$sample"_Complete" >>
/Integron_Finder/extraction_scripts/"$sample".sh
echo "rm -f "$sample"_tmp1" >> /Integron_Finder/extraction_scripts/"$sample".sh
```

II Appendix II. Supplementary Data for Chapter 2

Source Group	*Keyword*
Environmental	Air, Almonds, Animal, Animal feed, Animal-calf-formula-fed veal, Animal-cattle-beef cow, Animal-cattle-dairy cow, Animal-cattle-heifer, Animal-cattle-steer, Animal-swine-market swine, Animal-swine-roaster swine, Animal-swine-sow, Aquatic, Aquatic animal, Avian, Avian (carcass), Bacon flavored bones, Basil, Bearded dragon feces, Beef chicken soy composite, Beef cutlets, Beef meat, Biological tissue and/or fluid -sus scrofa domesticus, Black pepper, Blood meal, Boneless pork, Bovine, Bovine (carcass), Bovine (feces), Bovine (feed-grain pellets), Bovine (feed), Bovine (intestine), Bovine (necropsy-mesenteric lymph node), Bovine (necropsy), Bovine feces, Bovine kidney (bos taurus), Bovine liver, Bovine tissue, Brewer's yeast, Bulk feed, Camel (feces), Camel (si), Camel spleen, Camelid, Canine, Canine (bal), Canine (feces), Canine (necropsy), Canine feces (canis lupus familiaris), Canine intestine (canis lupus familiaris), Cat's claw powder (herbal product), Catfish meal, Cattle, Cattle feed, Chicken cage, Chocolate candy pieces, Cilantro, Comminuted beef, Comminuted or otherwise nonintact-pork, Companion animal, Composite food, Compost, Cordon bleu, Coriander, Crawfish, Creek water, Cucumbers, Dairy, Distilled corn, Dog, Dog food, Dog treat, Drag swab, Drag swab through chicken house (gallus gallus domesticus), Dried parsley, Dried pork ears, Dry dog food, Duck meats for dogs, Environment, Environmental samples, Environmental sponge, Environmental swab, Environmental swabs, Environmental_other, Eq_horse, EQAD, EQAD UKNEQAS, Equine, Equine feces, Equus caballus feces, Farmed fresh water shrimp, Feces (bos taurus), Feed, Feeding stuffs, Feline, Feline intestine (felis catus), Fines, Finished feed, Finished pet food, Fish, Fish water (pisces), Food, Food investigation, Food source unknown, Fresh cheese, Fried charal (sardines), Frozen baby clam boiled, Frozen cut crab, Frozen iqf scallops adductors, Frozen lobster tails, Frz calamari rings Caprine (feces), Carcass swab, Cat, Cat food, Frz shrimp, Goat, Ground beef, Ground pork, Hog feed, Hogs, Horse, Invertebrates, IQA, Jalapeno peppers, Lemon grass tea, Lettuce, Livestock, Lizard, Mdh garam masala, Meat, Meat & bone meal, Meat and bone meal, Meat feed, Meat/nonmeat combination-combination species, Mouse (necropsy), Natural pig ears, Nonmeat-other, Nutmeg, Offal, Other mammal, Other_veg_mineral, Ovine, Ovine (feces), Papaya, Parmesan cheese, shelf stable grated, Parsley, Pasilla peppers, Pasta, Pecan halves, Peppermint powder, Pet food - raw fresh, Pet food (kibble), Pet food ingredient, Pet foods, Pet treat, Pig, Pig fetus (sus scrofa domesticus), Pig stool, Pig's ears, Placenta (canis lupus familiaris), Plant, Porcine, Porcine (colon), Porcine (necropsy-intestine), Porcine feces(sus scrofa domesticus), Porcine liver, Porcine lung, Porcine lymph node (sus scrofa domesticus), Porcine spleen, Porcine tissue, Pork, Pork carcass, Pork chop, Pork meat, Pork sausage, Potato and meat (bovine), Poultry feed, Produce field - drag swab, Product-raw-ground, Product-raw-intact-beef, Product-raw-intact-pork, Product-raw-intact-siluriformes, Product-RTE-fully cooked, Product-swab-pork, Protein dairy feed, Pumpkin seed powder, Raw almonds, Raw almonds/nuts, Raw bone, Raw kale-red, Raw meat, Raw pork, Reptile, Rinse water, Rodent, Scallops, Seagull, Sediment, Sewage, Shelf stavble grated parmesan and romano cheese blend in a glass jar with a metal lid., Shellfish, Shinisaurus crocodilurus, Shrimp, Snails, Soil/dust, Soy lecithin fluid, Spinach, Sus scrofa, Sus scrofa domesticus, Swab, Sweet basil (dried), Swine, UK_GBRUSAL, Walnuts, Water, White pepper powder, Wild animal, Za'atar
Human	Anthropogenic, Ascitic fluid, Aspirate, Blood, Blood culture, Clinical, Clinical sample, CSF, CSF & stool, Faeces, Homo sapiens, Homo sapiens, Human, Human blood, Mesoscopic, Organic, Pus, Rectal swab, Stool, Stool, Swab superficial, Tissue, Urine, Wound, Wound drain, Wound swab
Poultry	Animal-Chicken-Young Chicken, Animal-Turkey-Turkey Carcass Sponge, Avian, Boneless skinless chicken breast, Broiler carcass, Broiler chicken, Cajun Chicken (Raw Meat), Chicken, Chicken breast, Chicken by-product, Chicken carcass, Chicken carcass rinse, Chicken cecum pre-harvest (Poultry), Chicken drag swab, Chicken kiev, Chicken leg, Chicken livers, Chicken meat, Chicken mince, Chicken survey, Chicken Wing, Comminuted Chicken, Comminuted Turkey, Dried egg powder, Duck, Egg yolks, Frozen raw whole chicken, Gallus gallus domesticus, Ground turkey, Guinea fowl, NRTE (Not-Ready-to-Eat) Comminuted Poultry Exploratory Sampling - Chickens, Osborne eggs, Poultry rinse, product-eggs-raw-whole, product-eggs-raw-yolks, Quail, Raw chicken, Raw chicken escalopes, Raw chicken kebab meat, Raw Egg Shell, Raw Intact Chicken, Raw poultry (chicken), Raw Shell Eggs, Retail Chicken Quarter Leg (Poultry), Slaughter chicken ceca, Slaughter comminuted chicken, Thai chicken mix, turkey, Young chicken rinse

Table II.1 Keywords used to stratify into source groups

All of the keywords associated with the *S.* Infantis strains and the source group they belonged to.

Source Group	*Keyword*
Africa	African Continent, Cameroon, Cote d'Ivoire, Egypt, Ethiopia, Gambia, Ghana, Kenya, Morocco, Nigeria, South Africa, Tunisia, Uganda, United Republic of Tanzania
Asia	Afghanistan, Asian Continent, Bangladesh, China, India, Indonesia, Iran, Iraq, Israel, Japan, Pakistan, Saudi Arabia, Thailand, Turkey, United Arab Emirates
Europe	Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, European Continent, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Luxembourg, Malta, Norway, Poland, Portugal, Romania, Slovakia, Spain, Ukraine, United Kingdom
North America	Canada, Caribbean, Costa Rica, Cuba, El Salvador, Jamaica, Mexico, Panama, Saint Vincent and the Grenadines, United States
South America	Bolivia, Brazil, Chile, Colombia, Costa Rica, Ecuador, Guyana, Peru, South American Continent, Suriname

Table II.2 Countries and continents *S.* Infantis was isolated from

The countries, and keywords defining continents, belonging to each continent that *S.* Infantis strains were isolated from.

Source Group	*Keyword*
Faeces or rectal swabs	Faeces, Stool, Rectal Swab
Blood	Blood, Blood culture
Urine	Urine
Chickens	Chickens, CHICKEN SURVEY, Young chicken rinse, Animal-Chicken-Young Chicken, Gallus gallus domesticus, Chicken, broiler chicken, Chicken drag swab
Chicken meat	Chicken leg, Chicken wing, Raw chicken, Raw Chicken Escalopes, Cajun Chicken (Raw Meat), THAI CHICKEN MIX, RAW POULTRY(CHICKEN), Raw Poultry, FROZEN RAW WHOLE CHICKEN, CHICKEN MINCE, RAW CHICKEN KEBAB MEAT, Chicken Wing, Chicken Breast, Chicken Breast, Boneless skinless chicken breast, Chicken Kiev, Chicken breasts, chicken meat, Chicken Wings, Comminuted Chicken, Chicken Legs, Raw Intact Chicken, Slaughter, chicken ceca, Slaughter, comminuted chicken, chicken by-product, Comminuted Turkey, Chicken Carcass, Chicken carcass rinse, Broiler carcass, Chicken livers, Chicken Cecum Pre-Harvest (Poultry), Retail Chicken Quarter Leg (Poultry), Poultry rinse
Eggs	Raw Egg Shell, OSBORNE EGGS, Raw Shell Eggs, egg yolks, product-eggs-raw-whole, dried egg powder, product-eggs-raw-yolks
Duck	Duck
Turkey	Ground Turkey, turkey, Animal-Turkey-Turkey Carcass Sponge,
Quail	Quail
Pigs	PIG, Pork Chops, Animal-Swine-Sow, pork chop, porcine spleen, raw pork, Swine, Porcine Liver, Porcine Tissue, Porcine (necropsy-intestine), Porcine (colon), Porcine, Porcine Lung, Pork Carcass, Animal-Swine-Roaster Swine, Animal-Swine-Market Swine, Hogs, Product-Swab-Pork, Pig Fetus (Sus scrofa domesticus), Product-Raw-Intact-Pork, Comminuted or Otherwise Nonintact-Pork, GROUND PORK, Product-Raw-Ground, biological tissue and/or fluid -Sus scrofa domesticus, Porcine feces(Sus scrofa domesticus), Boneless Pork, Pork Meat, Pork, Sus scrofa domesticus, Porcine Lymph node (Sus scrofa domesticus), Sus scrofa, pork sausage, Pig stool
Cattle	Animal-Cattle-Dairy Cow, Ground Beef, Animal-Cattle-Beef Cow, Product-Raw-Intact-Beef, Bovine (carcass), Bovine, Bovine (feces), Bovine (necropsy), Bovine (feed), Bovine (intestine), Bovine Tissue, Bovine Liver, Bovine (feed-grain pellets), Bovine Feces, Comminuted Beef, Animal-Calf-Formula-fed Veal, Animal-Cattle-Steer, potato and meat (bovine), Feces (bos taurus), Animal-Cattle-Heifer, Beef Cutlets, bovine kidney (Bos taurus), Bovine (necropsy-mesenteric lymph node), Beef Meat, CATTLE
Animal feed	dog treat, animal feed, dog food, bacon flavored bones, bulk feed, poultry feed, pet food (kibble), natural pig ears, Dry Dog Food, protein dairy feed, distilled corn, feeding stuffs, pet treat, hog feed, Blood Meal, pig's ears, Dried Pork Ears, cat food, Pet Food Ingredient, Raw Bone, Pet Foods, meat & bone meal, cattle feed, meat and bone meal, Finished Pet Food, Meat Feed, Finished Feed, Pet food - raw fresh, Duck meats for Dogs

Table II.3 Keywords used to stratify into source subgroups

All of the keywords associated with the *S.* Infantis strains and the source subgroup these belonged to.

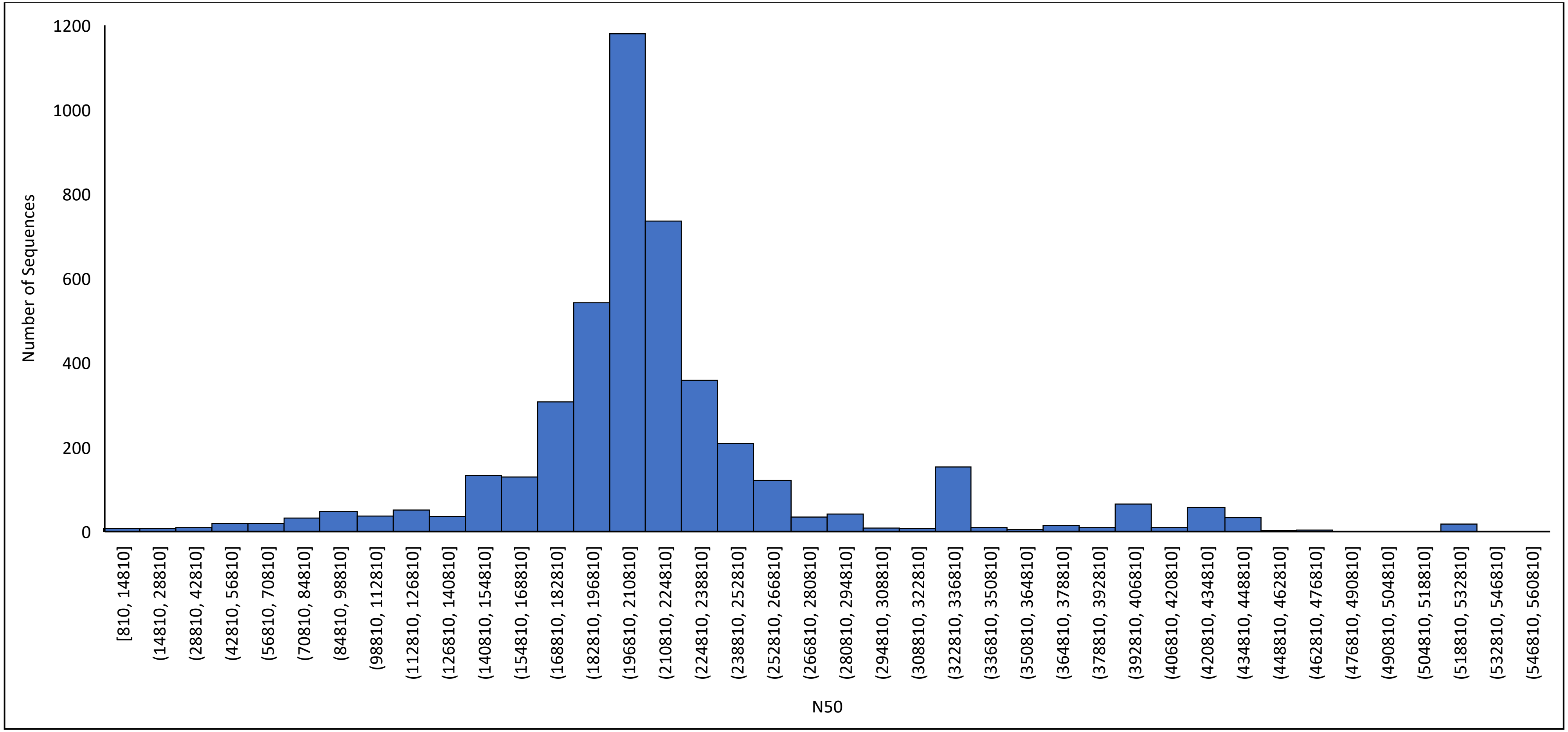


Figure II.1 N50 distribution in the *S.* Infantis genomes
Histogram of the N50 result for each of the first 4,438 genomes included in the project.

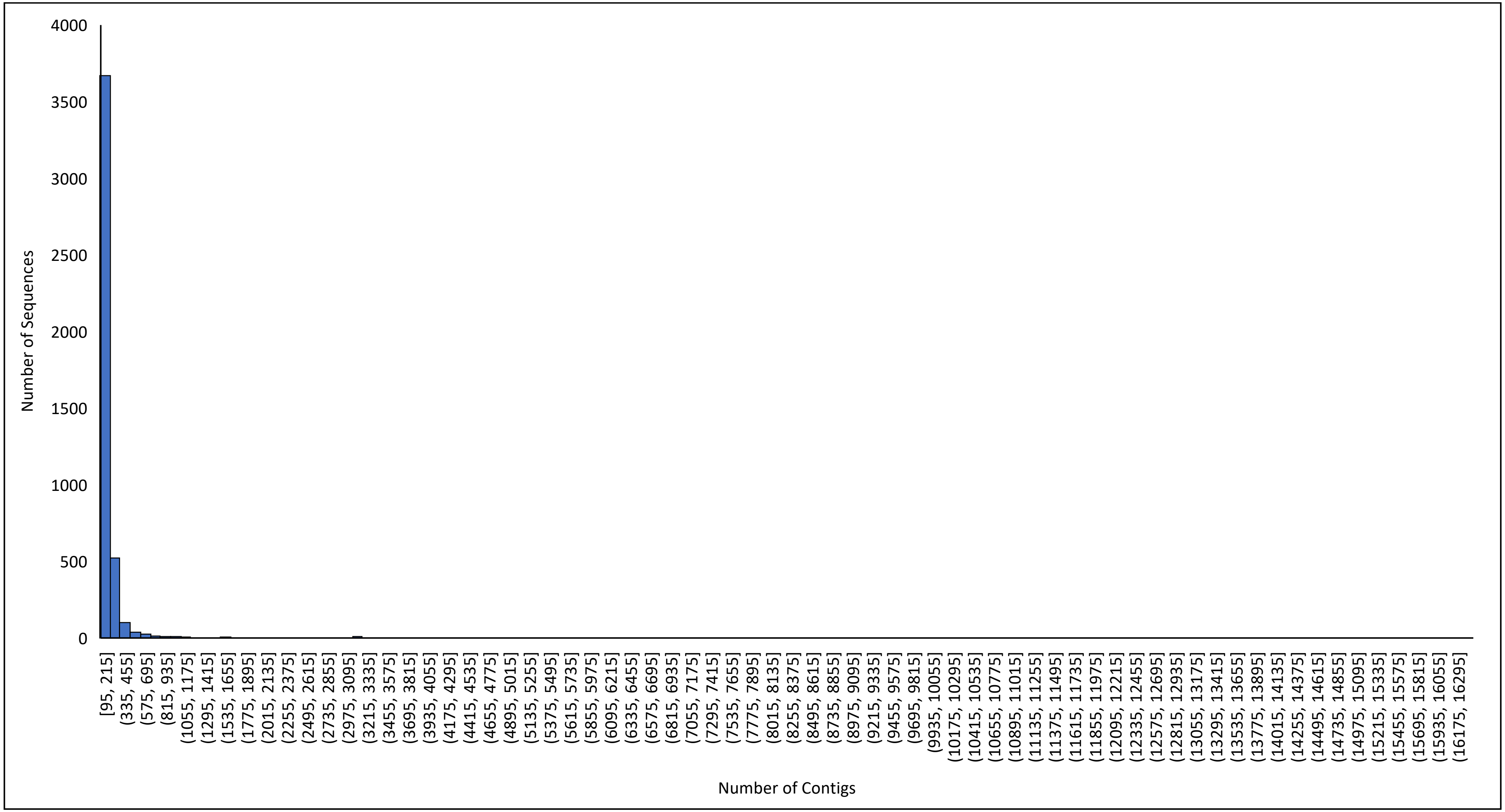


Figure II.2 Contig number distribution in the *S.* Infantis genomes

Histogram of the number of contigs for each of the first 4,438 genomes included in the project.

III Appendix III. Supplementary Data for Chapter 3

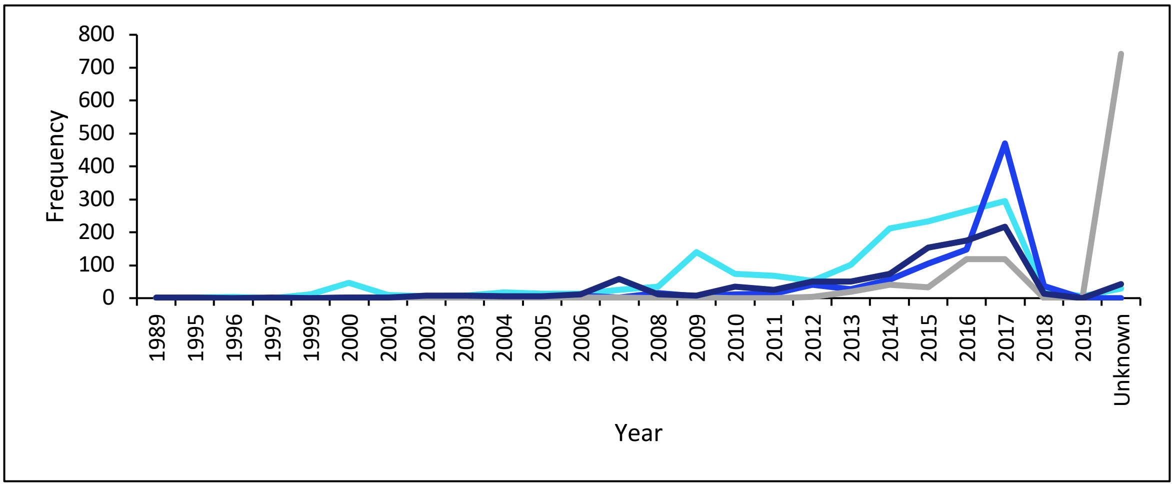


Figure III.1 Distribution of isolation source by year

Environmental n=956, Human n=1687, Poultry n=947, Unknown n=1080.

Environmental ■ Human ■ Poultry ■ Unknown □

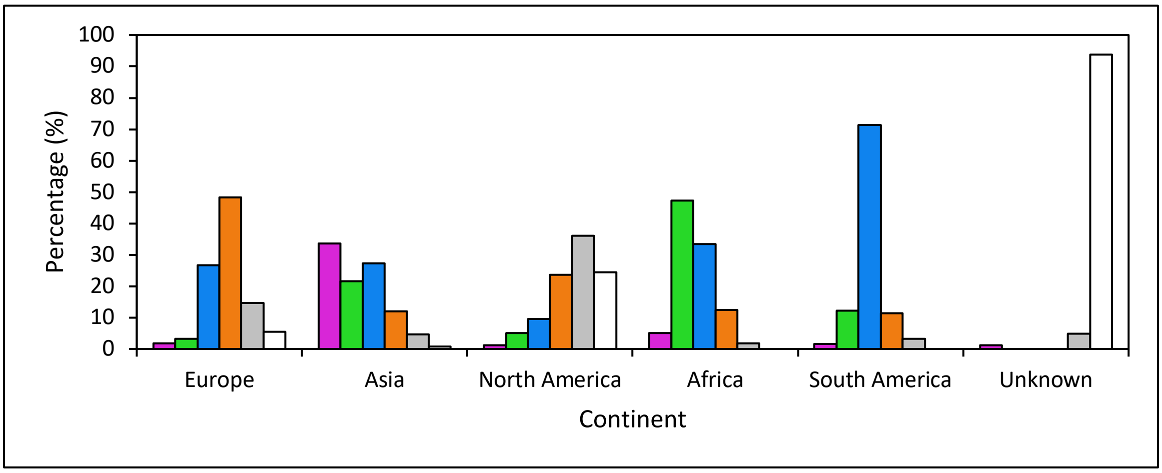


Figure III.2 Year distribution within each continent

Africa n=452, Asia n=241, Europe n=979, N.America n=2795, S.America n=122, Unknown n=81.

1989-2005 🟪 2006-2010 🟩 2011-2014 🟦
2015-2016 🟧 2017-2019 ■ Unknown □

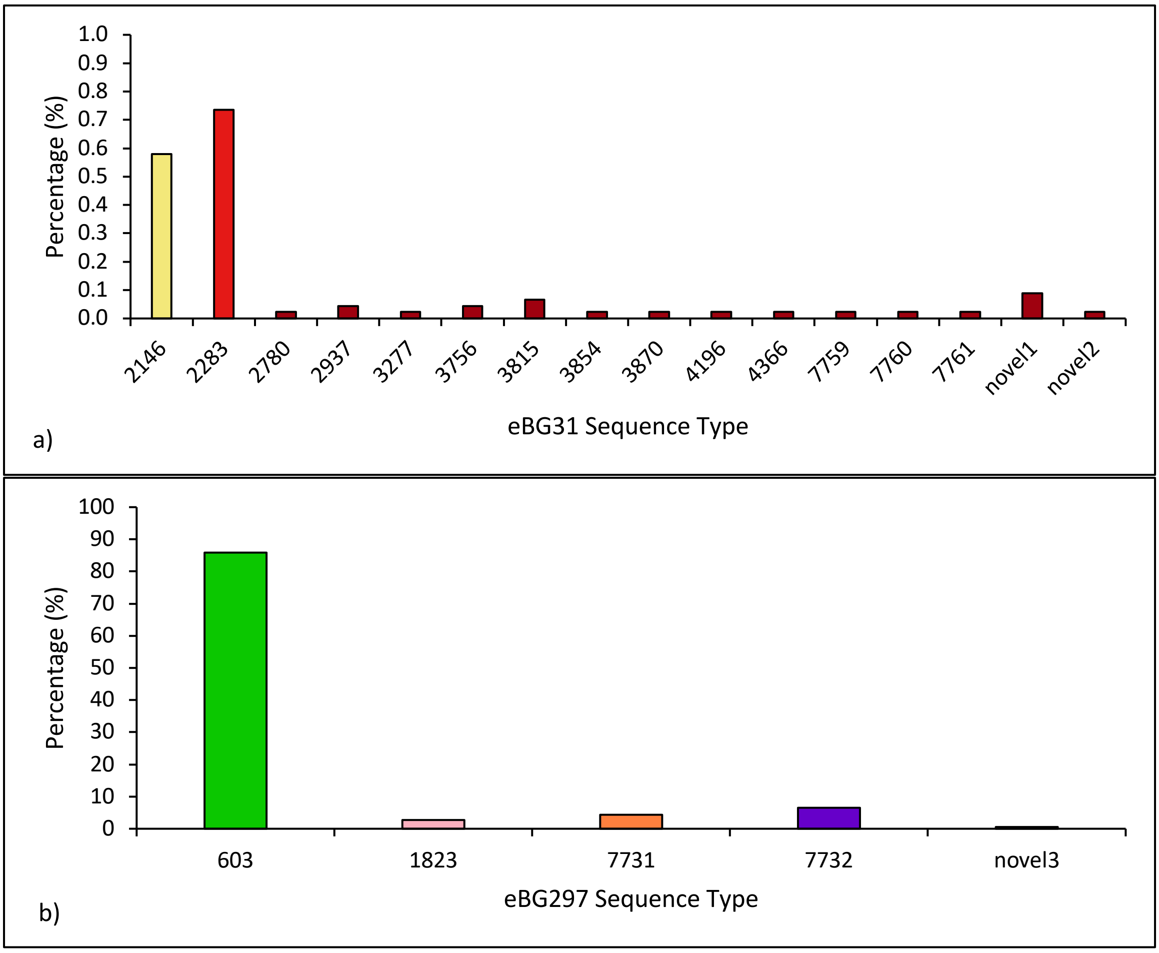


Figure III.3 Distribution of STs in *S.* Infantis

a) STs in eBG31 for n=80 strains, excluding ST32. ST groups: 2146 🟨 2283 🟥 32 🟧 Other 🟫

b) STs in eBG297 (n=184). STs: 603 🟩 1823 ⬜ 7731 🟧 7732 🟪 novel3 🟦

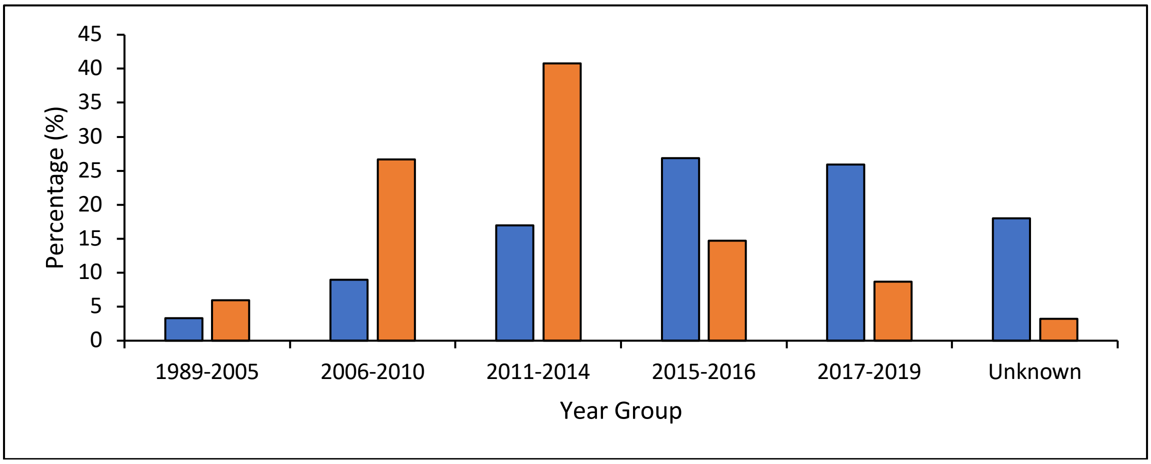


Figure III.4 Distribution of eBG by year

eBG31 n=4486 🟦 eBG297 n=184 🟧

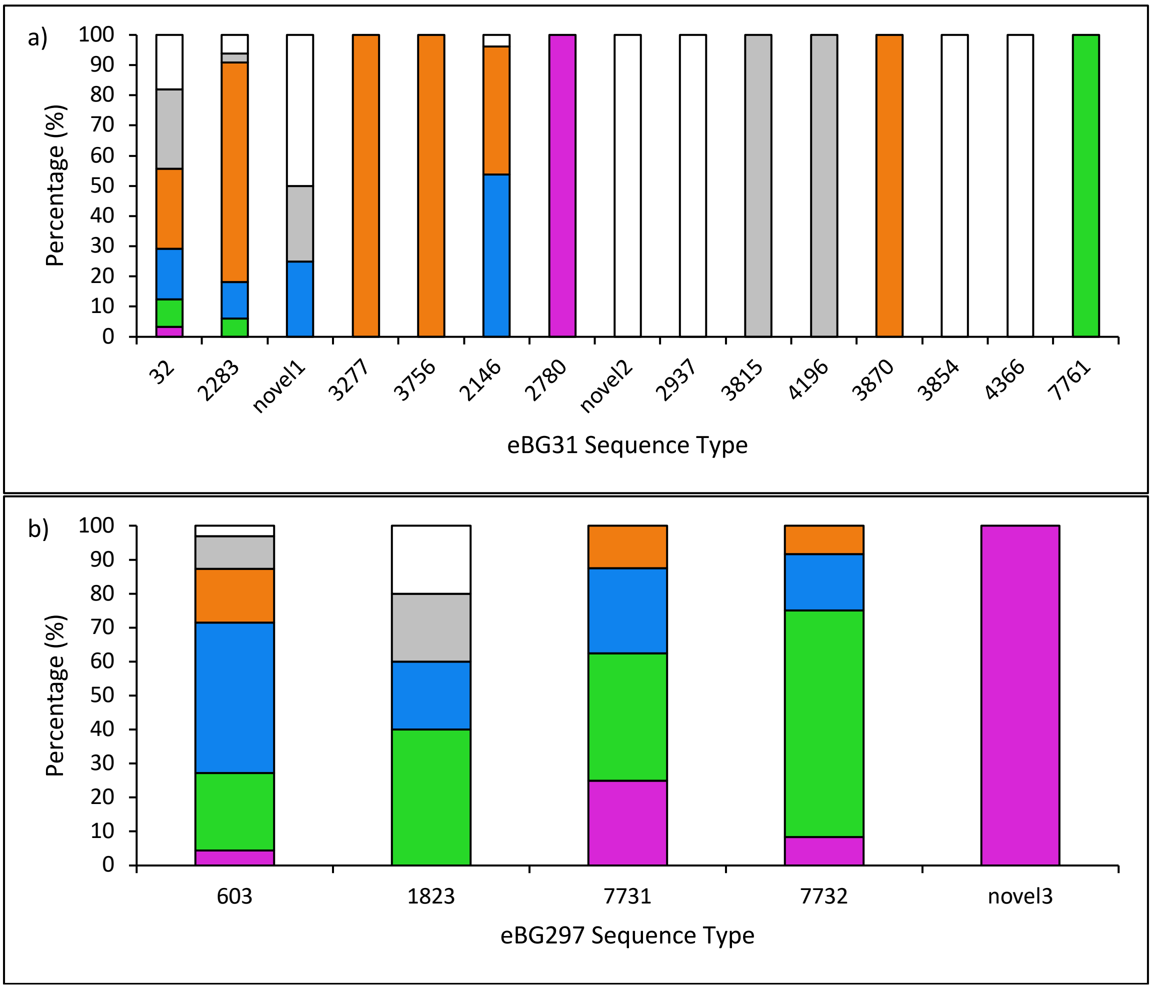


Figure III.5 Year distribution within each *S.* Infantis ST

a) eBG31: ST32 n=4406, ST2283 n=33, ST2146 n=26, ST2937 n=2, ST3756 n=2, ST3815 n=3, novel1 n=4, all other STs n=1

b) eBG297: ST603 n=158, ST1823 n=5, ST7731 n=8, ST7732 n=12, novel3 n=1

1989-2005 🟪 2006-2010 🟩 2011-2014 🟦
2015-2016 🟧 2017-2019 ⬜ Unknown □

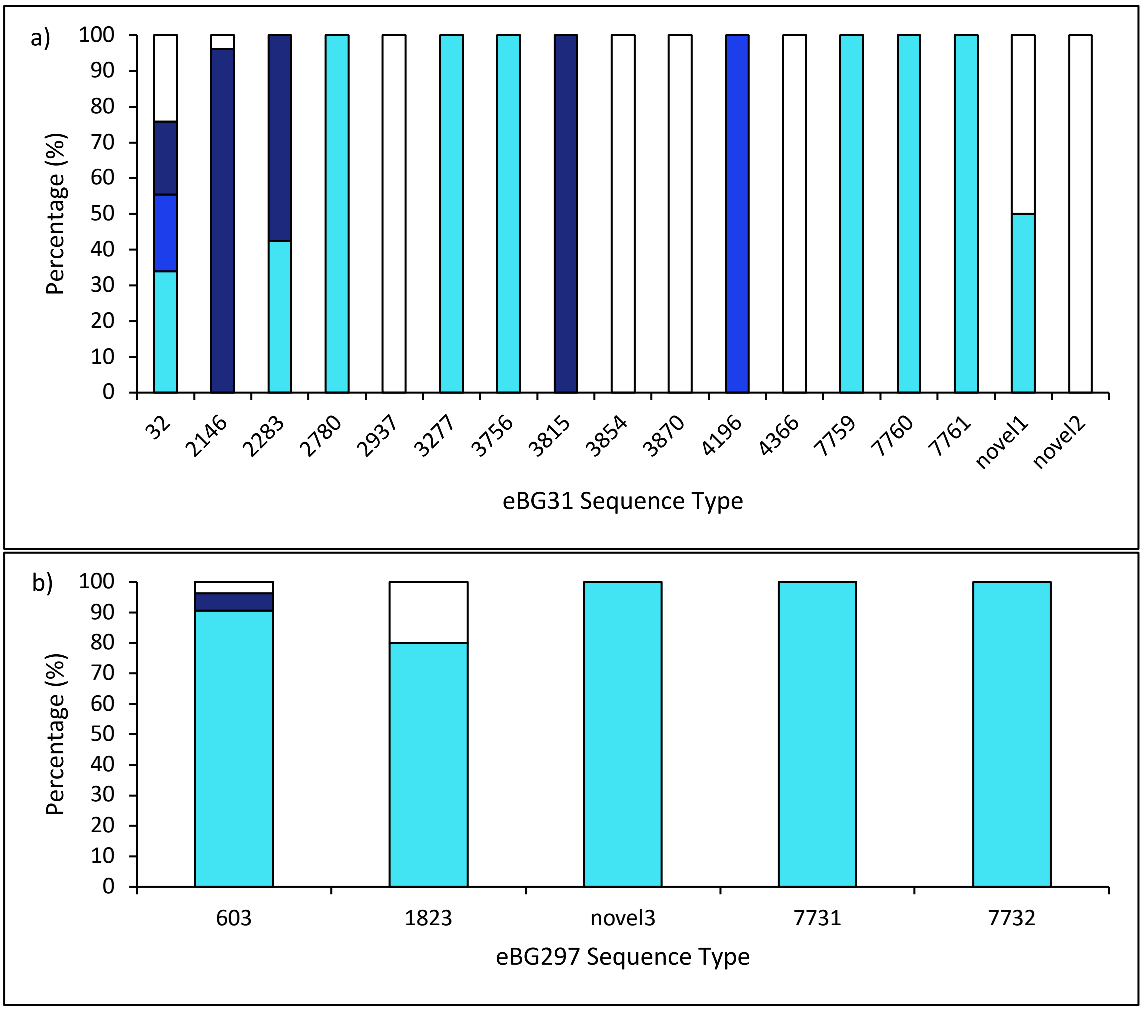


Figure III.6 Source distribution within each *S.* Infantis ST

a) eBG31: ST32 n=4406, ST2283 n=33, ST2146 n=26, ST2937 n=2, ST3756 n=2, ST3815 n=3, novel1 n=4, all other STs n=1

b) eBG297: ST603 n=158, ST1823 n=5, ST7731 n=8, ST7732 n=12, novel3 n=1

Environmental ■ Human ■ Poultry ■ Unknown □

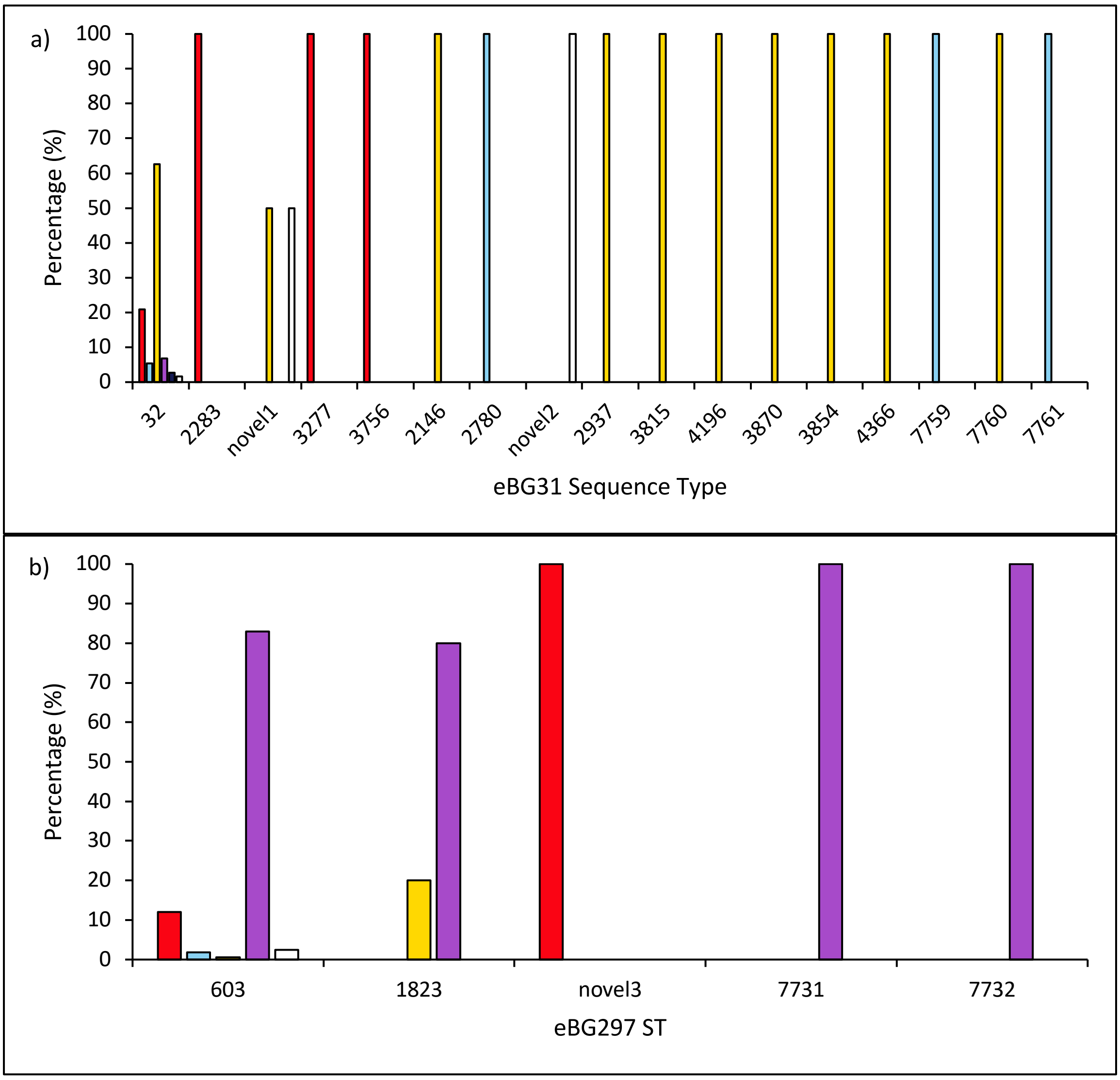


Figure III.7 Continent distribution within each *S.* Infantis ST

a) eBG31: ST32 n=4406, ST2283 n=33, ST2146 n=26, ST2937 n=2, ST3756 n=2, ST3815 n=3, novel1 n=4, all other STs n=1

b) eBG297: ST603 n=158, ST1823 n=5, ST7731 n=8, ST7732 n=12, novel3 n=1

Africa 🟪 Asia 🟦 Europe 🟥 N. America 🟨 S. America ⬛ Unknown □

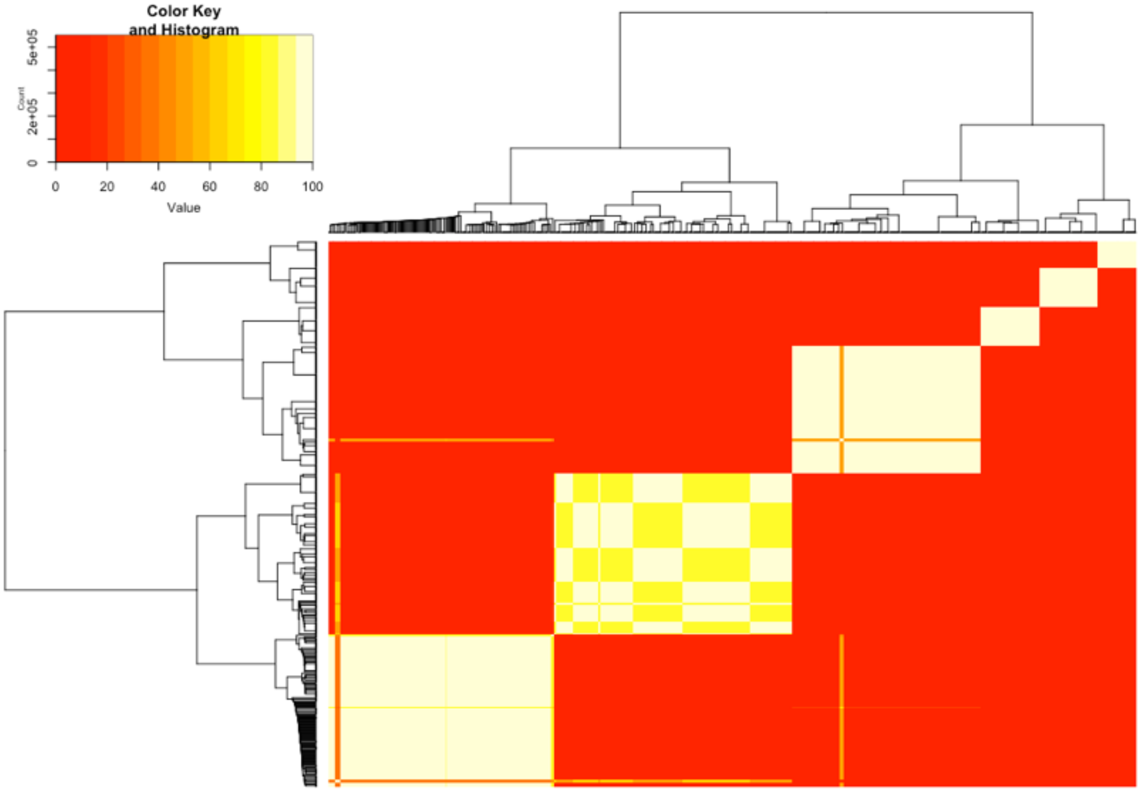


Figure III.8 Heatmap of the bootstrap results for the clustering within the eBG31 alignment
Dendrogram of the fastbaps clusters in the eBG31 soft-core SNP alignment with a heatmap of the bootstrap results. Six clusters were identified in the alignment with high bootstrap results seen between sequences within each fastbaps cluster.

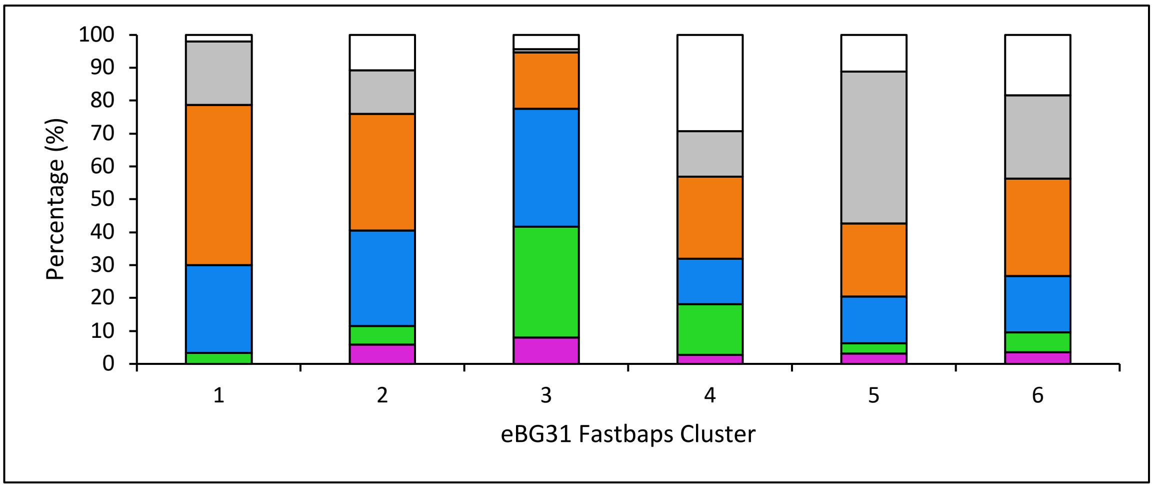


Figure III.9 Year distribution within eBG31 fastbaps clusters
Percentage of isolates from each eBG31 fastbaps cluster that were isolated from each year group.

1989-2005 🟪 2006-2010 🟩 2011-2014 🟦
2015-2016 🟧 2017-2018 ▩ Unknown □

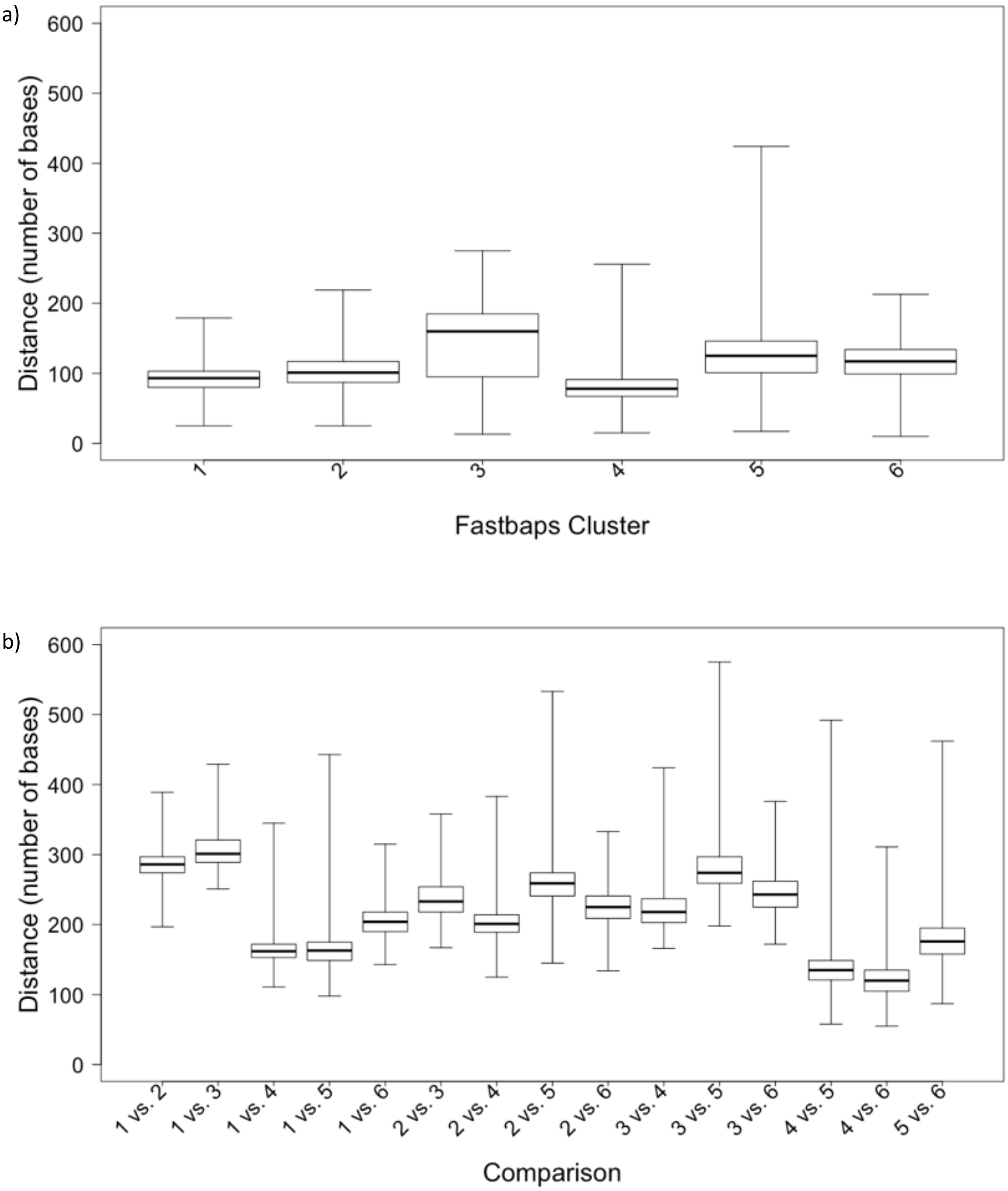


Figure III.10 Median pairwise SNP distribution within and between eBG31 fastbaps clusters
Box and whisker plot showing the minimum, 1st quartile, median, 3rd quartile and maximum pairwise SNP distances within and between each eBG31 fastbaps cluster.

a) Within each eBG31 fastbaps cluster
b) Between each eBG31 fastbaps cluster

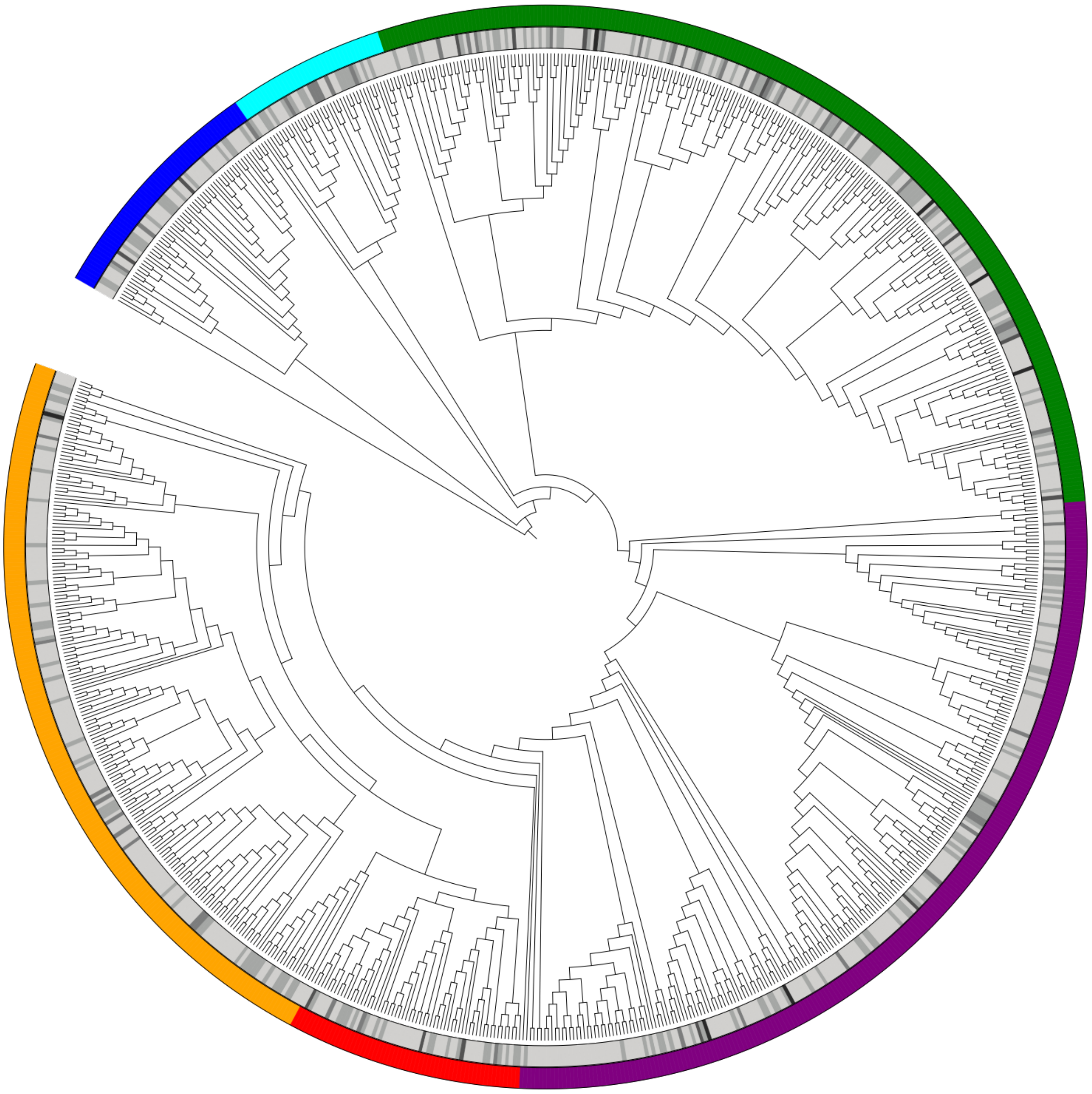


Figure III.11 eBG31 cladogram annotated with fastbaps cluster

Soft-core SNP Maximum Likelihood cladogram of 831 25SNP cluster representatives of eBG31.

Inner ring, Number of sequences in 25SNP cluster: ⬜ 1 ▩ 2-5 ▩ 6-20 ■ 21-50 ■ >50

Outer ring, fastbaps cluster: 1 🟥 2 🟦 3 🟦 4 🟪 5 🟧 6 🟩

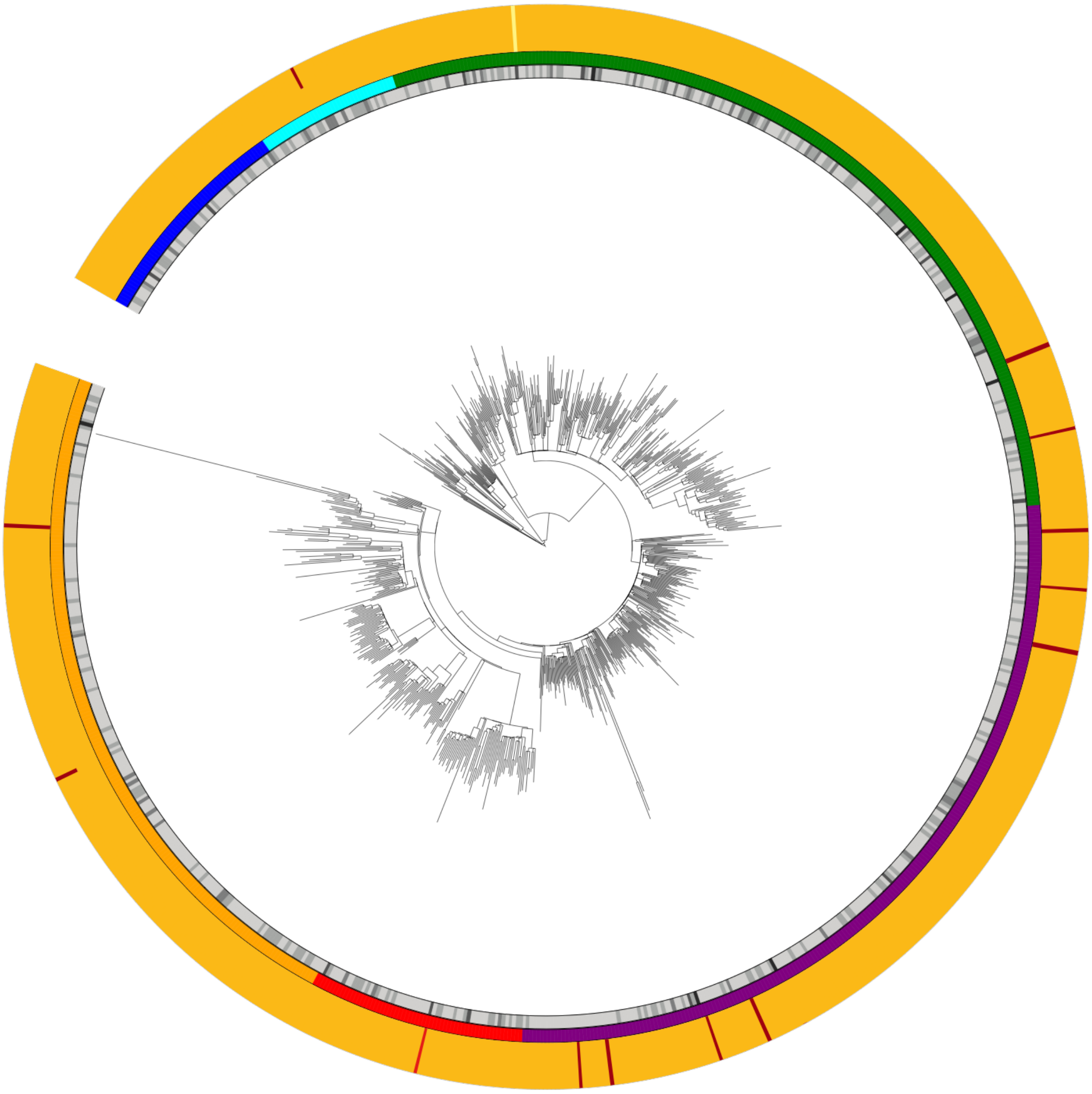


Figure III.12 eBG31 phylogeny annotated with ST

Soft-core SNP Maximum Likelihood Phylogeny of 831 25SNP cluster representatives of eBG31. The outer ring is annotated with the percentage of isolates in each 25SNP cluster that were isolated from each ST group.

Inner ring, Number of sequences in 25SNP cluster: ⬜ 1 ▨ 2-5 ▨ 6-20 ■ 21-50 ■ >50

Middle ring, fastbaps cluster: 1 🟥 2 🟦 3 🟦 4 🟪 5 🟧 6 🟩

Outer ring, ST: 2146 🟨 2283 🟥 32 🟧 Other 🟥

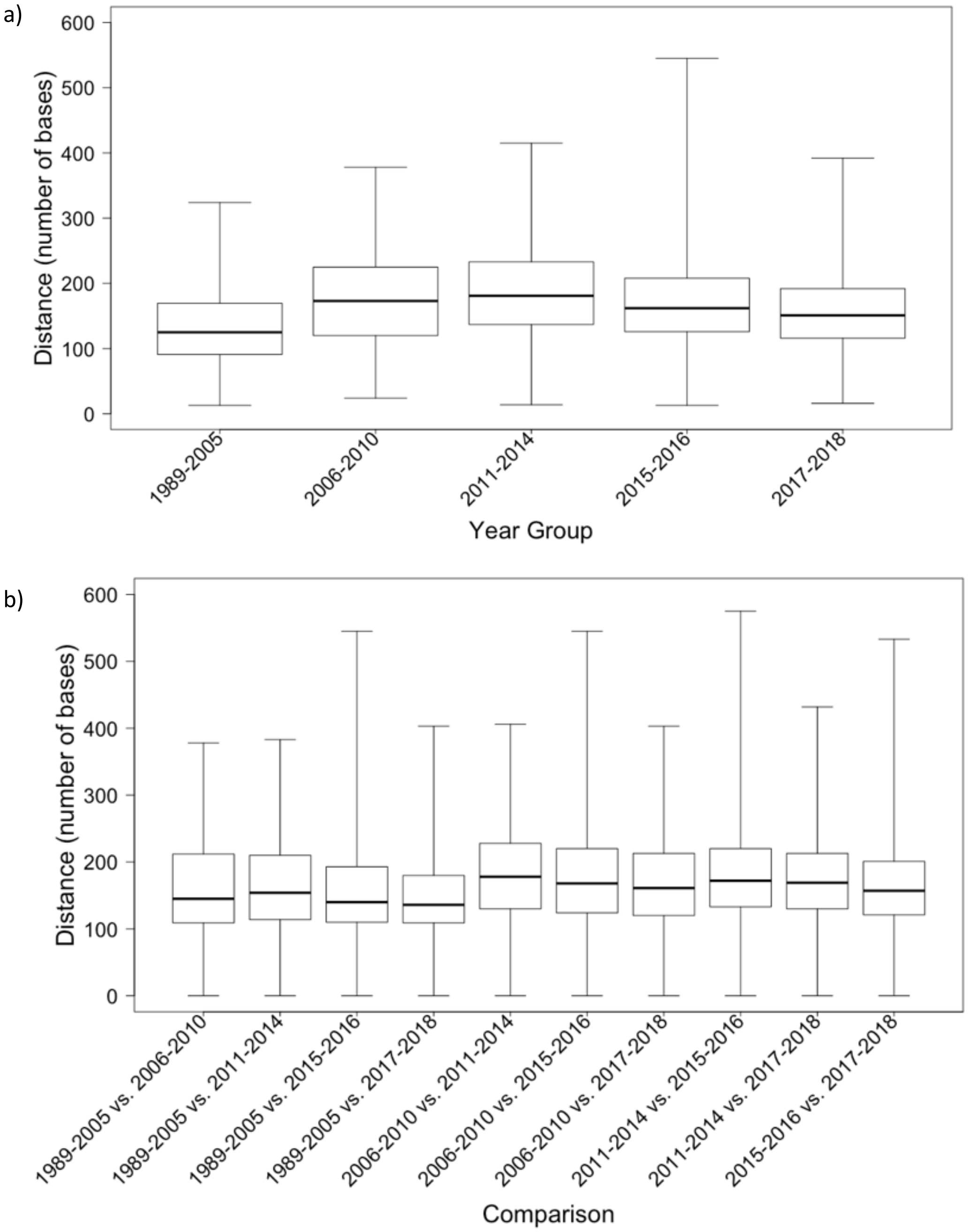


Figure III.13 Distribution of median pairwise SNP distance within and between each eBG31 year group
Box and whisker plot showing the minimum, 1st quartile, median, 3rd quartile and maximum pairwise SNP distances within and between each year group.

a) Within each eBG31 year group
b) Between each eBG31 year group

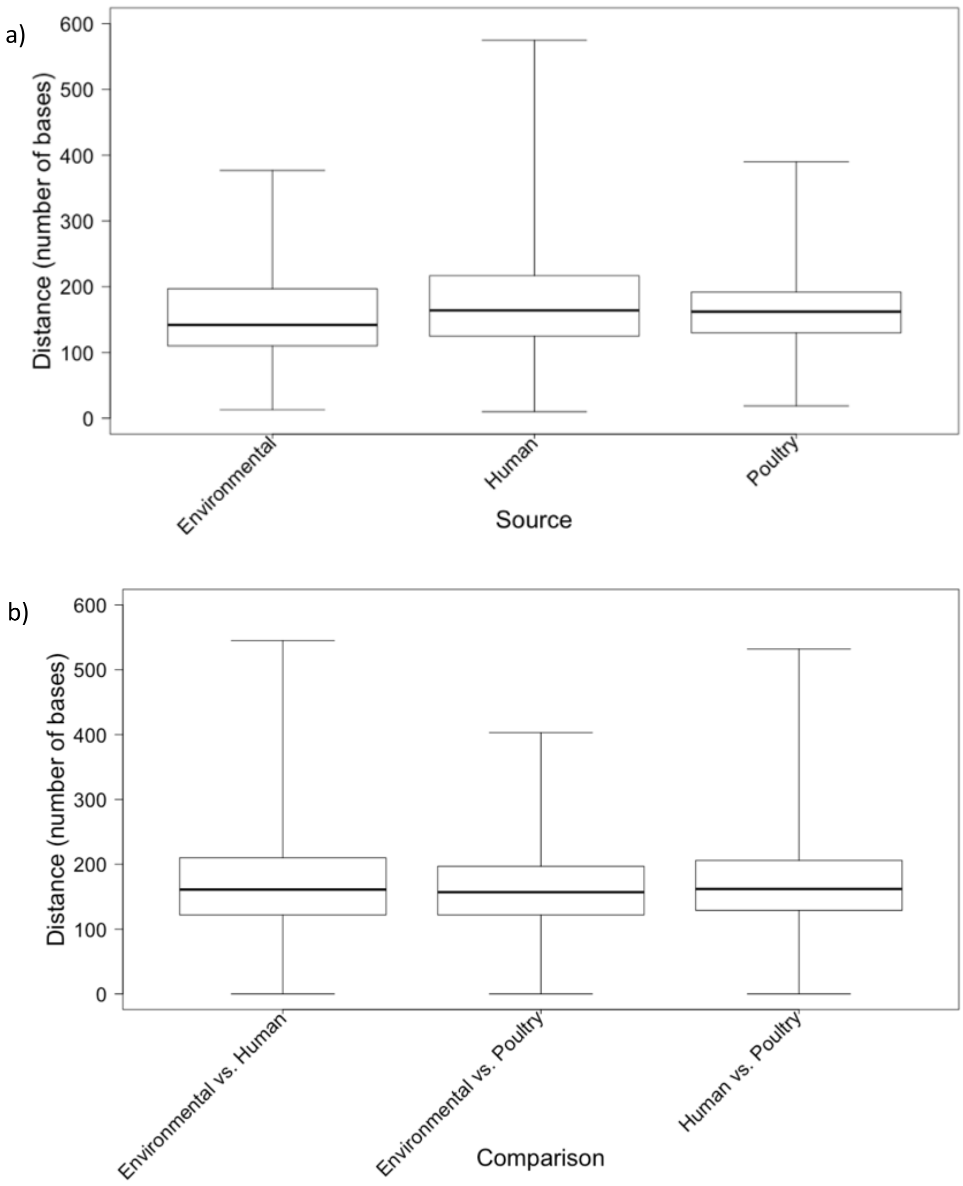


Figure III.14 Distribution of median pairwise SNP distance within and between each eBG31 source group

Box and whisker plot showing the minimum, 1st quartile, median, 3rd quartile and maximum pairwise SNP distances within and between each source.

a) Within each eBG31 source group
b) Between each eBG31 source group

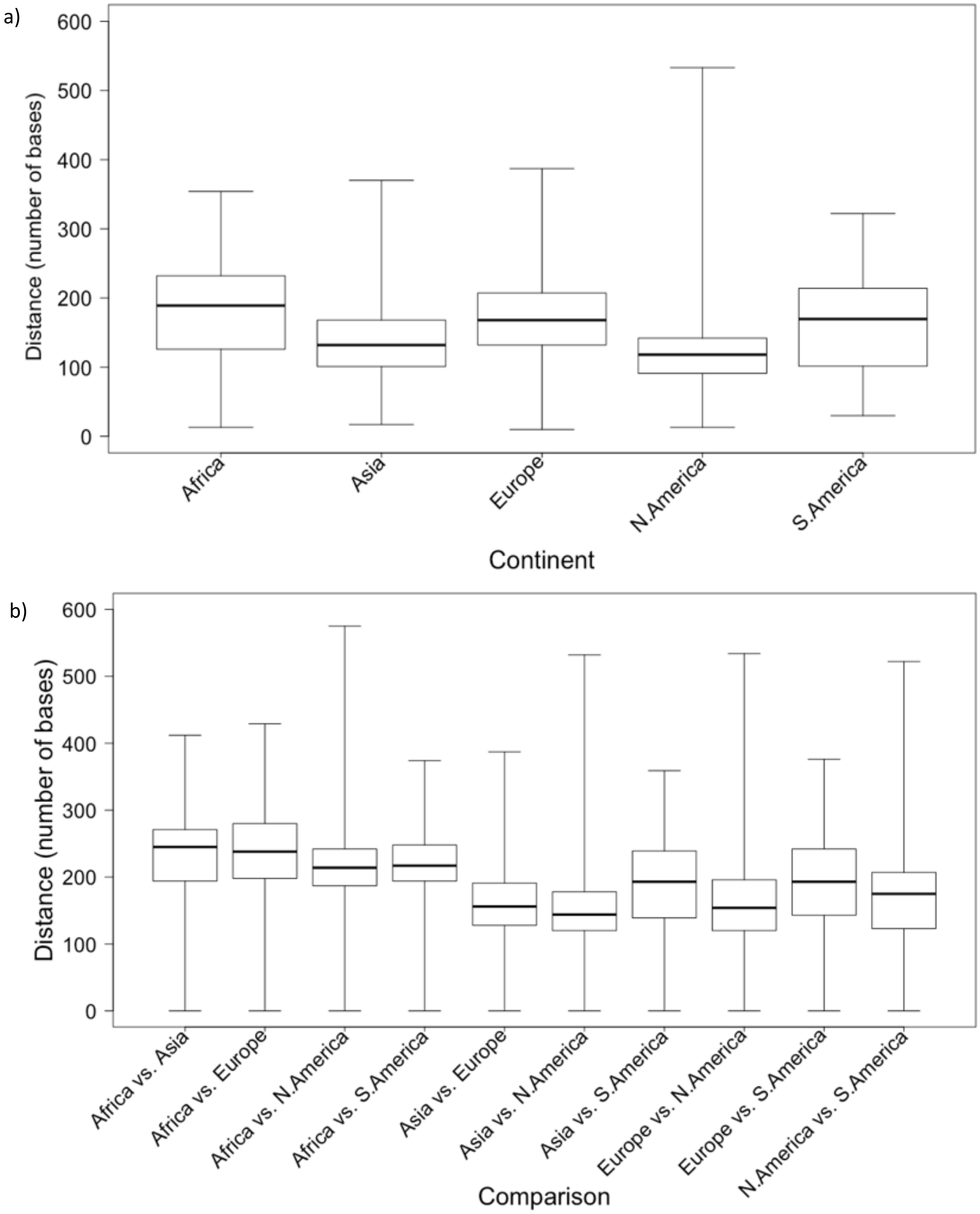


Figure III.15 Distribution of median pairwise SNP distance within and between each eBG31 continent
Box and whisker plot showing the minimum, 1st quartile, median, 3rd quartile and maximum pairwise SNP distances within and between each continent.

a) Within each continent
b) Between each continent

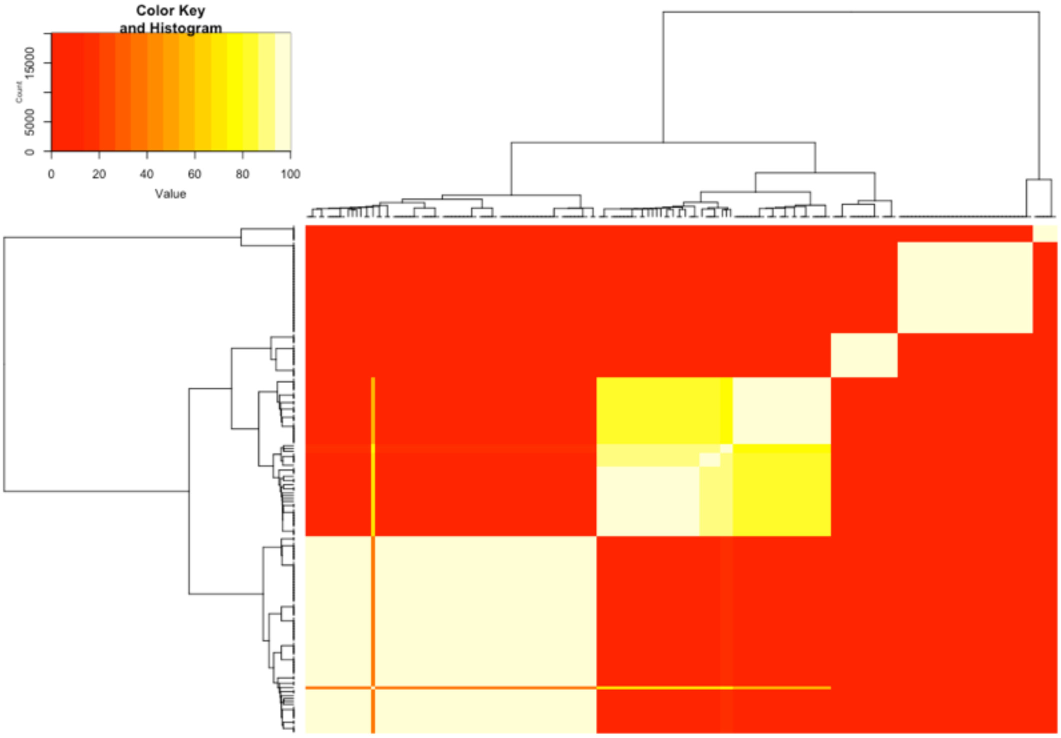


Figure III.16 Heatmap of the bootstrap results for the clustering within the eBG297 alignment
Dendrogram of the fastbaps clusters in the eBG297 soft-core SNP alignment with a heatmap of the bootstrap results. Five clusters were identified in the alignment with high bootstrap results seen between sequences within each fastbaps cluster.

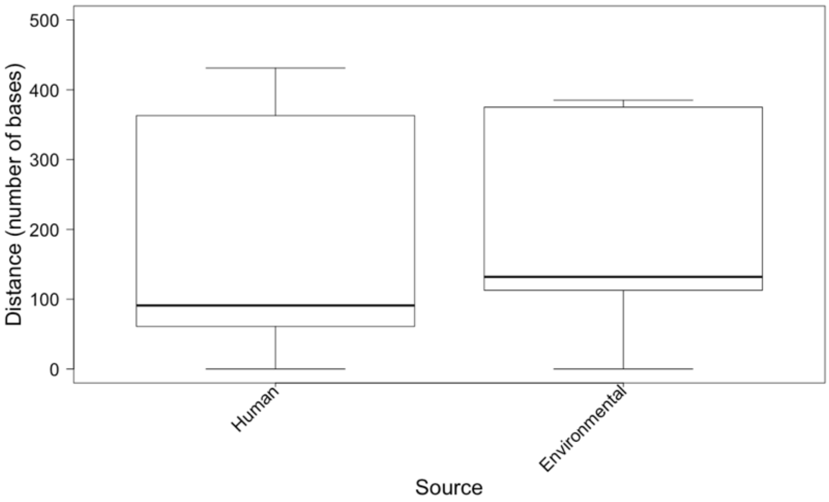


Figure III.17 Distribution of median pairwise SNP distance within each eBG297 source
Box and whisker plot showing the minimum, 1st quartile, median, 3rd quartile and maximum pairwise SNP distances within each source.

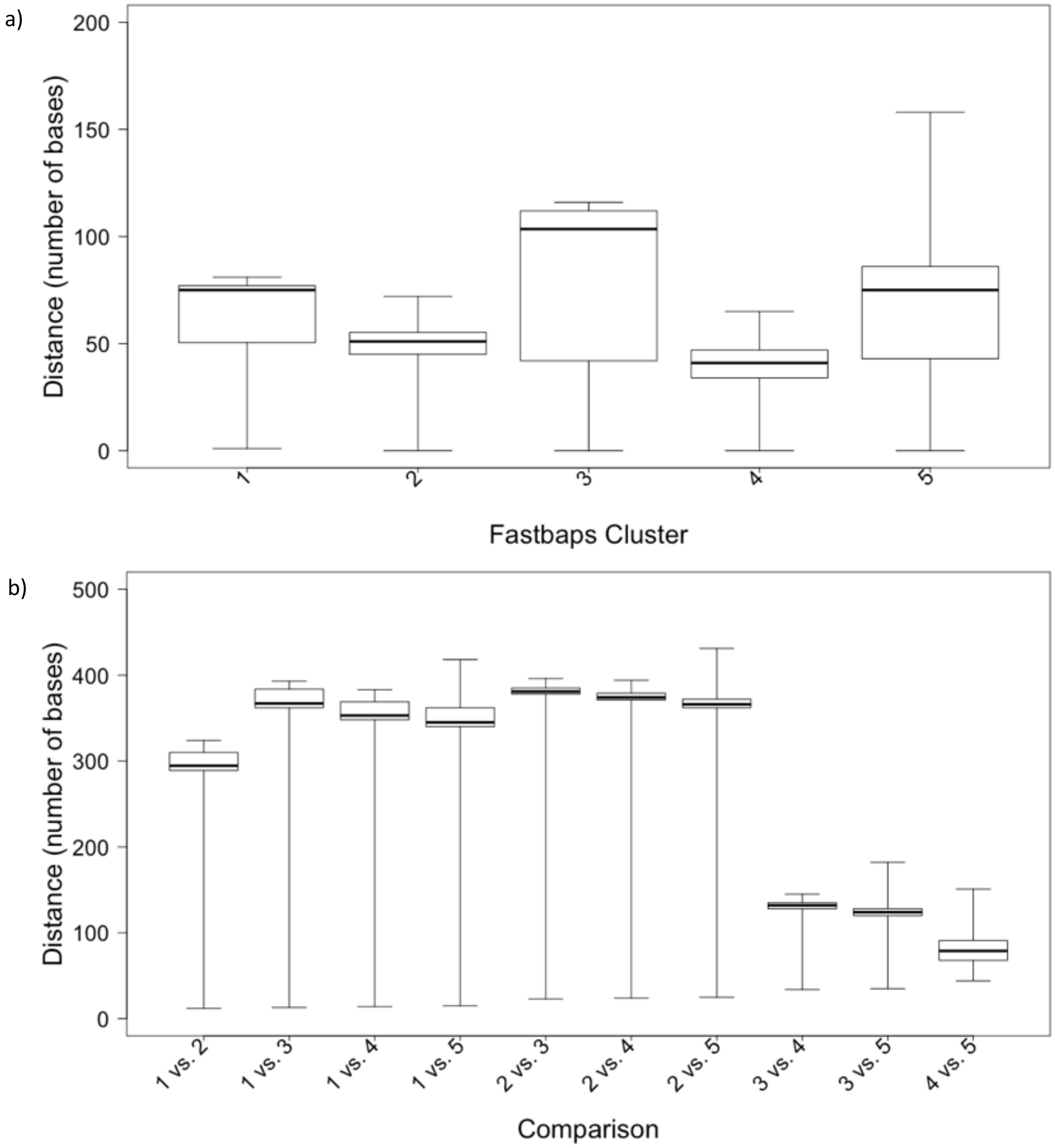


Figure III.18 Distribution of median pairwise SNP distance within and between each eBG297 fastbaps cluster

Box and whisker plot showing the minimum, 1st quartile, median, 3rd quartile and maximum pairwise SNP distances within and between each eBG297 fastbaps cluster.

a) Within each eBG297 fastbaps cluster

b) Between each eBG297 fastbaps cluster

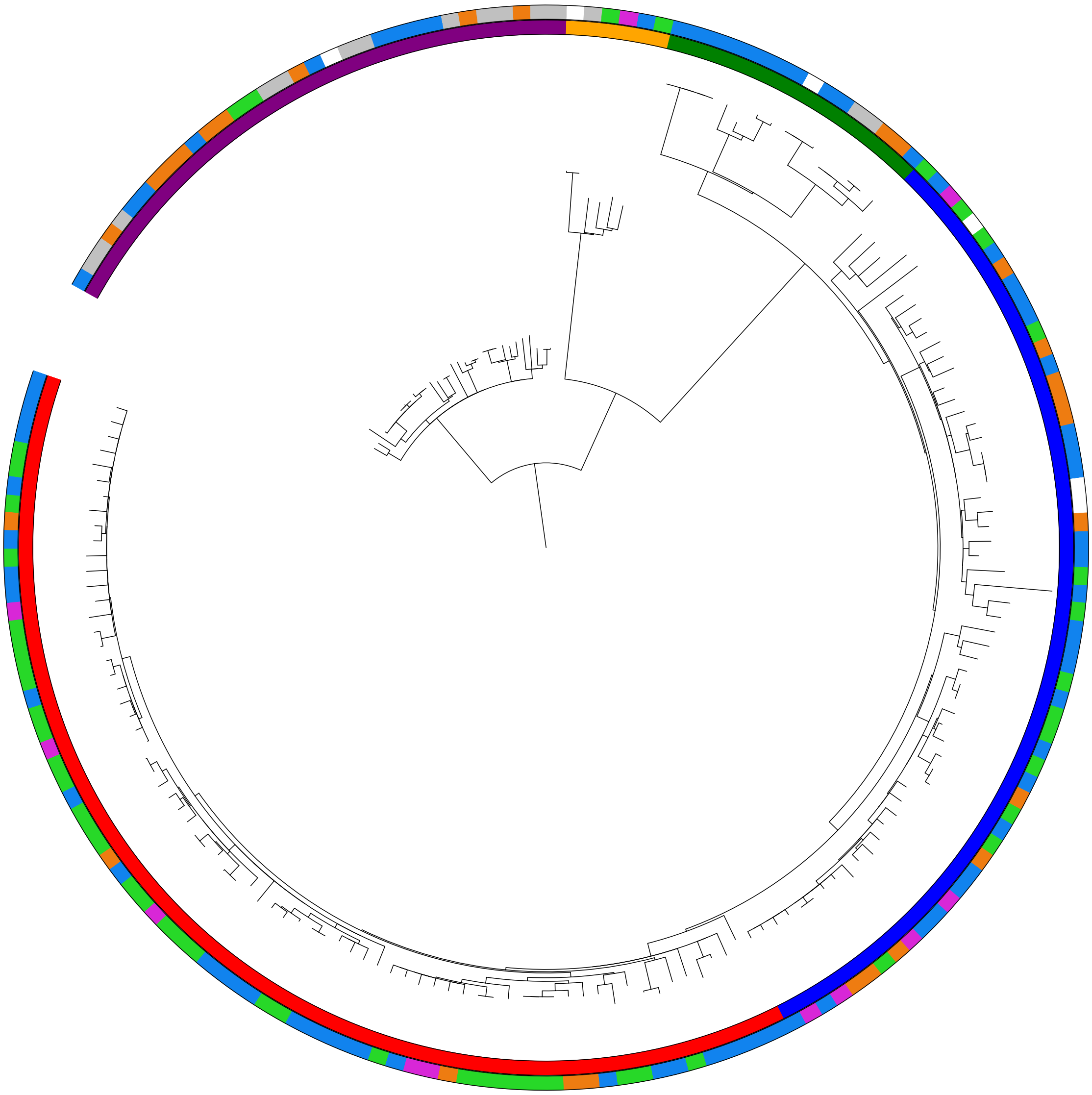


Figure III.19 eBG297 phylogeny annotated with year of isolation
Soft-core SNP Maximum Likelihood Phylogeny of 183 eBG297 isolates, annotated with the year group of isolation.

Inner ring, fastbaps cluster: 1 🟧 2 🟪 3 🟩 4 🟥 5 🟦

Outer ring, year group: 2003-2005 🟪 2006-2010 🟩 2011-2014 🟦
2015-2016 🟧 2017-2019 ▨ Unknown □

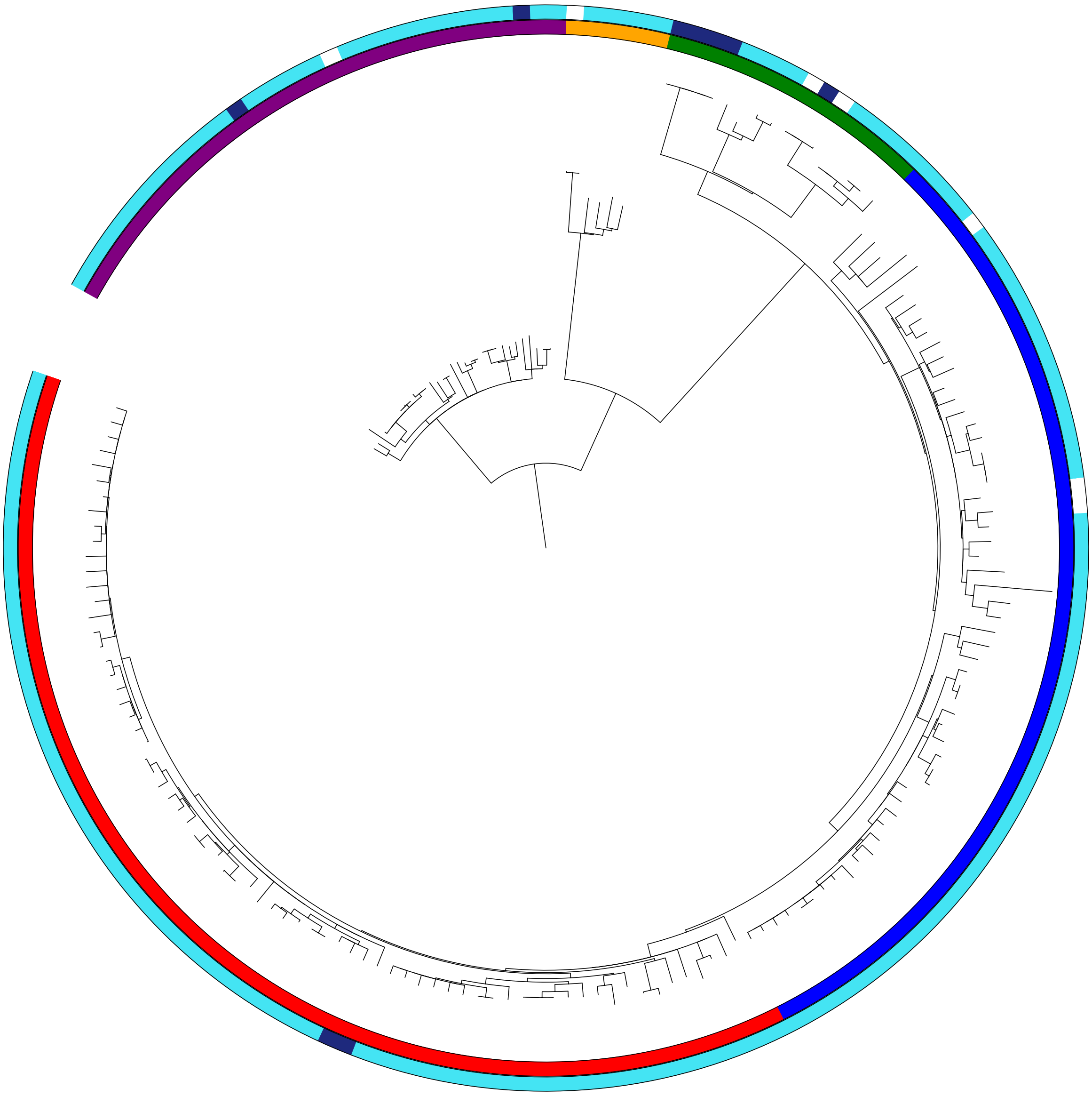


Figure III.20 eBG297 phylogeny annotated with isolation source

Soft-core SNP Maximum Likelihood Phylogeny of 183 eBG297 isolates, annotated with the source of isolation.

Inner ring, fastbaps cluster: 1 🟧 2 🟪 3 🟩 4 🟥 5 🟦

Outer ring, year group: Environmental ■ Human 🟦 Unknown □

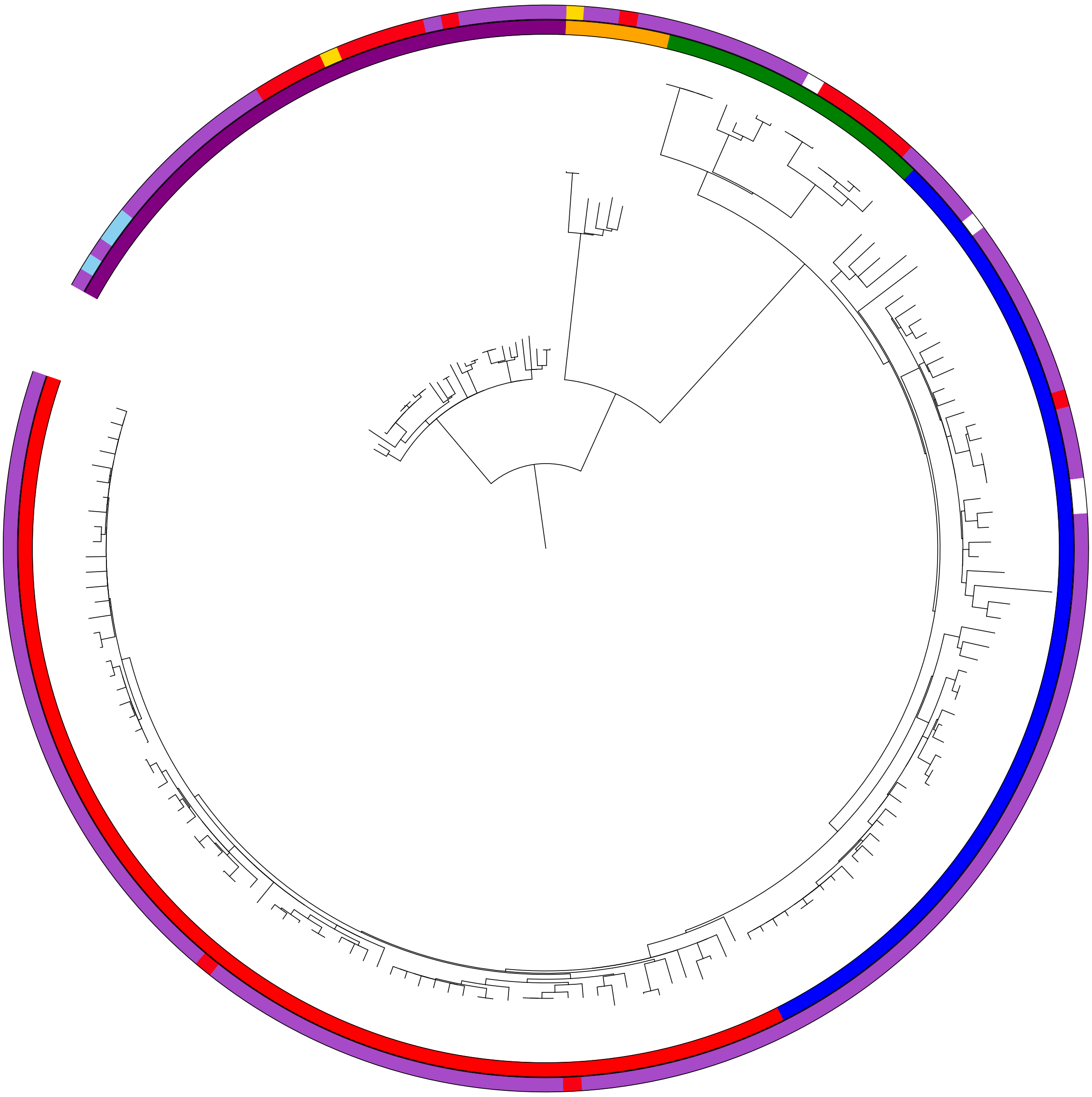


Figure III.21 eBG297 phylogeny annotated with continent of isolation

Soft-core SNP Maximum Likelihood Phylogeny of 183 eBG297 isolates, annotated with the continent of isolation.

Inner ring, fastbaps cluster: 1 🟧 2 🟪 3 🟩 4 🟥 5 🟦

Outer ring, continent: Africa 🟪 Asia 🟦 Europe 🟥 N. America 🟨 Unknown ⬜

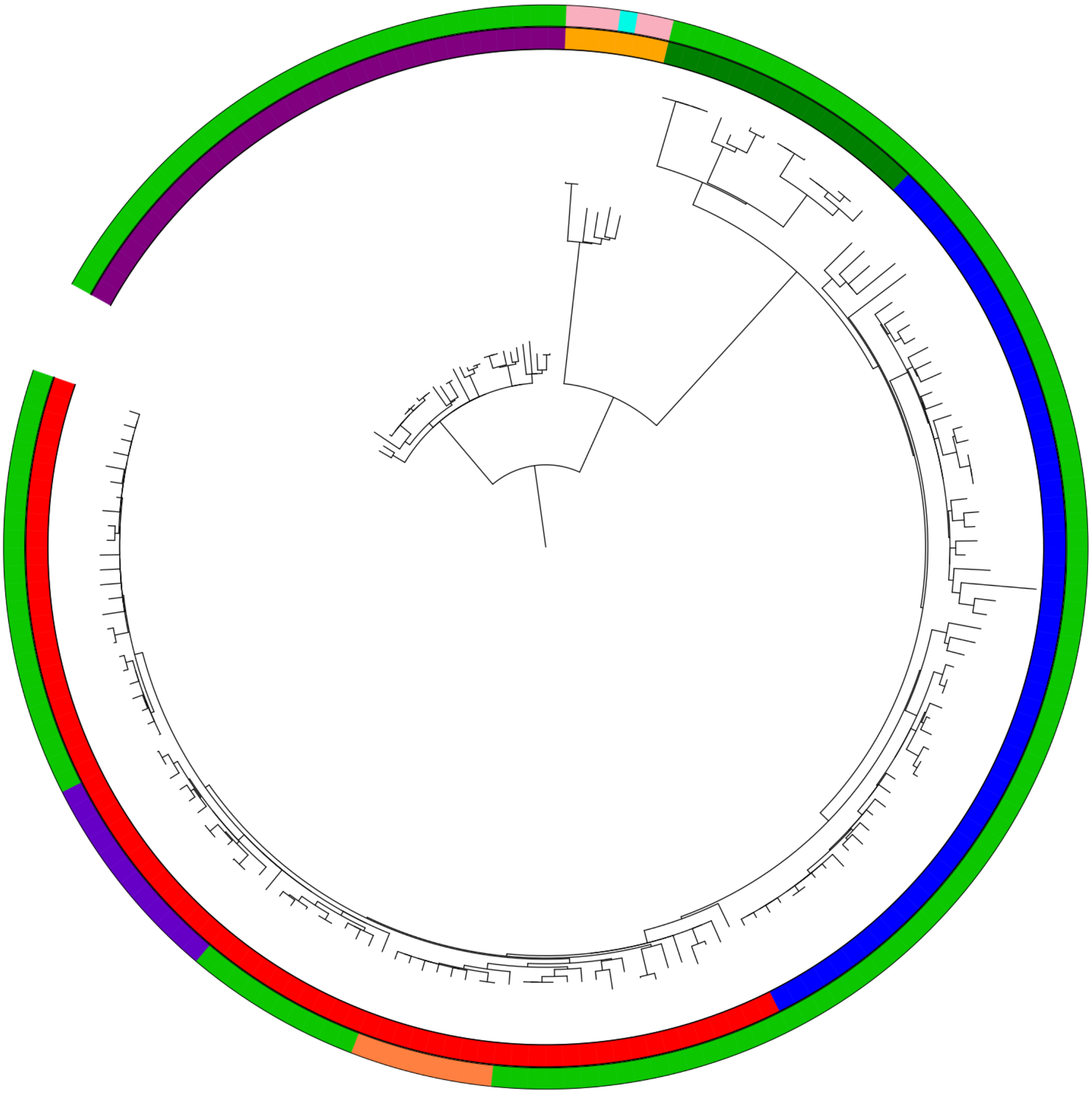


Figure III.22 eBG297 phylogeny annotated with ST
Soft-core SNP Maximum Likelihood Phylogeny of 183 eBG297 isolates, annotated with the ST.

Inner ring, fastbaps cluster: 1 🟧 2 🟪 3 🟩 4 🟥 5 🟦

Outer ring, ST: 603 🟩 1823 🟥 7731 🟧 7732 🟪 novel3 🟦

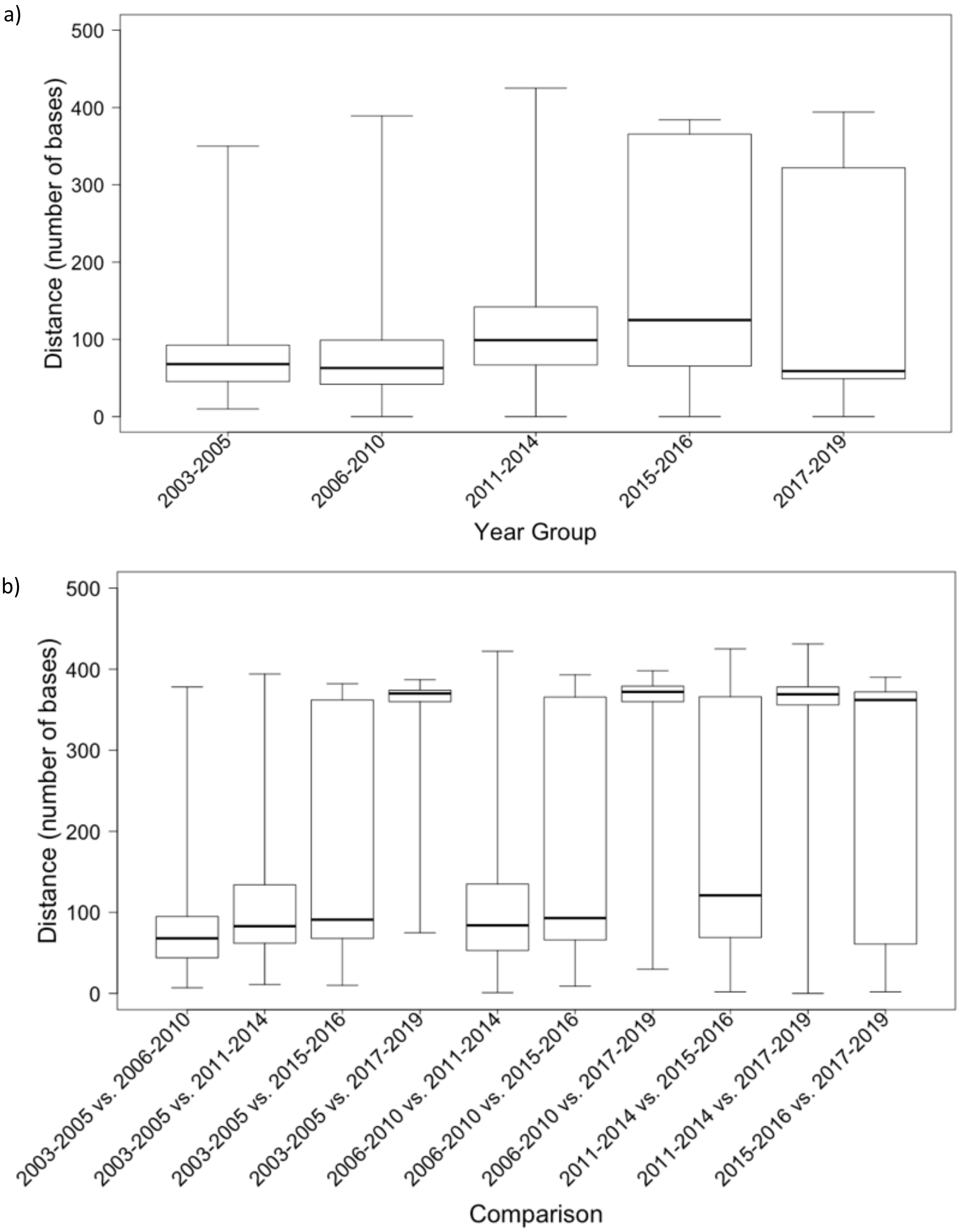


Figure III.23 Distribution of median pairwise SNP distance within and between each eBG297 year group Box and whisker plot showing the minimum, 1st quartile, median, 3rd quartile and maximum pairwise SNP distances within and between each year group.

a) Within each year group

b) Between each year group

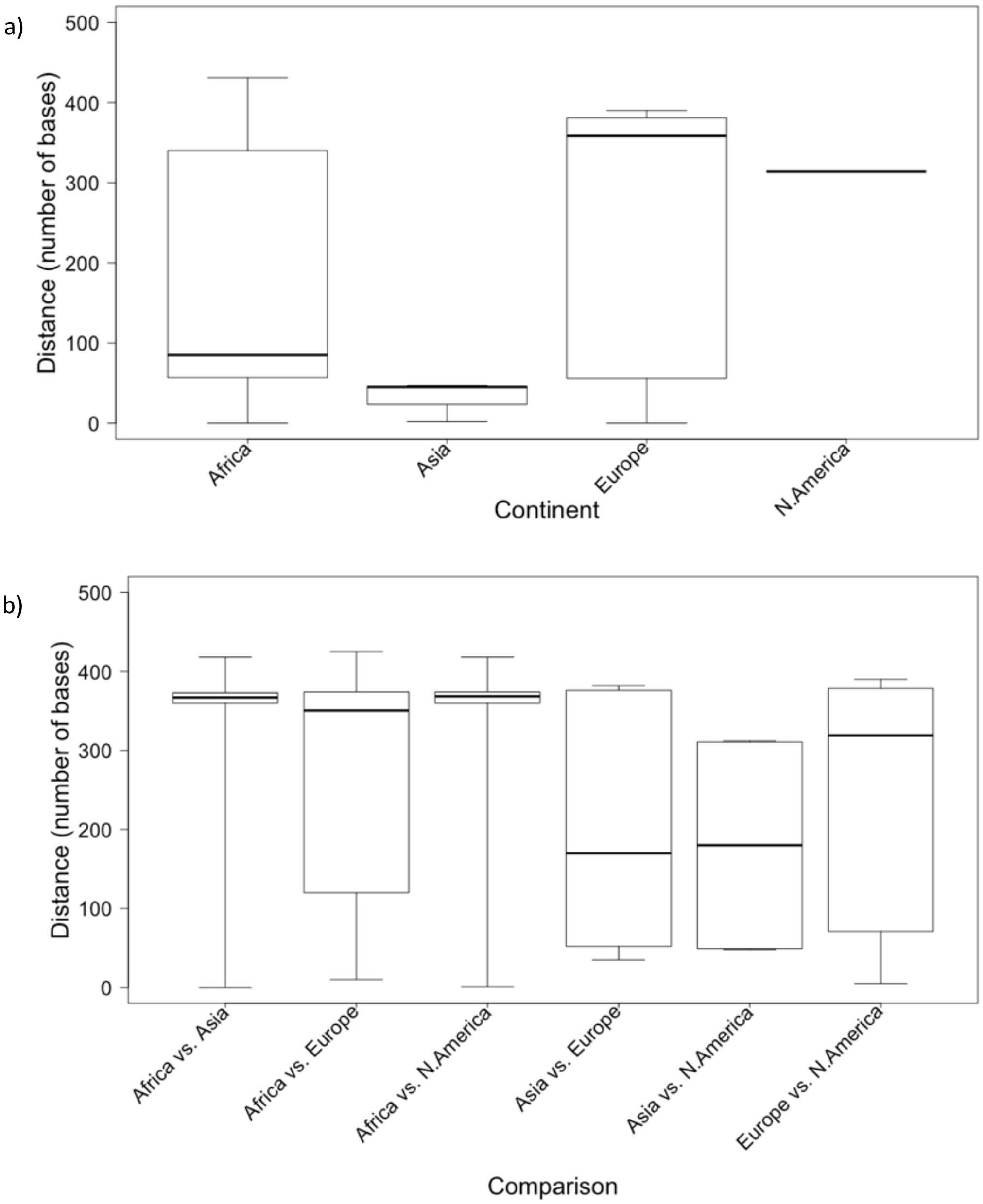


Figure III.24 Distribution of median pairwise SNP distance within and between each eBG297 continent
Box and whisker plot showing the minimum, 1st quartile, median, 3rd quartile and maximum pairwise SNP distances within and between each continent.

a) Within each continent
b) Between each continent

IV Appendix IV. Supplementary Data for Chapter 4

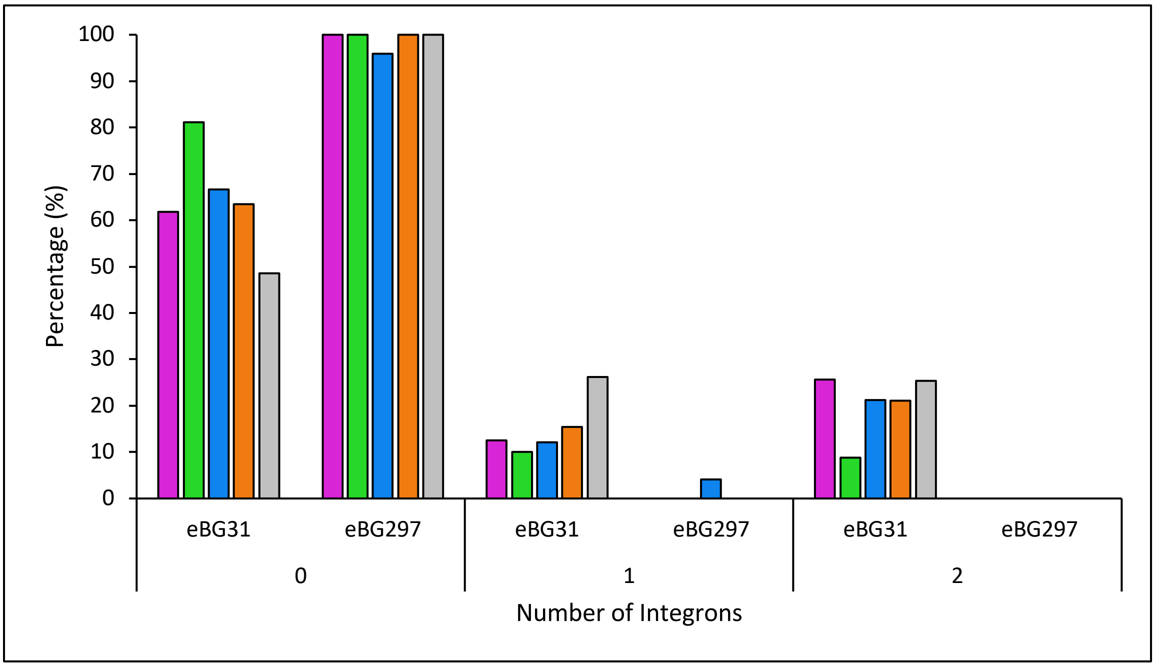


Figure IV.1 Difference in integron frequency across *S.* Infantis by year group
Shown as a percentage of isolates from each year group.
eBG31: 1989-2005 (n=144), 2006-2010 (n=39), 2011-2014 (n=715), 2015-2016 (n=1177) and 2017-2018 (n=1158)
eBG297: 2003-2005 (n=10), 2006-2010 (n=49), 2011-2014 (n=73), 2015-2016 (n=26) and 2017-2019 (n=16)
1989-2005 🟪 2006-2010 🟩 2011-2014 🟦 2015-2016 🟧 2017-2018 ⬜

		eBG31			eBG297		
		Lincosamides	Macrolides	Polymyxins	Lincosamides	Macrolides	Polymyxins
	eBG	0.38	0.27	0.002	0	2.17	0
Source	Environmental	0.53	0.53	0.1	0	0	0
	Human	0.39	0.39	0	0	2.38	0
	Poultry	0.53	0	0	0	0	0
Origin	Africa	0	0.67	0	0	2.58	0
	Asia	0.42	0.42	0	0	0	0
	Europe	1.67	0.42	0.1	0	0	0
	North America	0	0.18	0	0	0	0
	South America	0	0	0	0	0	0
Year Group	1989-2005	0	0.68	0	0	0	0
	2006-2010	0	0.99	0	0	0	0
	2011-2014	0.40	0.26	0	0	4.00	0
	2015-2016	0.75	0.33	0.1	0	3.70	0
	2017-2019	0.43	0.09	0	0	0	0

Table IV.1 Resistance to Lincosamides, Macrolides and Polymyxins

Percentage of eBG31 (n=4486) and eBG297 (n=184) isolates from each eBG, source, origin and year group with resistance to Lincosamides, Macrolides and Polymyxins.
eBG31 source: environmental (n=947); human (n=1519) and poultry (n=947)
eBG297 source: environmental (n=9); human (n=168) and poultry (n=0)
eBG31 origin: Africa (n=297); Asia (n=238); Europe (n=959); North America (n=2793) and South America (n=122)
eBG297 origin: Africa (n=155); Asia (n=3); Europe (n=20); North America (n=2) and South America (n=0)
eBG31 year: 1989-2005 (n=147), 2006-2010 (n=404), 2011-2014 (n=759), 2015-2016 (n=1203), 2017-2018 (n=1164)
eBG297 year: 2003-2005 (n=11), 2006-2010 (n=49), 2011-2014 (n=75), 2015-2016 (n=27), 2017-2019 (n=16)

	eBG31	eBG297
Col156	0.29	0.54
Col3M	0.02	0
Col440I	0.25	0
Col8282	0.16	0
Col(RNAI)	0.56	0
Col(BS512)	0.16	0.54
Col(IMGS31)	0	0.54
Col(MG828)	0.11	0
Col(pVC)	0.87	0
IncB/O/K/Z	0.13	0
IncH	0.40	0
IncL/M	0.11	0
IncN	0.45	0.54
IncQ	0.11	0
IncR	0.11	0
IncU	0.07	0
p0111	0.09	0
repA	0.09	0

Table IV.2 Rare plasmid distribution by eBG
Percentage of eBG31 (n=4486) and eBG297 (n=184) isolates containing each rare plasmid type with an overall isolation rate of less than 1% per eBG.

	eBG31			eBG297		
	Environmental	Human	Poultry	Environmental	Human	Poultry
Col156	0.11	0.33	0.11	0	0.60	0
Col3M	0	0.07	0	0	0	0
Col440I	0.11	0.20	0	0	0	0
Col8282	0.11	0.33	0	0	0	0
Col(RNAI)	0.21	0.92	0.42	0	0	0
Col(BS512)	0	0.13	0.11	0	0.60	0
Col(IMGS31)	0	0	0	0	0.60	0
Col(MG828)	0	0.13	0	0	0	0
Col(pVC)	0.63	0.46	0.84	0	0	0
IncB/O/K/Z	0	0.33	0	0	0	0
IncH	1.27	0	0	0	0	0
IncL/M	0.42	0	0	0	0	0
IncN	0	0.66	0.11	0	0	0
IncQ	0	0.13	0	0	0	0
IncR	0	0.20	0.21	0	0	0
IncU	0	0.20	0	0	0	0
p0111	0	0.26	0	0	0	0
repA	0.11	0.20	0	0	0	0

Table IV.3 Rare plasmid distribution by isolation source

Percentage of eBG31 and eBG297 isolates from each source containing each rare plasmid type with an overall isolation rate of less than 1% per eBG.

eBG31: environmental (n=947); human (n=1519) and poultry (n=947)

eBG297: environmental (n=9); human (n=168) and poultry (n=0)

	eBG31					eBG297				
	Africa	Asia	Europe	North America	South America	Africa	Asia	Europe	North America	South America
Col156	0.34	0	0.94	0.04	0.82	0.65	0	0	0	0
Col3M	0	0	0.10	0	0	0	0	0	0	0
Col440I	0	0	0.31	0.25	0	0	0	0	0	0
Col8282	1.01	0.42	0.21	0.04	0	0	0	0	0	0
Col(RNAI)	0.34	0.42	0.73	0.54	0	0	0	0	0	0
Col(BS512)	0	0	0.10	0.18	0.82	0.65	0	0	0	0
Col(IMGS31)	0	0	0	0	0	0.65	0	0	0	0
Col(MG828)	0	0	0.52	0	0	0	0	0	0	0
Col(pVC)	0.34	0	1.25	0.90	0.82	0	0	0	0	0
IncB/O/K/Z	1.01	0	0	0.11	0	0	0	0	0	0
IncH	0	0	0.21	0.57	0	0	0	0	0	0
IncL/M	0	0	0.21	0.11	0	0	0	0	0	0
IncN	0.67	0	0.73	0.39	0	0.65	0	0	0	0
IncQ	0	0	0.42	0.04	0	0	0	0	0	0
IncR	0	0	0.31	0	1.64	0	0	0	0	0
IncU	0.34	0	0.21	0	0	0	0	0	0	0
p0111	0	0	0.10	0.11	0	0	0	0	0	0
repA	0	0	0.21	0.07	0	0	0	0	0	0

Table IV.4 Rare plasmid distribution by continent

Percentage of eBG31 and eBG297 isolates from each continent containing each rare plasmid type with an overall isolation rate of less than 1% per eBG.

eBG31 origin: Africa (n=297); Asia (n=238); Europe (n=959); North America (n=2793) and South America (n=122)

eBG297 origin: Africa (n=155); Asia (n=3); Europe (n=20); North America (n=2) and South America (n=0)

	eBG31					eBG297				
	1989 to 2005	2006 to 2010	2011 to 2014	2015 to 2016	2017 to 2018	2003 to 2005	2006 to 2010	2011 to 2014	2015 to 2016	2017 to 2019
Col156	1.36	0	0.92	0.25	0	0	0	0	3.70	0
Col3M	0	0	0	0.08	0	0	0	0	0	0
Col440I	0	0	0.26	0.33	0.09	0	0	0	0	0
Col8282	0	0.25	0.26	0.33	0	0	0	0	0	0
Col(RNAI)	0.68	0.25	0.66	0.83	0.43	0	0	0	0	0
Col(BS512)	0	0	0	0.33	0.09	0	0	0	3.70	0
Col(IMGS31)	0	0	0	0	0	9.09	0	0	0	0
Col(MG828)	0	0	0.40	0.17	0	0	0	0	0	0
Col(pVC)	0	0.99	1.71	1.16	0.34	0	0	0	0	0
IncB/O/K/Z	0	0.50	0.13	0.08	0.09	0	0	0	0	0
IncH	1.36	0	1.05	0.50	0.17	0	0	0	0	0
IncL/M	0	0	0.53	0	0	0	0	0	0	0
IncN	0	0.25	0.53	0.42	0.34	0	0	0	3.70	0
IncQ	0	0	0.40	0.17	0	0	0	0	0	0
IncR	0	0	0.26	0.25	0	0	0	0	0	0
IncU	0	0	0.40	0	0	0	0	0	0	0
p0111	0	0	0.13	0.08	0.17	0	0	0	0	0
repA	0	0	0.26	0.08	0.09	0	0	0	0	0

Table IV.5 Rare plasmid distribution by year group

Percentage of eBG31 and eBG297 isolates from each year group containing each rare plasmid type with an overall isolation rate of less than 1% per eBG.

eBG31 year: 1989-2005 (n=147), 2006-2010 (n=404), 2011-2014 (n=759), 2015-2016 (n=1203), 2017-2018 (n=1164)

eBG297 year: 2003-2005 (n=11), 2006-2010 (n=49), 2011-2014 (n=75), 2015-2016 (n=27), 2017-2019 (n=16)

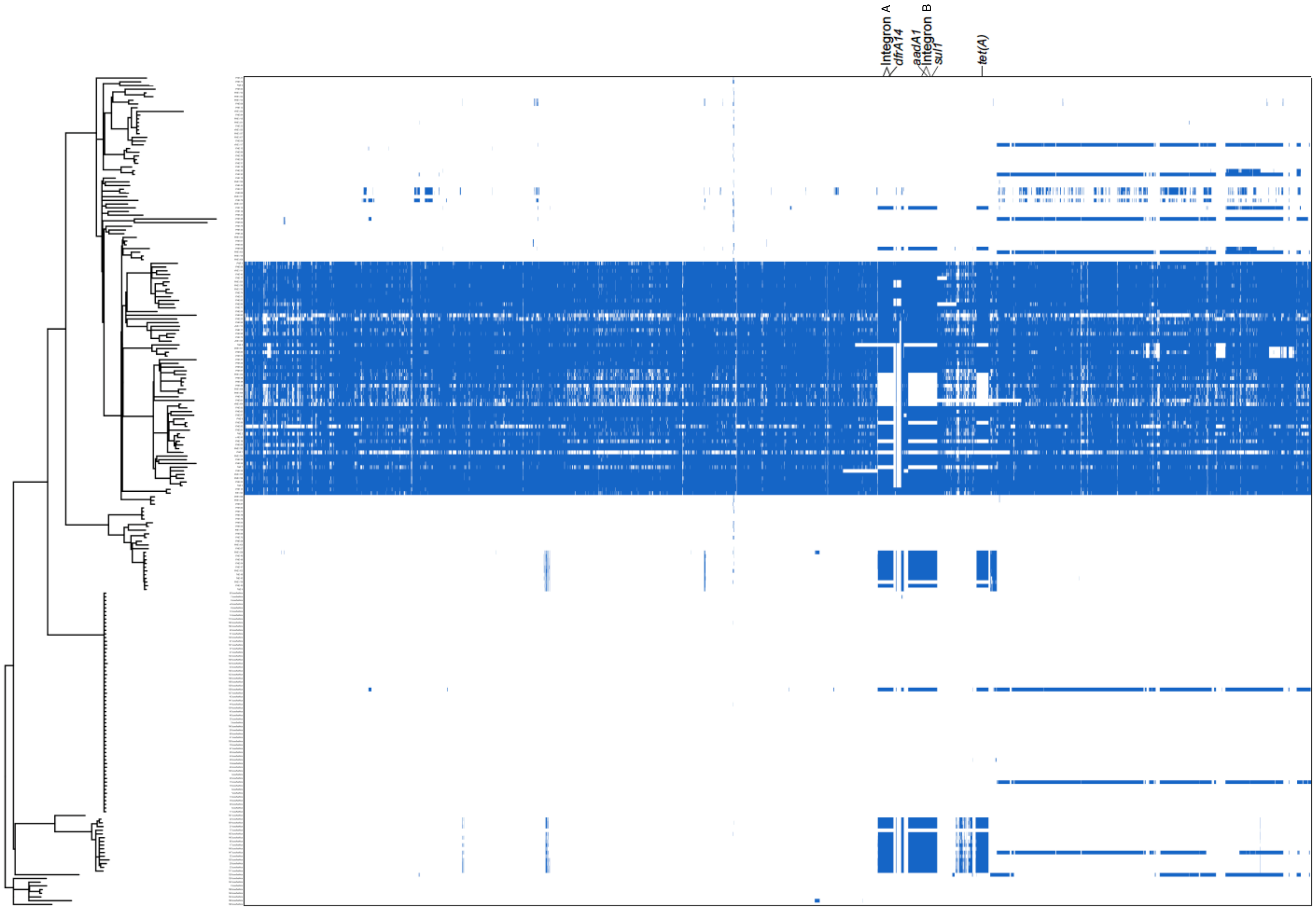


Figure IV.2 pESI presence in eBG31 from England & Wales and South Africa
Soft-core SNP maximum likelihood phylogeny of 139 England & Wales eBG31 sequences and 85 South African eBG31 sequences rooted to the most ancestral node. Heat map showing mapped sequence read coverage for the *S*. Infantis isolates to the pESI plasmid. The colour blue indicates a depth of ≥ 20 mapped reads.

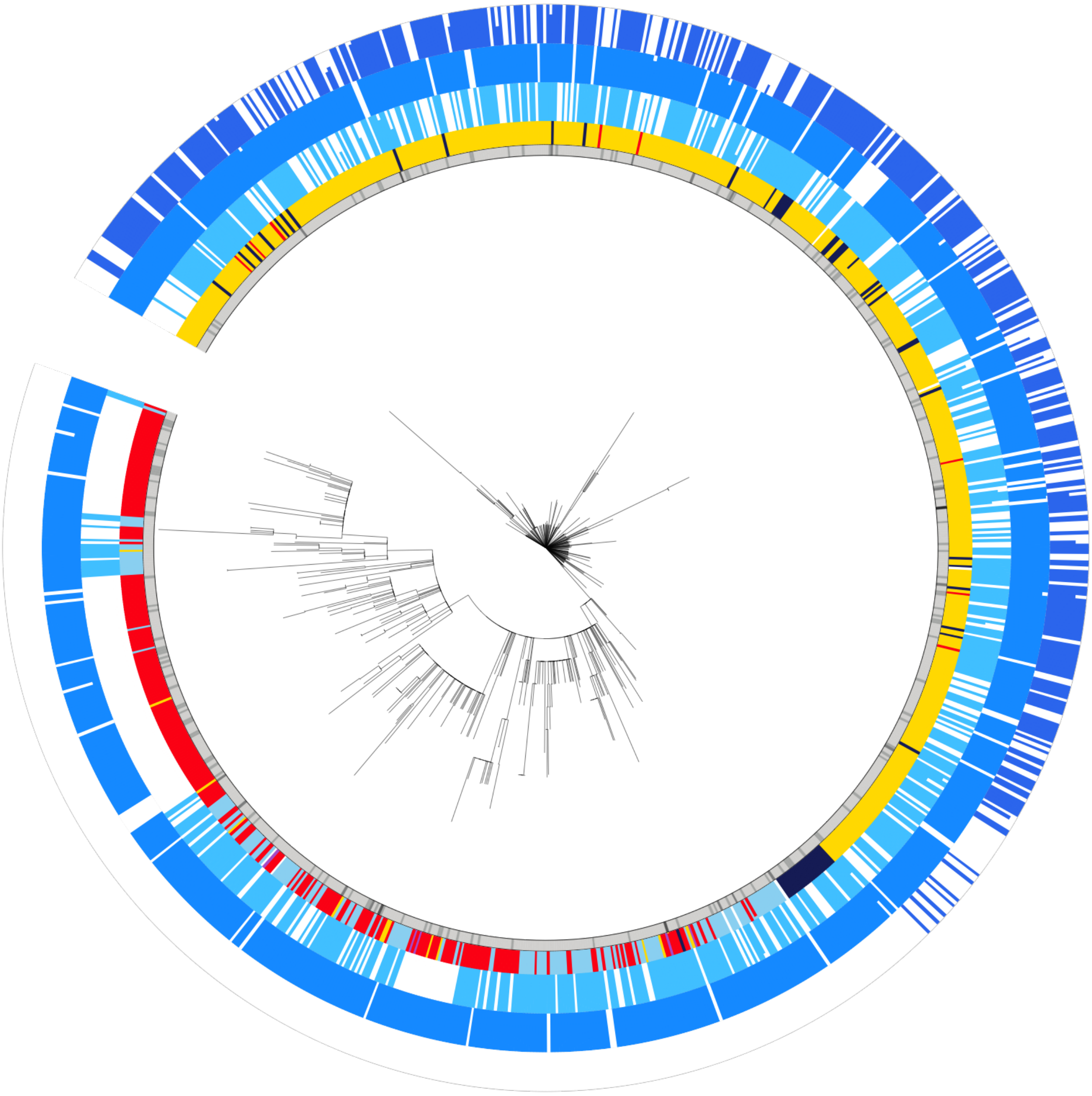


Figure IV.3 Phylogenetic distribution of pESI variants

Soft-core SNP Maximum Likelihood Phylogeny of 1013 5SNP cluster representatives of pESI. The outer rings are annotated with the percentage of isolates in each 5SNP cluster that were isolated from each continent or had each pESI variant.

Inner ring, number of sequences in 5SNP cluster: □ 1 ░ 2-5 ▒ 6-20 ▓ 21-50 ■ >50

Second ring, percentage of sequences in each cluster from each continent:
Africa 🟪 Asia 🟦 Europe 🟥 North America 🟨 South America ⬛ Unknown ⬜
Third ring, percentage of sequences in each cluster containing Integron 1:
Presence 🟦
Fourth ring, percentage of sequences in each cluster containing Integron 2:
Presence 🟦
Outer ring, percentage of sequences in each cluster containing *bla*$_{\text{CTX-M-65}}$ on pESI:
Presence 🟦

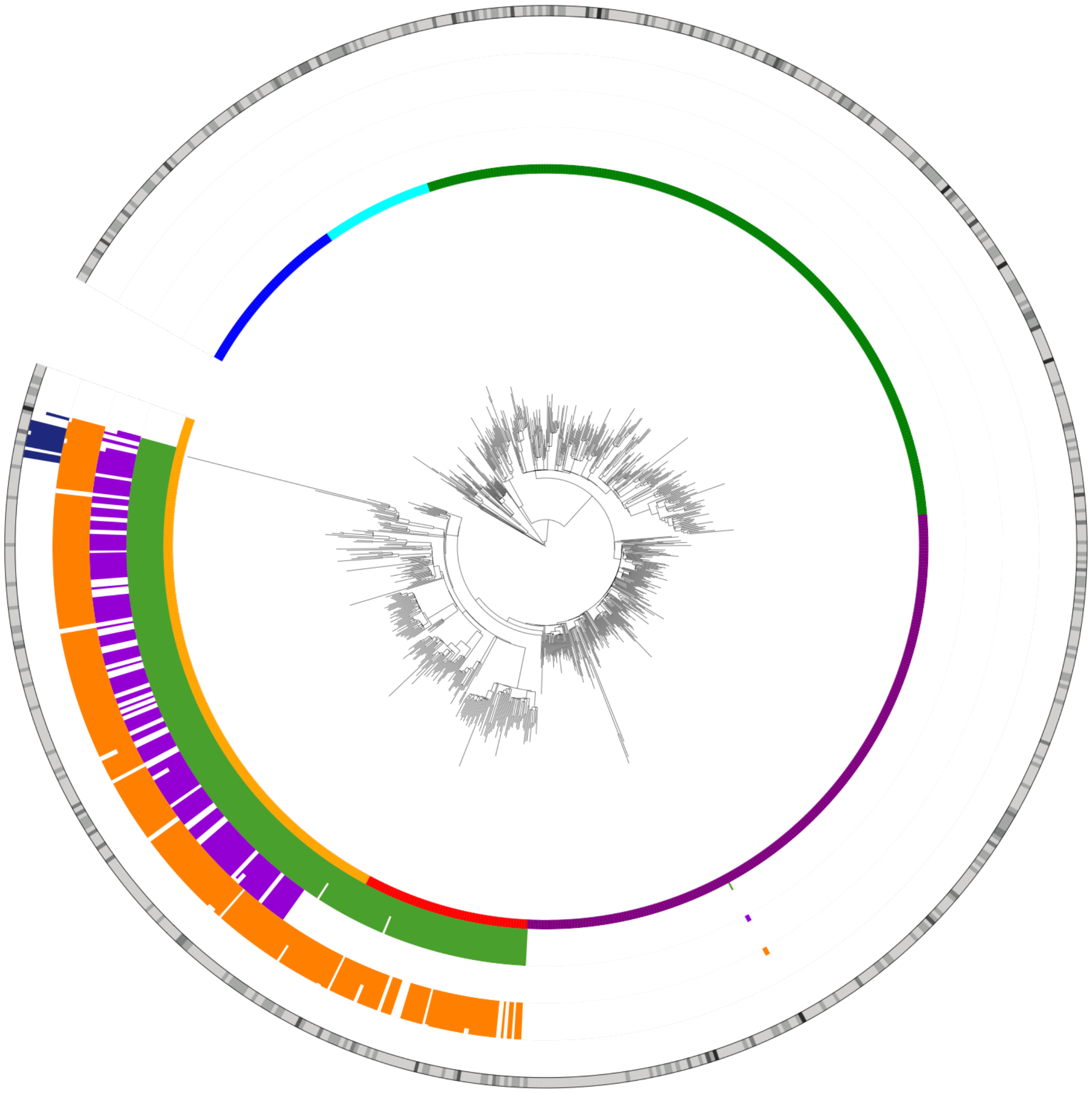


Figure IV.4 eBG31 phylogeny annotated with pESI variant

Soft-core SNP Maximum Likelihood Phylogeny of 831 25SNP cluster representatives of eBG31. The outer rings are annotated with the percentage of isolates in each 25SNP cluster that had each pESI variant.

Inner ring, fastbaps cluster: 1 🟥 2 🟦 3 🟦 4 🟪 5 🟧 6 🟩

Second ring, pESI presence: 🟩
Third ring, pESI with Integron 1: 🟪
Fourth ring, pESI with Integron 2: 🟧
Fifth ring, pESI with *bla*$_{\text{CTX-M-65}}$: 🟦
Outer ring, Number of sequences in 25SNP cluster: ⬜ 1 ▩ 2-5 ▩ 6-20 ▩ 21-50 ⬛ >50

V Appendix V. Supplementary Data for Chapter 5

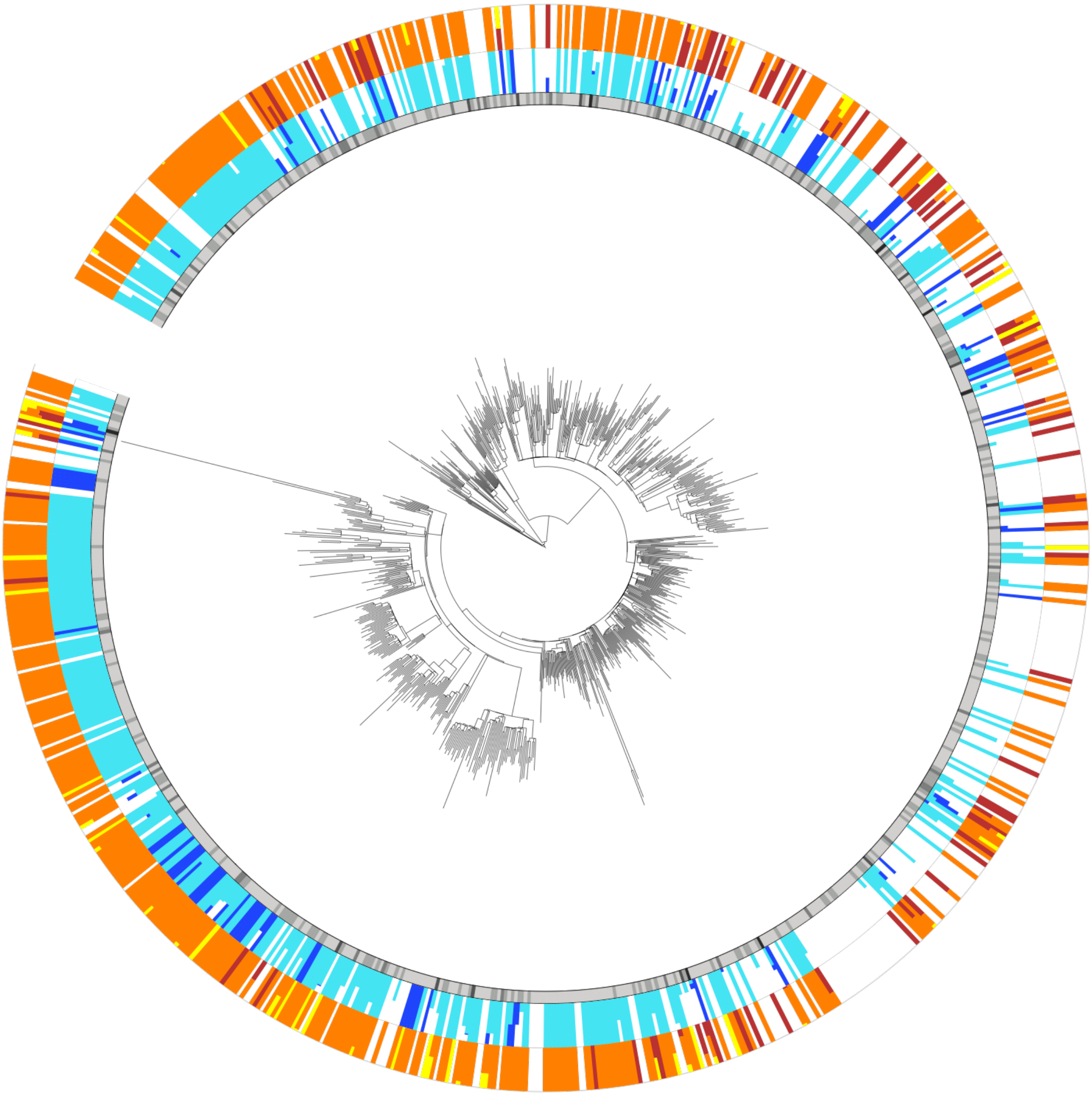


Figure V.1 eBG31 phylogeny annotated with *tufA* variant
Soft-core SNP Maximum Likelihood Phylogeny of 831 25SNP cluster representatives of eBG31. Each representative isolate is labelled with the percentage of isolates from humans or poultry sources in its cluster that contained *tufA*_1, *tufA*_2 or both variants.

Inner ring, Number of sequences in 25SNP cluster: ⬜ 1 ▒ 2-5 ▓ 6-20 ▓ 21-50 ⬛ >50

Second ring, Source: Human 🟦 Poultry 🟦

Outer ring, *tufA* variant: *tufA*_1 🟥 *tufA*_2 🟧 Both *tufA*_1 and *tufA*_2 🟨

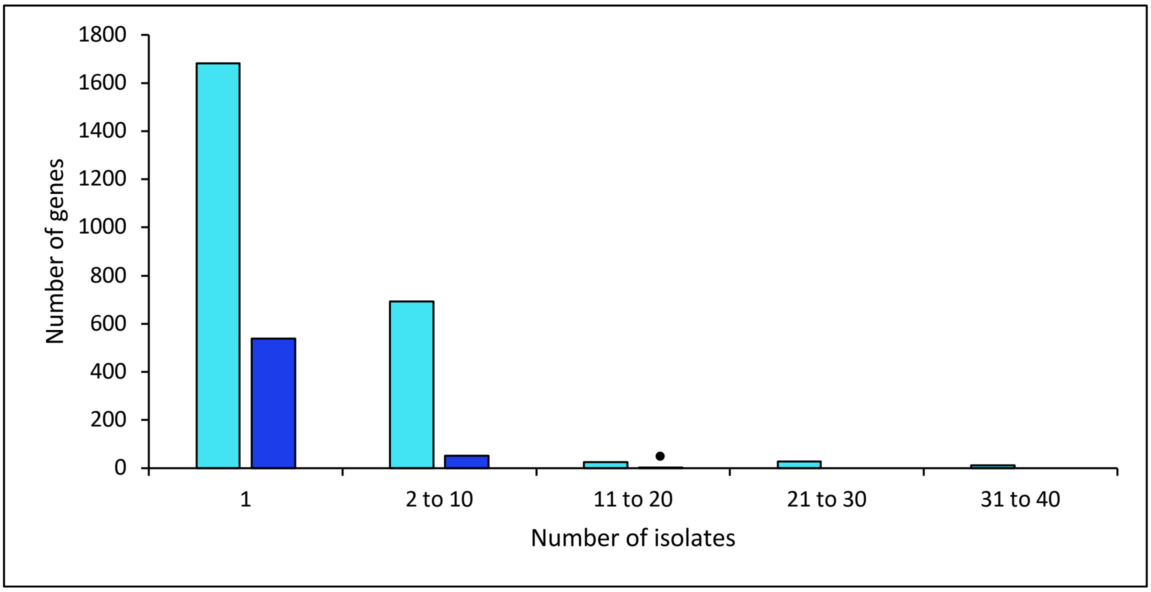


Figure V.2 Isolation frequency of known protein-coding genes unique to human or poultry eBG31 isolates

2,440 protein-coding genes with a known function were found exclusively in 1482 eBG31 human isolates. 593 protein-coding genes with a known function were found exclusively in 945 eBG31 poultry isolates.

• values are less than 10 but greater than 0.

Human □ Poultry ■

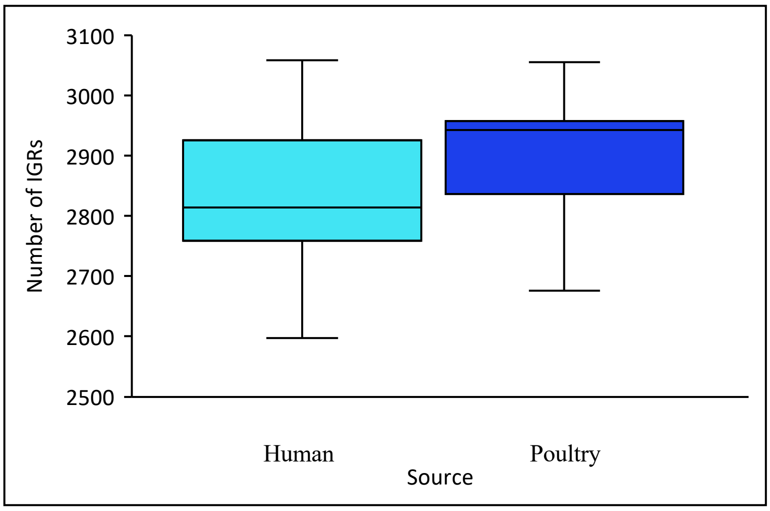


Figure V.3 Difference in the distribution in the number of IGRs in eBG31 human and poultry isolates.

Box plot showing the variation in the number of IGRs in each eBG31 isolate from human sources (n=1482) and poultry sources (n=945).

Human □ Poultry ■

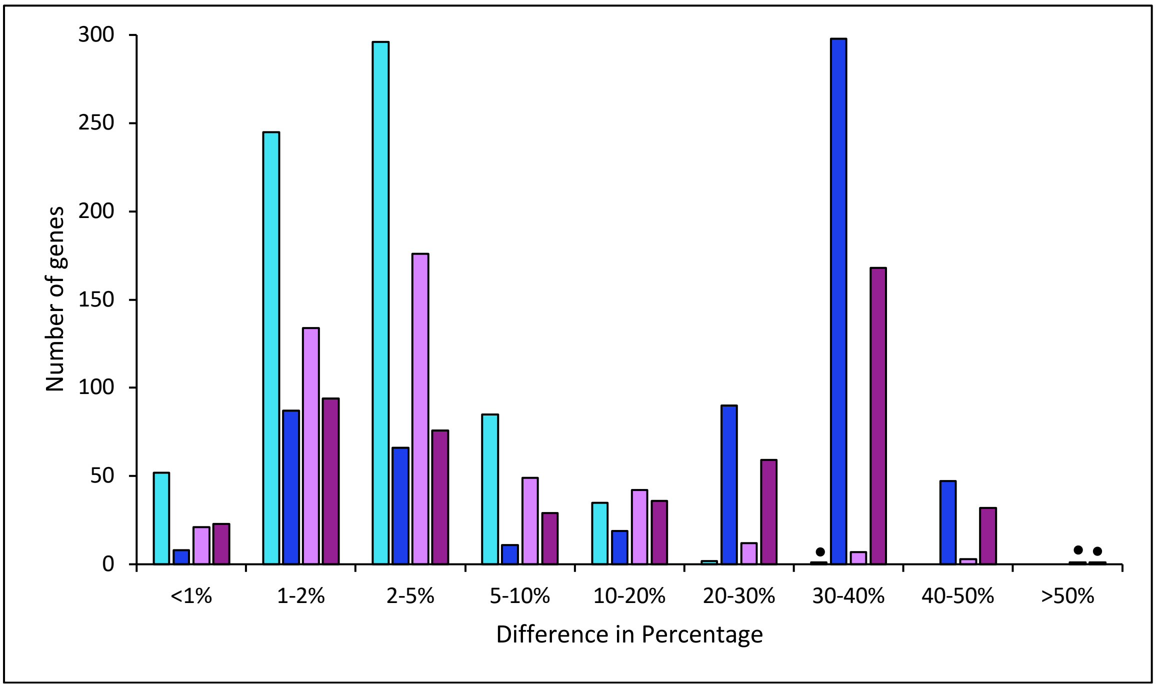


Figure V.4 Distribution of the difference in the percentage of associated genes and IGRs between sources The percentage of isolates with a lower proportion of each significantly associated gene/IGR was subtracted from the percentage of isolates with each gene/IGR from the other source. 1482 eBG31 human isolates and 945 poultry eBG31 isolates were compared. • values are less than 2 but greater than 0.

Human – Genes 🟦 Poultry – Genes 🟦 Human – IGRs 🟪 Poultry – IGRs 🟪

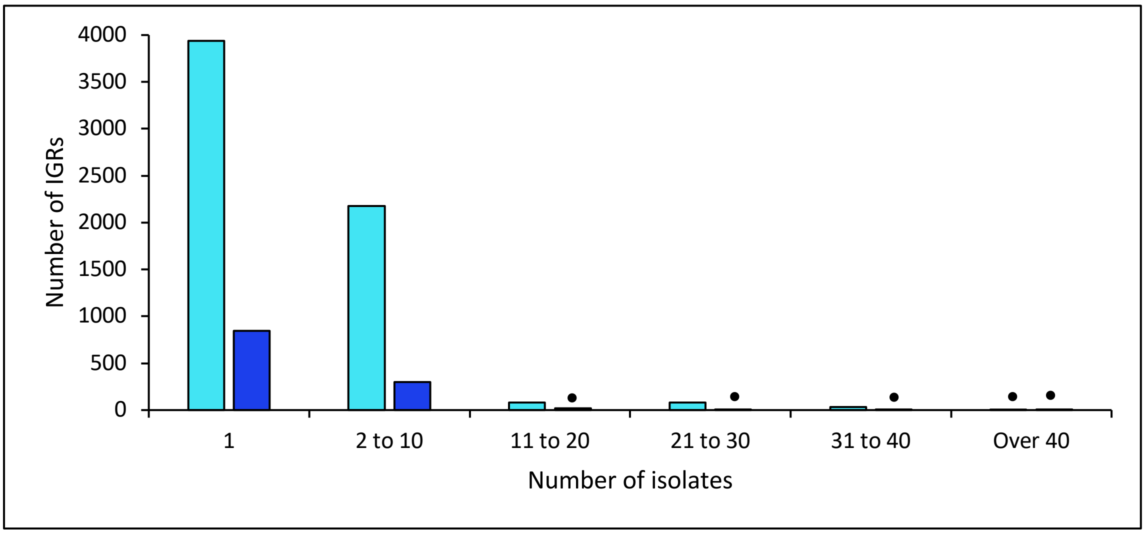


Figure V.5 Isolation frequency of IGRs unique to human or poultry eBG31 isolates
6,319 IGRs were found exclusively in 1,482 eBG31 human isolates. 1,172 IGRs were found exclusively in 945 eBG31 poultry isolates. • values are less than 20 but greater than 0.

Human 🟦 Poultry 🟦

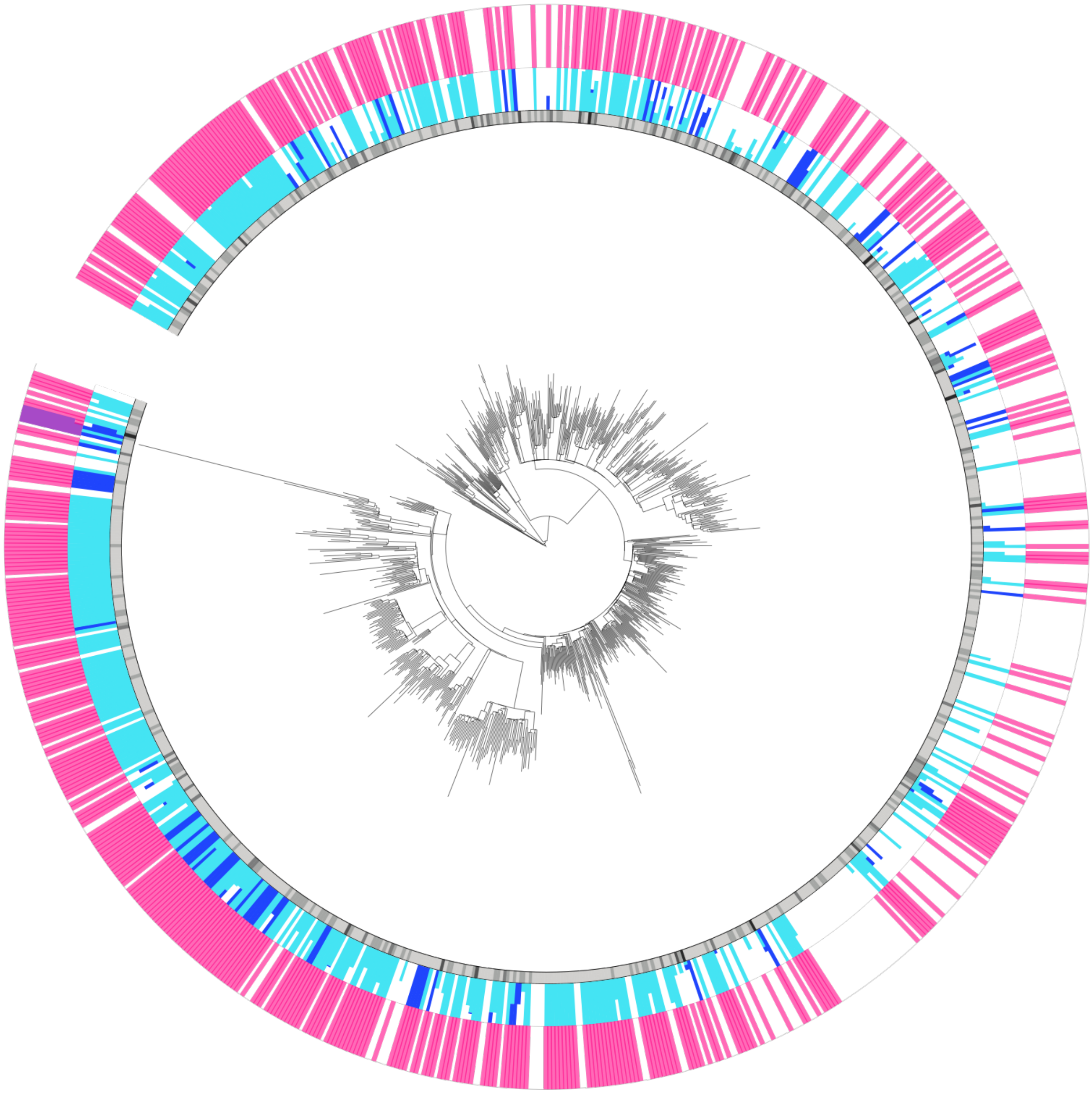


Figure V.6 eBG31 phylogeny annotated with source associated IGR variants
Soft-core SNP Maximum Likelihood Phylogeny of 831 25SNP cluster representatives of eBG31. Each representative isolate is labelled with the percentage of isolates from humans or poultry sources in its cluster that contained the poultry or human associated variant of the IGR between *yfkN* and a hypothetical protein.

Inner ring, Number of sequences in 25SNP cluster: ■ 1 ■ 2-5 ■ 6-20 ■ 21-50 ■ >50

Second ring, Source: Human ■ Poultry ■

Outer ring, IGR variant: Human associated variant ■ Poultry associated variant ■

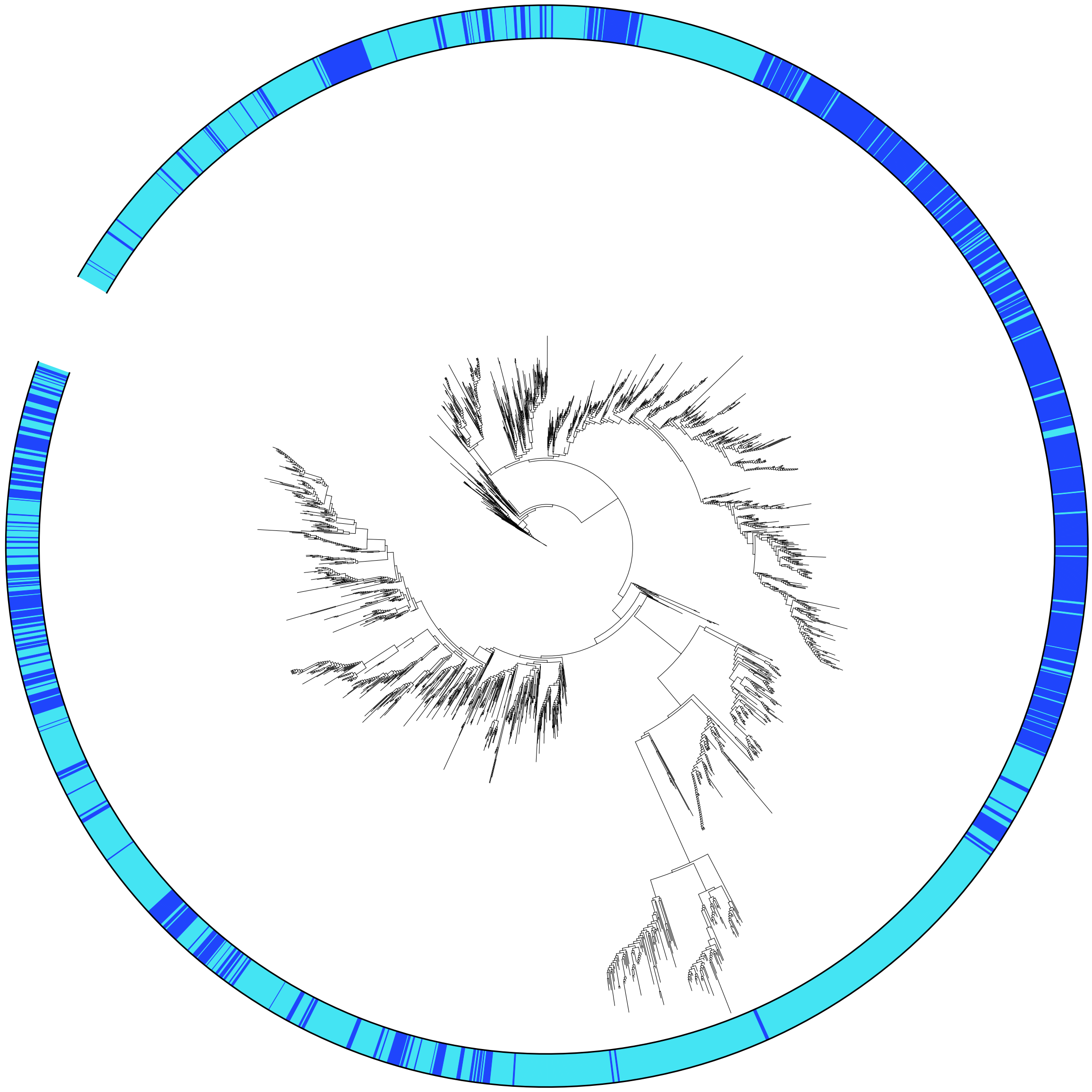


Figure V.7 Maximum likelihood phylogeny of core IGRs
Phylogeny of 2464 core IGRs from 1482 eBG31 human isolates and 945 eBG31 poultry isolates.

Human 🟦 Poultry 🟦

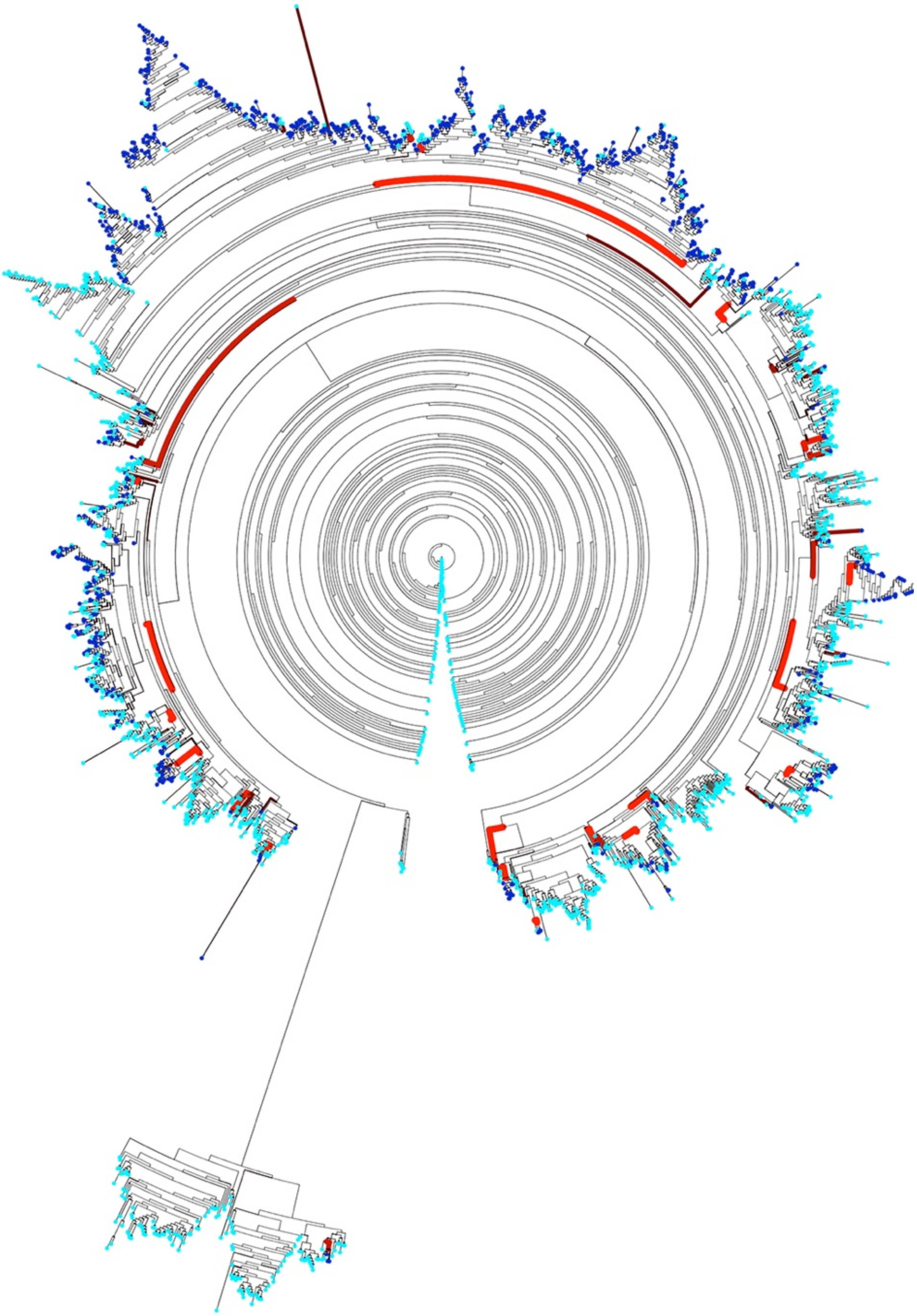


Figure V.8 Maximum likelihood phylogeny of core genes annotated with distinct phenotype distribution

Phylogeny of 4132 core genes from 1482 eBG31 human isolates and 945 eBG31 poultry isolates.

The branch thickness and redness is in proportion to the posterior probability ($p > 0.5$) that a change of isolation source has occurred.

Human 🟦 Poultry 🟦

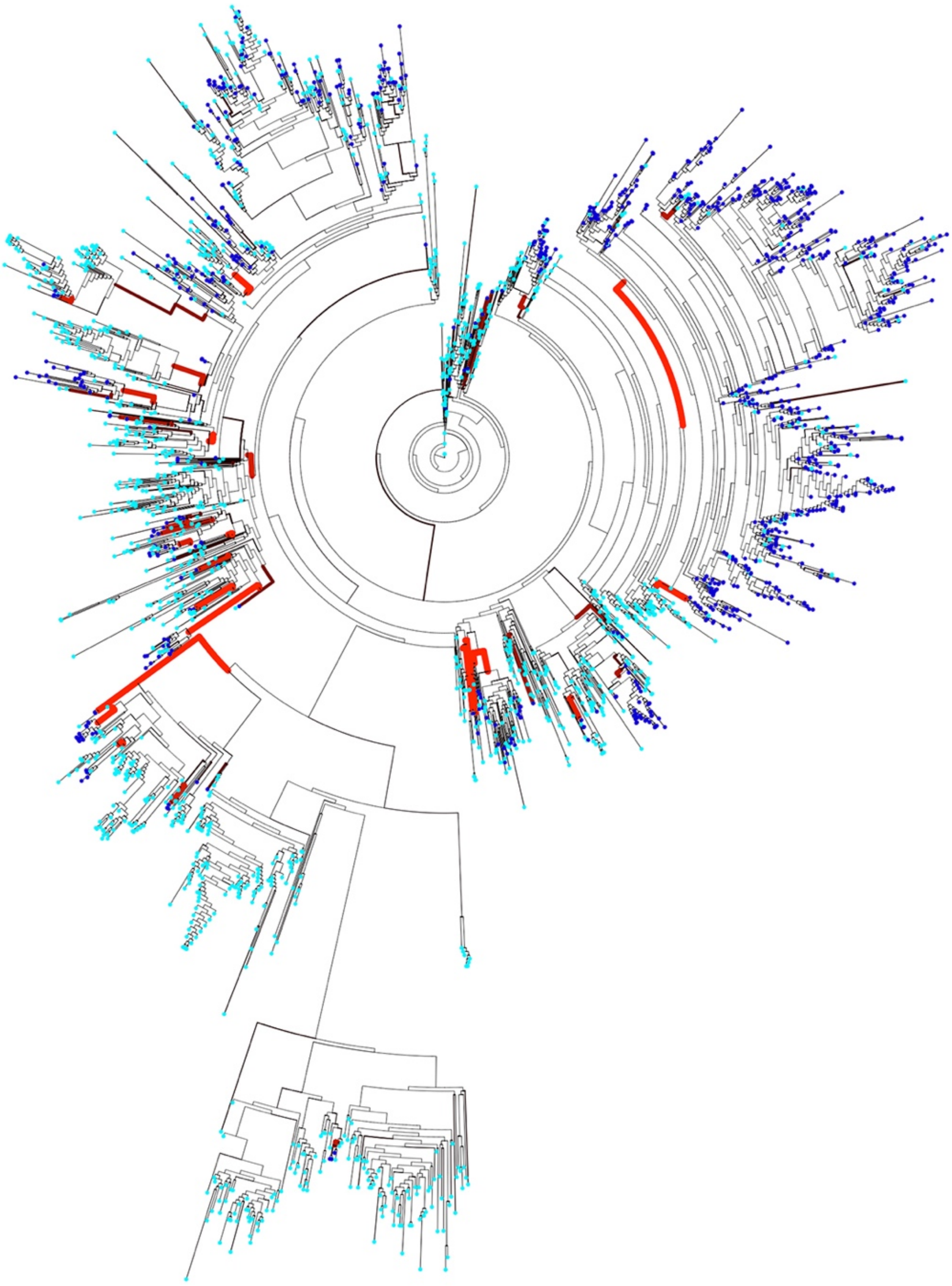


Figure V.9 Maximum likelihood phylogeny of core IGRs annotated with distinct phenotype distribution
Phylogeny of 2464 core IGRs from 1482 eBG31 human isolates and 945 eBG31 poultry isolates.
The branch thickness and redness is in proportion to the posterior probability (p > 0.5) that a change of isolation source has occurred.

Human 🟦 Poultry 🟦

VI Appendix VI. Electronic Appendices

VI.I Shorten node names for Prokka

“shorten_node_names.pl”

When given a list of names of fastas to shorten, this shortens all nodes names within each fasta, renaming the output to include ‘_shortcontigs’.

VI.II Get Representatives from SnapperDB script

“get_eBG31_representitives.py”

This script, when given a SnapperDB name, username and password, creates a list of all sequences in the database and a list of representatives for each 50SNP cluster in the database (Ashton *et al.*, 2017)

VI.III Metadata and Results

“Infantis_Metadata_and_Results.xlsx”

Table VI.1 *S*. Infantis Metadata

Continent, source group, year group and ST results for the *S*. Infantis collection (n=4670).

Table VI.2 *S*. Infantis Fastbaps Clusters

Fastbaps cluster results for the eBG31 (n=4485) and eBG297 (n=183) isolates.
. = isolate belongs to other eBG

Table VI.3 *S*. Infantis AMR Results

Antimicrobial resistance determinant presence for each of the 11 antimicrobial classes in the *S*. Infantis collection (n=4670).

Table VI.4 *S*. Infantis Plasmid Results

Plasmid group presence and pESI variants in the *S*. Infantis collection (n=4670).

Table VI.5 *S*. Infantis Integron Results

Number of integrons present in each *S*. Infantis isolate (n=4570).

VI.IV eBG31 phylogeny

“Annotated_eBG31_Phylogeny.pdf”

Figure VI.1 eBG31 phylogeny annotated with source, continent, year and ST

Soft-core SNP Maximum Likelihood Phylogeny of 831 25SNP cluster representatives of eBG31. Rings 3-6 represent the percentage of isolates in each 25SNP cluster that were isolated from each source, continent, year group and ST.

Inner ring, Number of sequences in 25SNP cluster: ⬜ 1 ⬜ 2-5 ⬛ 6-20 ⬛ 21-50 ⬛ >50

Second ring, fastbaps cluster: 1🟥 2🟦 3🟦 4🟪 5🟧 6🟩

Third ring, Source: Environmental🟦 Human🟦 Poultry🟦 Unknown⬜

Fourth ring, Continent: Africa🟪 Asia🟦 Europe🟥 N. America🟨 S. America🟦 Unknown⬜

Fifth ring, Year: 1989-2005🟪 2006-2010🟩 2011-2014🟦
2015-2016🟧 2017-2018⬜ Unknown⬜

Outer ring, ST: 2146🟨 2283🟥 32🟧 Other🟥

VI.V Significantly Associated Virulence Factors

“Associated_Virulence_Factors.xlsx”

Virulence factors identified in a comparison of 1,519 human and 947 poultry eBG31 isolates that were significantly associated with either isolation source

VI.VI Genes Present in Significantly Associated Unitigs

“Associated_Unitigs_Region1.gff, Associated_Unitigs_Region2.gff, Associated_Unitigs_Region3.gff”

Prokka outputs for the 3 regions of the eBG31 reference that contained high densities of unitigs significantly associated to an isolation source

VII Appendix VII. Publications Arising from this Thesis

VII.1 Posters

07.03.18, Global Diversity of *Salmonella* Infantis. HPRU GI Annual Conference, Norwich.

Authors: Jennifer Mattock[1], Marie Anne Chattaway[2], Hassan Hartman[2], Gemma Langridge[3], Paul Hunter[1], and John Wain[3]

20.03.18, Global Diversity of *Salmonella* Infantis. Public Health Research and Science Conference, Warwick.

Authors: Jennifer Mattock[1], Marie Anne Chattaway[2], Hassan Hartman[2], Tim Dallman[2], Gemma Langridge[3], and John Wain[3]

14.03.19, Global Diversity of *Salmonella* Infantis. HPRU GI Annual Conference, London. I won an award for the best overall poster at this conference.

Authors: Jennifer Mattock[1], Marie Anne Chattaway[2], Emma Manners[4], Hassan Hartman[2], Tim Dallman[2], Tina Duze[5], Shannon Smouse[6], Nomsa Tau[6], Karen Keddy[5], Anthony Smith[6], Gemma Langridge[3] and John Wain[3]

VII.2 Oral Presentations

01.03.17, "Reducing the Risk of Emerging Gastrointestinal Infection". HPRU GI Annual Conference, Liverpool.

03.10.17, "Impact of the chicken gut microbiota on *Salmonella* colonisation of the chicken caecum". Molecular Characterization of Foodborne and Waterborne Pathogens, including Whole-Genome Sequencing Analysis of Pathogen, South Africa.

17.11.17, "Diversity of Salmonella Infantis in England and Wales". GBRU, PHE.

28.06.18, "Global Diversity of *Salmonella* Infantis". GBRU, PHE.

25.09.18, "Global Diversity of *Salmonella* Infantis". International Symposium on *Salmonella* and Salmonellosis, France.

VII.III Journal Articles

Future publication: "Distinct genetic phylogeny in human *Salmonella* Infantis from South Africa and the United Kingdom: implications for management."

Authors: Jennifer Mattock[1], Marie Anne Chattaway[2], Karen Keddy[5], Hassan Hartman[2], Tim Dallman[2], Anthony Smith[6], Emma J. Manners[4], Oby Enwo[1], Tina Duze[5], Shannon Smouse[6], Nomsa Tau[6], Alison E. Mather[3], John Wain[3] and Gemma Langridge[3]

Publication in preparation for submission to Microbial Genomics: "Identification of a pESI-like plasmid and presence of multi-drug resistant clones found in the *S.* Infantis UK population"

Authors: Jennifer Mattock*[1], Winnie Lee*[2,7], David Greig[2,8], Gemma Langridge[3], Samuel Bloomfield[3], Alison Mather[3], Andrew Edwards[7], John Wain[3], Hassan Hartman[2], Tim Dallman[2], Marie Anne Chattaway[2], Satheesh Nair[2]
*Joint first co-authors

Location: [1]Norwich Medical School, University of East Anglia, UK. [2]Gastrointestinal Bacteriology Reference Unit, Public Health England, UK. [3]Quadram Institute Bioscience, Norwich, UK. [4]European Bioinformatics Institute, UK. [5]Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. [6]Centre for Enteric Diseases, National Institute for Communicable Disease, Johannesburg, South Africa. [7]Imperial College London, London, UK. [8]University of Edinburgh, Edinburgh, UK.

Abbreviations

AMR	antimicrobial resistance
APHA	Animal and Plant Health Agency
CDC	Centers for Disease Control and Prevention
CSF	cerebrospinal fluid
CSV	Comma-separated value
DEFRA	Department for Environment, Food & Rural Affairs
DLV	double-locus variant
DNA	deoxyribonucleic acid
eBG	eBurstGroup
ECDC	European Centre for Disease Prevention and Control
EFSA	European Food Safety Authority
ENA	European Nucleotide Archive
ESBLs	extended-spectrum beta-lactamases
EU	European Union
FDA	Food and Drug Administration
FSA	Food Standards Agency
GBRU	Gastrointestinal Bacterial Reference Unit
GWAS	Genome wide association studies
HIV	human immunodeficiency virus
hPHE	historical Public Health England
HUS	haemolytic uraemic syndrome
IGR	intergenic region
IID	infectious intestinal disease
IL	Interleukin
Inc	incompatibility
Integron A	The integron in the ASRF01000104.1_contig_55 contig of the pESI assembly
Integron B	The integron in the ASRF01000099.1_contig_4 contig of the pESI assembly
iNTS	invasive non-typhoidal salmonellosis
MDR	Multidrug resistance
MLST	Multi-locus sequence typing

MLV	multiple-locus variant
MRCA	most recent common ancestor
NCEZID	National Center for Emerging and Zoonotic Infectious Diseases
ncRNAs	non-coding ribonucleic acids
NICD	National Institute for Communicable Diseases
NTS	Non-Typhoidal Salmonella
PCR	polymerase chain reaction
pESI	plasmid of emerging S. Infantis
PHE	Public Health England
QIB	Quadram Institute Bioscience
QRDR	quinolone resistance determining region
Rep	Replicon
SLV	single-locus variant
SNP	single nucleotide polymorphism
SPI	*Salmonella* Pathogenicity Island
SPV	*Salmonella* plasmid virulence
SRA	Sequence Read Archive
ST	Sequence Type
STEC	Shiga toxin-producing *E. coli*
STFP	Secure File Transfer Protocol
STM	signature tagged mutagenesis
subsp.	subspecies
T3SS	Type III Secretion System
TMDH	transposon-mediated differential hybridisation
TraDIS	transposon directed insertion-site sequencing
TraSH	transposon-site hybridisation
UEA	University of East Anglia
UK	United Kingdom
USA	United States of America
WHO	World Health Organization

Glossary of Terms

Basecalling – The process of interpreting the output of a sequencing run as nucleotides.

Contig – Contiguous sections of the genome (Narzisi and Mishra, 2011).

Crunch file – The format used by software such as ACT containing information comparing genomes (Carver *et al.*, 2005).

De Bruijn Graph – A method used by genome assemblers to resolve differences between reads. Reads of length k are plotted onto the graph, when they diverge this is plotted as a node and edges are plotted until the k-mers converge again. The assembled genome is determined by walking through the graph.

Demultiplexing – To allow multiple isolates to be sequenced at once, each sample is labelled with unique indexes. Demultiplexing is the process of pulling out the sequence data out for each isolate from the sequencing run output.

Dcm methylation – DNA methylation of the second cytosine of CC(A/T)GG sequences (Gomez-Eichelmann, Levy-Mustri and Ramirez-Santos, 1991).

eBurstGroup (eBG) - defined as multi locus sequence types (ST) linked by single locus variants (Achtman *et al.*, 2012).

eBG31 collection – All eBG31 sequences that successfully passed into SnapperDB.

Indel – An indel is a short insertion or deletion of nucleotides (Sehn, 2015).

k-mer – A sequence of length k.

Manhattan plot – A scatter plot commonly used to present GWAS results. The x-axis represents the SNPs position in the genome and the y-axis shows the negative log of the p-value for each SNP. The lowest p-values appear at the top of the graph (Modena *et al.*, 2019).

Multidrug Resistance (MDR) – Resistance to at least 3 classes of antimicrobials

N50 – The number at which the combined length of sequences greater than this number makes up at least 50% of the whole genome (Narzisi and Mishra, 2011).

pESI pseudomolecule – The eBG31 reference genome concatenated with the pESI reference.

Properly paired reads – When read pairs are aligned with a distance equal to the distance between the ends of the reads, they are properly paired (Thankaswamy-Kosalai, Sen and Nookaew, 2017).

QC max percentage non consensus base of each locus – An output of *MOST* showing the highest percentage of non-consensus bases in each allele (Tewolde *et al.*, 2016) .

Scaffolds – Contigs that have been ordered and spaced by N's, predicting their orientation in the genome.

Scree plot – a plot of the results of a principal component analysis, a multivariate statistical test where data is split into components showing variation, measured in eigenvalues (Abdi and Williams, 2010; Lewith, Jonas and Walach, 2010).

SNP address - SnapperDB calculates and stores the distances between all sequences added to the database (Ashton *et al.*, 2017). It clusters the sequences on seven levels of SNP distance: 250, 100, 50, 25, 10, 5 and 0. For a sequence to be added to a cluster it needs to be within that SNP distance of any isolate in the cluster. The clusters that each sequence belongs to are used to give it a seven-digit code. This SNP address is then used to provide real-time clustering of sequences and identify outbreaks.

Tagmentation – A step in the library prep reaction which uses a transposon to cleave the DNA and adds a primer to each piece of double stranded DNA (Illumina, 2015).

Unitig – A compacted De Bruijn graph is made using all the genomes in the association study (Jaillard *et al.*, 2018). A node on the graph is a unitig and represents a sequence that is shared by genomes.

UpSet Plot – An alternative to a Venn diagram, each row in the plot represents a segment of a Venn diagram, with the value of each intersection shown as a bar chart (Lex *et al.*, 2014).

Software and Hardware Glossary

Albacore (Oxford Nanopore Technologies) – An Oxford Nanopore Technologies (ONT) basecaller that identifies the DNA sequences from the raw output of a MinION sequencing run.

awk (Dougherty and Robbins, 1997) – a Unix programming language for the manipulation of text files.

ape (Paradis and Schliep, 2018) - An R package containing tools to work with phylogenies.

ARIBA (Hunt *et al.*, 2017) – Identifies the presence of genes in fastq files from a selection of databases, with the option to create a database.

Artemis Comparison Tool (ACT) (Carver *et al.*, 2005) – A java based tool to compare two or more DNA sequences and visualise the similarities.

Artemis (Carver *et al.*, 2012) – Allows visualisation of genomes, including annotation files.

Bandage (Wick *et al.*, 2015) – Allows visualisation of the assembly graphs produced by assemblers.

bcl2fastq (Illumina, 2019b) – Converts the bcl output of an Illumina sequencing run into fastq files.

BRIG (Alikhan *et al.*, 2011) – Uses BLAST to generate images showing the similarity of multiple prokaryotic genomes to a reference.

BWA (Li, 2013) – Burrows-Wheeler Aligner, a tool that maps sequences against a reference sequence. BWA-MEM is used to map sequences with reads, varying from 70bp-1Mbp in length, faster and more accurately than the other algorithm for that read length, BWA-SW.

Canu (Koren *et al.*, 2017) – A long read assembler that uses adaptive k-mer weighting and can accurately assemble large repeats.

Circlator (Hunt *et al.*, 2015) – A tool that circularises assemblies.

data.table (Dowle and Srinivasan, 2018) – An R package that enables reading and manipulation of large data.

devtools (Wickham *et al.*, 2018) – An R package that facilitates R package development and installation of other packages.

Entrez Direct (Kans, 2019) – Gives access to the NCBI's database via command line.

Fastbaps (Tonkin-Hill *et al.*, 2019) – Provides hierarchical clustering of sequence data similar to hierBAPS, but much faster.

FastQC (Andrews, 2010) – Provides information on the quality of raw sequence data, producing an overview and summary graphs.

Genome Analysis Toolkit (GATK) (Van der Auwera *et al.*, 2013) – A variant caller for high-throughput sequencing data.

ggplot2 (Wickham, 2016) – An R package for creating graphics.

grep (Dougherty and Robbins, 1997) – a Unix command for displaying and editing lines of text.

Gubbins (Nicholas J Croucher *et al.*, 2015) – Identifies recombination within an alignment.

hierBAPS (Cheng *et al.*, 2013) – Determines clustering of DNA sequences using a hierarchical approach of Bayesian inference.

Integron Finder (Cury *et al.*, 2016a) – Identifies integrons in bacterial genomes.

iTOL (Letunic and Bork, 2016) - An online tool for visualisation and annotation of phylogenies.

Mash (Ondov *et al.*, 2016) – Calculates the genetic distance between sequence data by creating MinHash sketches and determining the proportion of k-mers shared.

MATLAB (MathWorks, 2014) – A programming platform for data analysis.

MEGA7 (Kumar, Stecher and Tamura, 2016) – A graphical user interface that analyses sequence data, producing alignments, phylogenies and calculating molecular evolution.

MegaBLAST (McGinnis and Madden, 2004) - Searches a nucleotide query against nucleotide databases, efficiently finding long alignments between highly similar sequences.

Metric Oriented Sequence Typer (MOST) (Tewolde *et al.*, 2016) – Assigns ST profile.

MUMmer dnadiff (Kurtz *et al.*, 2004) – Quantifies the differences between two genomes.

NanoFilt (De Coster *et al.*, 2018) – Trims and filters the reads of long read sequencing data from ONT sequencing runs.

Nanopolish (Loman, Quick and Simpson, 2015) – When provided with Oxford Nanopore fast5 files and a draft assembly it produces a consensus sequence.

NanoStat (De Coster *et al.*, 2018) – Produces summary statistics of long read sequencing data from ONT sequencing runs.

Nucleotide BLAST (McGinnis and Madden, 2004) – Searches a nucleotide query against nucleotide databases.

PHASTER (Arndt *et al.*, 2016) – Identifies and annotates prophage sequences in bacterial genomes.

PHEnix (Jironkin *et al.*, 2017) – The SNP calling pipeline used by PHE. It takes paired end fastqs as input, maps to a reference genome, variant calls, and filters the resulting variant call format (VCF) file.

phylobase (Michonneau *et al.*, 2019) – An R package that allows manipulation of phylogenies.

PHYLOViZ (Ribeiro-Gonçalves *et al.*, 2016) – Generates minimum spanning trees of allelic data.

phytools (Revell, 2012) – An R package containing tools to work with phylogenies.

Piggy (Thorpe *et al.*, 2018) – Uses the output of Roary to analyse the intergenic regions in bacterial genomes.

Pilon (Walker *et al.*, 2014) – A tool for improving draft assemblies by correcting bases and variant detection using read alignment analysis.

PlasmidFinder (Carattoli *et al.*, 2014) - An online tool that identifies the presence of plasmids in uploaded sequences. The database can be downloaded and used as input for ARIBA.

Prokka (Seemann, 2014) – Performs prokaryote genome annotation.

Protein BLAST (McGinnis and Madden, 2004) – Searches an protein query against protein databases.

pyseer (Lees *et al.*, 2018) – Estimates the genetic variation within a bacterial population associated with a phenotype; taking into account the effect of population structure.

QUAST (Gurevich *et al.*, 2013) – A tool to assess the quality of assemblies.

R (R Core Team, 2018) – A programming language for statistical computing and graphics.

Racon (Vaser *et al.*, 2017) – Corrects assemblies using alignments of the assembly to either short or long read sequence data.

RAxML (Stamatakis, 2014) – Produces a maximum likelihood phylogeny from an alignment.

ResFinder (Zankari *et al.*, 2012) – An online tool that identifies the presence of AMR genes in uploaded sequences. The database can be downloaded and used as input for ARIBA.

rhierBAPS (Tonkin-Hill *et al.*, 2018) – An r package of the hierBAPS software.

Roary (Page *et al.*, 2015) – Produces the pan genome from annotated assemblies.

RStudio (RStudio, 2018) – An environment for R.

SAMtools (Li *et al.*, 2009) – Has various functions to process alignments in the SAM format. SAMtools sort is used to sort alignment, SAMtools index to index them and

SAMtools depth calculates the depth at each position. SAMtools flagstat prints statistics for the alignment.

Scoary (Brynildsrud *et al.*, 2016) – Uses the output of Roary to calculate the association between traits and genes.

sed (Dougherty and Robbins, 1997) – A Unix stream editor, used for editing multiple text files.

SMALT (Ponstingl and Ning, 2014) – Maps sequence data to a reference genome.

SnapperDB (Ashton *et al.*, 2017) – A database that stores a pairwise distance matrix of SNP distances. This can be used to produce alignments and a SNP address for each isolate.

SPAdes (Bankevich *et al.*, 2012) – A genome assembler.

SRA Toolkit fastq-dump (National Center for Biotechnology Information, 2014) – Allows access and conversion of the data in the SRA to fastq format.

Tablet (Milne *et al.*, 2013) – Allows visualisation of assemblies and alignments.

The Cloud Infrastructure for Microbial Bioinformatics (CLIMB) (Connor *et al.*, 2016) – A high performance computing cluster for UK microbial bioinformaticians.

TreeBreaker (Azim and Didelot, 2018) – With a phylogeny and phenotype information it uses Bayesian inference to determine whether phenotypes are distributed evenly across the phylogeny and if branches are associated with a phenotype.

Trimmomatic (Bolger, Lohse and Usadel, 2014) – A pre-processing tool that trims and filters reads from either single end or paired end sequence data.

Unicycler (Wick *et al.*, 2017) – A tool that assembles bacterial genomes using a combination of long read and short read sequence data.

unitig-counter (Jaillard *et al.*, 2018; Lees, 2019) - Takes assemblies as input and counts the unitigs in a bacterial population using a compressed De-Bruijn graph. The output is used as input for pyseer.

VFDB (Chen *et al.*, 2016) – An online database containing information about virulence factors of bacterial pathogens. The database can be downloaded and used as input for ARIBA.

References

Abdi, H. and Williams, L. J. (2010) ‘Principal component analysis’, *WIREs Computational Statistics* . John Wiley & Sons, Ltd, 2(4), pp. 433–459. doi: 10.1002/wics.101.

Acar, S. *et al.* (2019) ‘Genome analysis of antimicrobial resistance, virulence, and plasmid presence in Turkish *Salmonella* serovar Infantis isolates’, *International Journal of Food Microbiology*, 307. doi: https://doi.org/10.1016/j.ijfoodmicro.2019.108275.

Achtman, M. *et al.* (2012) ‘Multilocus sequence typing as a replacement for serotyping in Salmonella enterica’, *PLoS Pathogens*, 8(6). doi: 10.1371/journal.ppat.1002776.

Afema, J. A., Mather, A. E. and Sischo, W. M. (2015) ‘Antimicrobial Resistance Profiles and Diversity in *Salmonella* from Humans and Cattle , 2004 – 2011’, *Zoonoses and Public Health*, pp. 506–517. doi: 10.1111/zph.12172.

Akiba, M. *et al.* (2007) ‘Changes in antimicrobial susceptibility in a population of *Salmonella enterica* serovar Dublin isolated from cattle in Japan from 1976 to 2005’, *Journal of Antimicrobial Chemotherapy*, 60(6), pp. 1235–1242. doi: 10.1093/jac/dkm402.

Akullian, A. *et al.* (2018) ‘Multi-drug resistant non-typhoidal *Salmonella* associated with invasive disease in western Kenya’, *PLOS Neglected Tropical Diseases*. Public Library of Science, 12(1). Available at: https://doi.org/10.1371/journal.pntd.0006156.

Alali, W. Q. *et al.* (2010) ‘Prevalence and Distribution of *Salmonella*’, *Foodborne Pathogens and Disease*, 7(11). doi: 10.1089/fpd.2010.0566.

Ali, M. M. *et al.* (2014) ‘Fructose-Asparagine is a Primary Nutrient during Growth of *Salmonella* in the Inflamed Intestine’, *PLOS Pathogens*. Public Library of Science, 10(6). Available at: https://doi.org/10.1371/journal.ppat.1004209.

Alikhan, N. *et al.* (2011) ‘BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons.’, *BMC genomics*. England, 12, p. 402. doi: 10.1186/1471-2164-12-402.

Alikhan, N. *et al.* (2018) ‘A genomic overview of the population structure of *Salmonella*’, *PLOS Genetics*. Public Library of Science, 14(4). Available at: https://doi.org/10.1371/journal.pgen.1007261.

Almeida, F. *et al.* (2013) ‘Molecular epidemiology and virulence markers of *Salmonella* Infantis isolated over 25years in São Paulo State, Brazil’, *Infection, Genetics and Evolution*. Elsevier B.V., 19, pp. 145–151. doi: 10.1016/j.meegid.2013.07.004.

Andoh, L. A. *et al.* (2016) ‘Prevalence and antimicrobial resistance of *Salmonella* serovars isolated from poultry in Ghana’, *Epidemiol. Infect.*, 144, pp. 3288–3299.

Andrews-Polymenis, H. L., Santiviago, C. A. and McClelland, M. (2009) ‘Novel genetic tools for studying food-borne *Salmonella*', *Current Opinion in Biotechnology*, 20(2), pp. 149–157. doi: https://doi.org/10.1016/j.copbio.2009.02.002.

Andrews, S. (2010) *FastQC: a quality control tool for high throughput sequence data.* Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed: 12 July 2019).

Animal and Plant Health Agency (2017) *About us*. Available at: https://www.gov.uk/government/organisations/animal-and-plant-health-agency/about (Accessed: 16 August 2019).

Ansari, M. A. and Didelot, X. (2016) 'Bayesian Inference of the Evolution of a Phenotype Distribution on a Phylogenetic Tree', *Genetics*. Genetics, 204(1), pp. 89–98. doi: 10.1534/genetics.116.190496.

Ao, T. T. *et al.* (2015) 'Global Burden of Invasive Nontyphoidal *Salmonella* Disease, 2010', *Emerging Infectious Disease journal*, 21(6). doi: 10.3201/eid2106.140999.

Arndt, D. *et al.* (2016) 'PHASTER: a better, faster version of the PHAST phage search tool.', *Nucleic acids research*. England, 44(W1). doi: 10.1093/nar/gkw387.

Asgharpour, F. *et al.* (2014) 'Investigation of Class I Integron in *Salmonella* Infantis and Its Association With Drug Resistance', *Jundishapur J Microbiol.*, 7(5), pp. 1–5. doi: 10.5812/jjm.10019.

Ashton, P. *et al.* (2017) 'SnapperDB: A database solution for routine sequencing analysis of bacterial isolates', *Bioinformatics*. doi: 10.1101/033225.

Asten, A. J. A. M. Van and Dijk, J. E. Van (2005) 'Distribution of "classic" virulence factors among *Salmonella* spp.', *FEMS Immunology and Medical Microbiology*, 44, pp. 251–259. doi: 10.1016/j.femsim.2005.02.002.

Van der Auwera, G. A. *et al.* (2013) 'From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.', *Current protocols in bioinformatics*. United States, 43. doi: 10.1002/0471250953.bi1110s43.

Aviv, G. *et al.* (2014) 'A unique megaplasmid contributes to stress tolerance and pathogenicity of an emergent *Salmonella enterica* serovar Infantis strain', *Environmental Microbiology*, 16(4), pp. 977–994. doi: 10.1111/1462-2920.12351.

Aviv, G. *et al.* (2019) 'The emerging *Salmonella* Infantis expresses lower levels of SPI-1 genes and causes milder colitis in mice and lower rates of invasive disease in humans than *Salmonella* Typhimurium', *Infectious Diseases Society of America*, pp. 1–27.

Aviv, G., Rahav, G. and Gal-Mor, O. (2016) 'Horizontal Transfer of the *Salmonella enterica* Serovar Infantis Resistance and Virulence Plasmid pESI to the Gut Microbiota of Warm-

Blooded Hosts', *American Society for Microbiology*, 7(5), pp. 1–12. doi: 10.1128/mBio.01395-16.Editor.

Azim, M. and Didelot, X. (2018) *TreeBreaker*. Available at: https://github.com/ansariazim/treeBreaker (Accessed: 10 June 2019).

Bankevich, A. *et al.* (2012) 'SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing', *Journal of computational biology : a journal of computational molecular cell biology*. Mary Ann Liebert, Inc., 19(5), pp. 455–477. doi: 10.1089/cmb.2012.0021.

Barak, J. D. and Liang, A. S. (2008) 'Role of soil, crop debris, and a plant pathogen in *Salmonella enterica* contamination of tomato plants.', *PloS one*. United States, 3(2). doi: 10.1371/journal.pone.0001657.

Barrow, P. A. *et al.* (1987) 'Observations on the pathogenesis of experimental *Salmonella* Typhimurium infection in chickens', *Research in Veterinary Science*, 42, pp. 194–199.

Barquist, L. *et al.* (2013) 'A comparison of dense transposon insertion libraries in the *Salmonella* serovars Typhi and Typhimurium', *Nucleic Acids Research*, 41(8), pp. 4549–4564. doi: 10.1093/nar/gkt148.

Barua, S. *et al.* (2002) 'Involvement of surface polysaccharides in the organic acid resistance of Shiga Toxin-producing *Escherichia coli* O157:H7', *Molecular Microbiology*. John Wiley & Sons, Ltd, 43(3), pp. 629–640. doi: 10.1046/j.1365-2958.2002.02768.x.

Barza, M. and Travers, K. (2002) 'Excess Infections Due to Antimicrobial Resistance: The "Attributable Fraction"', *Clinical Infectious Diseases*, 34(Supplement_3). doi: 10.1086/340250.

Behravesh, C. B. *et al.* (2010) 'Human *Salmonella* infections linked to contaminated dry dog and cat food, 2006-2008.', *Pediatrics*. United States, 126(3), pp. 477–483. doi: 10.1542/peds.2009-3273.

Benson, D. A. *et al.* (2005) 'GenBank', *Nucleic acids research*. Oxford University Press, 33(Database issue). doi: 10.1093/nar/gki063.

Blaser, M. J. and Feldman, R. A. (1981) '*Salmonella* Bacteremia: Reports to the Centers for Disease Control, 1968-1979', *The Journal of Infectious Diseases*. Oxford University Press, 143(5), pp. 743–746. Available at: http://www.jstor.org/stable/30113298.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: a flexible trimmer for Illumina sequence data.', *Bioinformatics (Oxford, England)*. England, 30(15), pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.

Borck Høg, B. *et al.* (2018) *DANMAP 2017 - Use of antimicrobial agents and occurrence of* *antimicrobial resistance in bacteria from food animals, food and humans in Denmark*. Available at: https://www.danmap.org/-/media/arkiv/projekt-sites/danmap/danmap-reports/danmap-2017/danmap2017.pdf?la=en (Accessed: 01 December 2019).

Boucher, Y. *et al.* (2007) 'Integrons: mobilizable platforms that promote genetic diversity in bacteria', *Trends in Microbiology*, 15(7), pp. 301–309. doi: https://doi.org/10.1016/j.tim.2007.05.004.

Brandis, G. and Hughes, D. (2016) 'The Selective Advantage of Synonymous Codon Usage Bias in *Salmonella'*, *PLOS Genetics*. Public Library of Science, 12(3), pp. 1–16. doi: 10.1371/journal.pgen.1005926.

Briones, V. *et al.* (2004) '*Salmonella* diversity associated with wild reptiles and amphibians in Spain', *Environmental Microbiology*. John Wiley & Sons, Ltd (10.1111), 6(8), pp. 868–871. doi: 10.1111/j.1462-2920.2004.00631.x.

Brown, A. C. *et al.* (2018) 'CTX-M-65 Extended-Spectrum *Salmonella enterica* Serotype Infantis, United States', *Emerging Infectious Diseases*, 24(12).

Brynildsrud, O. *et al.* (2016) 'Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary', *Genome Biology*, 17(1). doi: 10.1186/s13059-016-1108-8.

Buchwald, D. S. and Blaser, M. J. (1984) 'A Review of Human Salmonellosis: II. Duration of Excretion Following Infection with Nontyphi *Salmonella*', *Reviews of Infectious Diseases*, 6(3), pp. 345–356.

Buncic, S. and Sofos, J. (2012) 'Interventions to control *Salmonella* contamination during poultry, cattle and pig slaughter', *Food Research International*, 45(2), pp. 641–655. doi: https://doi.org/10.1016/j.foodres.2011.10.018.

Burke, L. *et al.* (2013) 'Resistance to third-generation cephalosporins in human non-typhoidal *Salmonella enterica* isolates from England and Wales, 2010–12', *Journal of Antimicrobial Chemotherapy*, 69(4), pp. 977–981. doi: 10.1093/jac/dkt469.

Byrne, C. M., Clyne, M. and Bourke, B. (2007) '*Campylobacter jejuni* adhere to and invade chicken intestinal epithelial cells in vitro', *Microbiology*, 153(2), pp. 561–569. doi: 10.1099/mic.0.2006/000711-0.

Cabezón, E. *et al.* (2014) 'Towards an integrated model of bacterial conjugation', *FEMS Microbiology Reviews*, 39(1), pp. 81–95. doi: 10.1111/1574-6976.12085.

Carattoli, A. *et al.* (2014) 'In Silico Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing', *Antimicrobial Agents and Chemotherapy*. American Society for Microbiology Journals, 58(7), pp. 3895–3903. doi:

10.1128/AAC.02412-14.

Carattoli, A. and Elena, V. R. (2009) ‘Resistance Plasmid Families in *Enterobacteriaceae*’, *Antimicrobial Agents and Chemotherapy*, 53(6), pp. 2227–2238. doi: 10.1128/AAC.01707-08.

Card, R. *et al.* (2016) ‘Virulence Characterisation of *Salmonella enterica* Isolates of Differing Antimicrobial Resistance Recovered from UK Livestock and Imported Meat Samples’, 7(May), pp. 1–11. doi: 10.3389/fmicb.2016.00640.

Carfora, V. *et al.* (2018) ‘Colistin Resistance Mediated by mcr-1 in ESBL-Producing, Multidrug Resistant *Salmonella* Infantis in Broiler Chicken Industry, Italy (2016-2017)’, *Frontiers in microbiology*. Frontiers Media S.A., 9(1880). doi: 10.3389/fmicb.2018.01880.

Cartelle Gestal, M. *et al.* (2016) ‘Characterization of a small outbreak of *Salmonella enterica* serovar Infantis that harbour CTX-M-65 in Ecuador’, *Brazilian Journal of Infectious Diseases*. Elsevier Editora Ltda, 20(4), pp. 406–407. doi: 10.1016/j.bjid.2016.03.007.

Carver, T. *et al.* (2012) ‘Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data.’, *Bioinformatics (Oxford, England)*, 28(4), pp. 464–9. doi: 10.1093/bioinformatics/btr703.

Carver, T. J. *et al.* (2005) ‘ACT : the Artemis comparison tool’, *Bioinformatics*, 21(16), pp. 3422–3423. doi: 10.1093/bioinformatics/bti553.

Center for Algorithmic Biotechnology (2013) *Quast*. Available at: http://quast.bioinf.spbau.ru (Accessed: 8 March 2018).

Centers for Disease Control and Prevention (2013) *Infection with Salmonella*, *Salmonella in the Caribbean*. Available at: https://www.cdc.gov/training/SIC_CaseStudy/Infection_Salmonella_ptversion.pdf (Accessed: 2 February 2020).

Centers for Disease Control and Prevention (2017) *Foodborne Diseases Active Surveillance Network (FoodNet): FoodNet 2015 Surveillance Report (Final Data)*.

Centers for Disease Control and Prevention (2018) *Estimates of Foodborne Illness in the United States*. Available at: https://www.cdc.gov/foodborneburden/2011-foodborne-estimates.html (Accessed: 12 June 2019).

Centers for Disease Control and Prevention (2019) *Outbreak of Multidrug-Resistant Salmonella Infections Linked to Raw Chicken Products*. Available at: https://www.cdc.gov/salmonella/infantis-10-18/index.html (Accessed: 20 August 2019).

Centre for Disease Control and Prevention (CDC) (2005) ‘Outbreaks of *Escherichia coli*

O157:H7 associated with petting zoos--North Carolina, Florida, and Arizona, 2004 and 2005.’, *MMWR. Morbidity and mortality weekly report*. United States, 54(50), pp. 1277–1280.

Chambers, J. R. and Gong, J. (2011) ‘The intestinal microbiota and its modulation for *Salmonella* control in chickens’, *Food Research International*. Guelph Food Research Centre, Agriculture and Agri-Food Canada, 93 Stone Road West, Guelph, ON N1G 5C9, Canada, 44(10), pp. 3149–3159. doi: 10.1016/j.foodres.2011.08.017.

Chaudhuri, R. R., Allen, A. G., *et al.* (2009) ‘Comprehensive identification of essential *Staphylococcus aureus* genes using Transposon-Mediated Differential Hybridisation (TMDH)’, *BMC genomics*. BioMed Central, 10, p. 291. doi: 10.1186/1471-2164-10-291.

Chaudhuri, R. R., Peters, S. E., *et al.* (2009) ‘Comprehensive Identification of *Salmonella enterica* Serovar Typhimurium Genes Required for Infection of BALB/c Mice’, *PLOS Pathogens*. Public Library of Science, 5(7). Available at: https://doi.org/10.1371/journal.ppat.1000529.

Chen, H. M. *et al.* (2013) ‘Nontyphoid *Salmonella* infection: Microbiology, clinical features, and antimicrobial therapy’, *Pediatrics and Neonatology*. Elsevier Taiwan LLC, 54(3), pp. 147–152. doi: 10.1016/j.pedneo.2013.01.010.

Chen, L. *et al.* (2016) ‘VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on.’, *Nucleic acids research*. England, 44(D1). doi: 10.1093/nar/gkv1239.

Cheng, L. *et al.* (2013) ‘Hierarchical and spatially explicit clustering of DNA sequences with BAPS software’, *Molecular biology and evolution*. 2013/02/13. Oxford University Press, 30(5), pp. 1224–1228. doi: 10.1093/molbev/mst028.

Cheng, Y. *et al.* (2015) ‘rpoS-Regulated core genes involved in the competitive fitness of *Salmonella enterica* Serovar Kentucky in the intestines of chickens’, *Applied and environmental microbiology*. 2014/10/31. American Society for Microbiology, 81(2), pp. 502–514. doi: 10.1128/AEM.03219-14.

Cherubin, C. E. *et al.* (1974) ‘Septicemia with Non-Typhoid *Salmonella*’, *Medicine*, 53(5), pp. 365–376.

Cloeckaert, A. *et al.* (2007) ‘Dissemination of an extended-spectrum-beta-lactamase blaTEM-52 gene-carrying IncI1 plasmid in various *Salmonella enterica* serovars isolated from poultry and humans in Belgium and France between 2001 and 2005’, *Antimicrobial agents and chemotherapy*. 2007/02/26. American Society for Microbiology (ASM), 51(5), pp. 1872–1875. doi: 10.1128/AAC.01514-06.

Colavecchio, A. *et al.* (2017) ‘Bacteriophages Contribute to the Spread of Antibiotic

Resistance Genes among Foodborne Pathogens of the *Enterobacteriaceae* Family – A Review', *Frontiers in Microbiology*, 8, p. 1108. doi: 10.3389/fmicb.2017.01108.

Colobatiu, L. *et al.* (2015) 'First description of plasmid-mediated quinolone resistance determinants and β-lactamase encoding genes in non-typhoidal *Salmonella* isolated from humans, one companion animal and food in Romania', *Gut Pathogens*. BioMed Central, 7(1), pp. 1–11. doi: 10.1186/s13099-015-0063-3.

Connor, T. R. *et al.* (2016) 'CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community.', *Microbial genomics*. England, 2(9). doi: 10.1099/mgen.0.000086.

Cooper, J. E. and Feil, E. J. (2004) 'Multilocus sequence typing – what is resolved?', *Trends in Microbiology*, 12(8), pp. 373–377. doi: https://doi.org/10.1016/j.tim.2004.06.003.

De Coster, W. *et al.* (2018) 'NanoPack: visualizing and processing long-read sequencing data', *Bioinformatics*, 34(15), pp. 2666–2669. doi: 10.1093/bioinformatics/bty149.

Croucher, Nicholas J *et al.* (2015) 'Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins', *Nucleic Acids Research*, 43(3). Available at: http://dx.doi.org/10.1093/nar/gku1196.

Croucher, Nicholas J. *et al.* (2015) 'Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins', *Nucleic Acids Research*, 43(3). doi: 10.1093/nar/gku1196.

Crump, J. A. and Wain, J. (2017) '*Salmonella*', in Quah, S.R (eds) *International Encyclopedia of Public Health*. 2nd edn. Oxford: Academic Press, pp. 425–433. doi:https://doi.org/10.1016/B978-0-12-803678-5.00394-5.

Cui, S. *et al.* (2005) 'Prevalence and Antimicrobial Resistance of Campylobacter spp . and *Salmonella* Serovars in Organic Chickens from Maryland Retail Stores Prevalence and Antimicrobial Resistance of *Campylobacter* spp. and *Salmonella* Serovars in Organic Chickens from Maryland', *Applied and Environmental Microbiology*, 71(7), pp. 4108–4111. doi: 10.1128/AEM.71.7.4108.

Cury, J. *et al.* (2016a) 'Identification and analysis of integrons and cassette arrays in bacterial genomes', *Nucleic Acids Research*, 44(10), pp. 4539–4550. doi: 10.1093/nar/gkw319.

Cury, J. *et al.* (2016b) *Integron Finder*. Available at: https://galaxy.pasteur.fr/root?tool_id=toolshed.pasteur.fr%2Frepos%2Fkhillion%2Fintegron_finder%2Fintegron_finder%2F1.5.1 (Accessed: 9 September 2019).

Cuypers, W. L. *et al.* (2018) ‘Fluoroquinolone resistance in *Salmonella*: insights by whole-genome sequencing’, *Microbial Genomics*, pp. 1–9. doi: 10.1099/mgen.0.000195.

Czerwiński, J. *et al.* (2012) ‘Effects of sodium butyrate and salinomycin upon intestinal microbiota, mucosal morphology and performance of broiler chickens’, *Archives of Animal Nutrition*. The Kielanowski Institute of Animal Physiology and Nutrition, Polish Academy of Sciences, Jablonna, Poland, 66(2), pp. 102–116. doi: 10.1080/1745039X.2012.663668.

Dahshan, H. *et al.* (2010) ‘Characterization of Antibiotic Resistance and the Emergence of AmpC-Producing *Salmonella* Infantis from Pigs’, *J. Vet. Med. Sci.*, 72(11), pp. 1437–1442.

Dallman, T. J. (2018) ‘Genomic Surveillance of *Salmonella* Enteritidis reveals emergence of new epidemic clone circulating in European poultry industry.’, in *International Symposium on Salmonella and salmonellosis*. Saint-Malo.

Davies, R. H. and Wray, C. (1996) ‘Persistence of *Salmonella* Enteritidis in poultry units and poultry food’, *British Poultry Science*, 37. doi: 10.1080/00071669608417889.

DEFRA (2016) *Salmonella: get your broiler flock chickens tested*. Available at: https://www.gov.uk/guidance/salmonella-get-your-broiler-flock-chickens-tested (Accessed: 19 November 2019).

DEFRA (2018) *Code of practice for the welfare of meat chickens and meat breeding chickens*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/694013/meat-chicken-code-march2018.pdf (Accessed: 19 November 2019).

DEFRA (2019a) *Livestock numbers in England and the UK*. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/603892/structure-livestockpop-UK-29Mar17.xls (Accessed: 19 November 2019).

DEFRA (2019b) *Number of poultry slaughtered per year in the UK*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/839973/poultry-slaughter-17oct19.ods (Accessed: 19 November 2019).

Dembski, T., Patynski, M. and Ciesla, P. (1995) ‘Spondylitis caused by *Salmonella* Infantis--case report’, *Chirurgia narzadow ruchu i ortopedia polska*. Poland, 60(5), pp. 365–367.

Denbow, D. M. (2000) ‘Chapter 12. Gastrointestinal Anatomy and Physiology’, in Whittow, G. C. (ed.) *Sturkie’s Avian Physiology*. 5th edn, p. 314.

Deng, Y. *et al.* (2015) ‘Resistance integrons : class 1 , 2 and 3 integrons’, *Annals of Clinical Microbiology and Antimicrobials*. BioMed Central, 14(45), pp. 1–11. doi: 10.1186/s12941-015-0100-6.

Desai, P. T. *et al.* (2013) ‘Evolutionary Genomics of *Salmonella enterica* Subspecies’, *mBio*. Edited by B. B. Finlay, 4(2). doi: 10.1128/mBio.00579-12.

Desin, T. S., Köster, W. and Potter, A. A. (2013) ‘*Salmonella* vaccines in poultry: past, present and future’, *Expert Review of Vaccines*. Taylor & Francis, 12(1), pp. 87–96. doi: 10.1586/erv.12.138.

Dhillon, R, H, P. and Clark, J. (2012) ‘ESBLs: A Clear and Present Danger?’, *Critical Care Research and Practice*, 2012.

Dougherty, D. and Robbins, A. (1997) *sed & awk: UNIX Power Tools*. 2nd edn. O’Reilly Media (Nutshell Handbooks). Available at: https://books.google.co.uk/books?id=Xu0G31e-4gIC.

Dowle, M. and Srinivasan, A. (2018) *data.table*. Available at: https://github.com/Rdatatable/data.table/wiki (Accessed: 17 November 2018).

Duffy, L. L., Dykes, G. A. and Fegan, N. (2012) ‘A review of the ecology, colonization and genetic characterization of *Salmonella enterica* serovar Sofia, a prolific but avirulent poultry serovar in Australia’, *Food Research International*, 45(2), pp. 770–779. doi: https://doi.org/10.1016/j.foodres.2011.04.024.

Eaves, D. J. *et al.* (2004) ‘Prevalence of Mutations within the Quinolone Resistance-Determining Region of gyrA , gyrB , parC , and parE and Association with Antibiotic Resistance in Quinolone-Resistant *Salmonella enterica*’, 48(10), pp. 4012–4015. doi: 10.1128/AAC.48.10.4012.

Economic Research Service (ERS) and U.S. Department of Agriculture (USDA) (2014) *Cost of foodborne illness estimates for Salmonella (non-typhoidal*, *Cost Estimates of Foodborne Illnesses*. Available at: https://www.ers.usda.gov/data-products/cost-estimates-of-foodborne-illnesses/ (Accessed: 2 February 2020).

Eddy, S. R. (2011) ‘Accelerated Profile HMM Searches’, *PLOS Computational Biology*. Public Library of Science, 7(10), pp. 1–16. doi: 10.1371/journal.pcbi.1002195.

Edwards, K. *et al.* (2001) ‘Genetic variability among archival cultures of *Salmonella* Typhimurium’, *FEMS Microbiology Letters*, 199(2), pp. 215–219. doi: 10.1111/j.1574-6968.2001.tb10677.x.

Eeckhaut, V. *et al.* (2018) ‘Oral vaccination with a live *Salmonella* Enteritidis / Typhimurium bivalent vaccine in layers induces cross-protection against caecal and internal organ colonization by a *Salmonella* Infantis strain’, *Veterinary Microbiology*. Elsevier, 218(September 2017), pp. 7–12. doi: 10.1016/j.vetmic.2018.03.022.

EFSA and ECDC (2011) ‘European Union summary report on antimicrobial resistance in zoonotic and indicator bacteria from animals and food in the European Union in 2009', *EFSA Journal*. John Wiley & Sons, Ltd, 9(7). doi: 10.2903/j.efsa.2011.2154.

EFSA and ECDC (2012) 'The European Union Summary Report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2010', *EFSA Journal*. John Wiley & Sons, Ltd, 10(3). doi: 10.2903/j.efsa.2012.2598.

EFSA and ECDC (2013a) 'The European Union Summary Report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2011', *EFSA Journal*. John Wiley & Sons, Ltd, 11(5). doi: 10.2903/j.efsa.2013.3196.

EFSA and ECDC (2013b) 'The European Union Summary Report on Trends and Sources of Zoonoses, Zoonotic Agents and Food-borne Outbreaks in 2011', *EFSA Journal*, 11(4).

EFSA and ECDC (2014a) 'The European Union Summary Report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2012', *EFSA Journal*. John Wiley & Sons, Ltd, 12(3). doi: 10.2903/j.efsa.2014.3590.

EFSA and ECDC (2014b) 'The European Union Summary Report on Trends and Sources of Zoonoses, Zoonotic Agents and Food-borne Outbreaks in 2012', *EFSA Journal*. John Wiley & Sons, Ltd, 12(2). doi: 10.2903/j.efsa.2014.3547.

EFSA and ECDC (2015a) 'EU Summary Report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2013', *EFSA Journal*. John Wiley & Sons, Ltd, 13(2). doi: 10.2903/j.efsa.2015.4036.

EFSA and ECDC (2015b) 'The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2013', *EFSA Journal*, 13(1). doi: 10.2903/j.efsa.2015.3991.

EFSA and ECDC (2015c) 'The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2014', *EFSA Journal*, 13(December 2015). doi: 10.2903/j.efsa.2015.4329.

EFSA and ECDC (2016a) 'The European Union summary report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2014', *EFSA Journal*. John Wiley & Sons, Ltd, 14(2). doi: 10.2903/j.efsa.2016.4380.

EFSA and ECDC (2016b) 'The European Union summary report on trends and sources of zoonoses , zoonotic agents and food-borne outbreaks in 2015', 14(November). doi: 10.2903/j.efsa.2016.4634.

EFSA and ECDC (2017a) 'The European Union summary report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2015', *EFSA Journal*. John Wiley & Sons, Ltd, 15(2). doi: 10.2903/j.efsa.2017.4694.

EFSA and ECDC (2017b) 'The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2016', *EFSA Journal*, 15(12). doi: 10.2903/j.efsa.2017.5077.

EFSA and ECDC (2018a) 'The European Union summary report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2016', *EFSA Journal*. John Wiley & Sons, Ltd, 16(2), p. e05182. doi: 10.2903/j.efsa.2018.5182.

EFSA and ECDC (2018b) 'The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2017', *EFSA Journal*, 16(December). doi: 10.2903/j.efsa.2018.5500.

EFSA and ECDC (2019a) 'The European Union One Health 2018 Zoonoses Report', *EFSA Journal*. John Wiley & Sons, Ltd, 17(12). doi: 10.2903/j.efsa.2019.5926.

EFSA and ECDC (2019b) 'The European Union summary report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2017', *EFSA Journal*. John Wiley & Sons, Ltd, 17(2). doi: 10.2903/j.efsa.2019.5598.

Eikmeier, D., Medus, C. and Smith, K. (2018) 'Incubation period for outbreak-associated, non-typhoidal salmonellosis cases, Minnesota, 2000–2015', *Epidemiology and Infection*. 2018/02/07. Cambridge University Press, 146(4), pp. 423–429. doi: DOI: 10.1017/S0950268818000079.

Ejemot-Nwadiaro RI, E. J. E. A. D. M. M. M. and Critchley, J. A. (2015) 'Hand washing promotion for preventing diarrhoea', *Cochrane Database of Systematic Reviews*. John Wiley & Sons, Ltd, (9). doi: 10.1002/14651858.CD004265.pub3.

Ekman, P. *et al.* (1999) 'Detection of *Salmonella* infantis in synovial fluid cells of a patient with reactive arthritis.', *The Journal of Rheumatology*. Canada, 26(11), pp. 2485–2488.

Elder, J. R. *et al.* (2018) 'Genomic organization and role of SPI-13 in nutritional fitness of *Salmonella*', *International Journal of Medical Microbiology*, 308(8), pp. 1043–1052. doi: https://doi.org/10.1016/j.ijmm.2018.10.004.

Eng, S. *et al.* (2015) '*Salmonella* : A review on pathogenesis , epidemiology and antibiotic resistance', *Frontiers in Life Science*, 8(3). doi: 10.1080/21553769.2015.1051243.

Espinoza, R. A. *et al.* (2017) 'Differential roles for pathogenicity islands SPI-13 and SPI-8 in the interaction of *Salmonella* Enteritidis and *Salmonella* Typhi with murine and human macrophages', *Biological research*. BioMed Central, 50(1). doi: 10.1186/s40659-017-0109-8.

Essack, S. Y. *et al.* (2017) 'Antimicrobial resistance in the WHO African region: current status and roadmap for action', *Journal of Public Health*, 39(1), pp. 8–13. Available at:

http://dx.doi.org/10.1093/pubmed/fdw015.

Eswarappa, S. M. *et al.* (2008) ‘Differentially Evolved Genes of *Salmonella* Pathogenicity Islands: Insights into the Mechanism of Host Specificity in *Salmonella*’, *PLOS ONE*. Public Library of Science, 3(12). Available at: https://doi.org/10.1371/journal.pone.0003829.

European Commission (2006) *Reducing Salmonella: Commission sets EU targets for laying hens and adopts new control rules*. Available at: http://europa.eu/rapid/press-release_IP-06-1082_en.htm (Accessed: 25 November 2019).

European Commission (2011) *COMMISSION REGULATION (EU) No 200/2010 of 10 March 2010 implementing Regulation (EC) No 2160/2003 of the European Parliament and of the Council as regards a Union target for the reduction of the prevalence of Salmonella serotypes in adult breeding flocks* . Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02010R0200-20110529&from=EN (Accessed: 25 November 2019).

European Commission (2012) *COMMISSION REGULATION (EU) No 200/2012 of 8 March 2012 concerning a Union target for the reduction of Salmonella enteritidis and Salmonella typhimurium in flocks of broilers, as provided for in Regulation (EC) No 2160/2003 of the European Parliament and of the Council Available at: https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:071:0031:0036:EN:PDF* (Accessed: 25 November 2019).

Farthing, M. J. G. and Kelly, P. (2007) ‘Infectious diarrhoea’, *Medicine*, 35(5), pp. 251–256. doi: https://doi.org/10.1016/j.mpmed.2007.02.009.

Faucher, S. P. *et al.* (2008) ‘The *prpZ* gene cluster encoding eukaryotic-type Ser/Thr protein kinases and phosphatases is repressed by oxidative stress and involved in *Salmonella enterica* serovar Typhi survival in human macrophages’, *FEMS Microbiology Letters*, 281(2), pp. 160–166. doi: 10.1111/j.1574-6968.2008.01094.x.

Feasey, N. A. *et al.* (2012) ‘Invasive non-typhoidal *Salmonella* disease: An emerging and neglected tropical disease in Africa’, *The Lancet*. Elsevier Ltd, 379(9835), pp. 2489–2499. doi: 10.1016/S0140-6736(11)61752-2.

Feasey, N. A. *et al.* (2016) ‘Distinct *Salmonella* Enteritidis lineages associated with enterocolitis in high-income settings and invasive disease in low-income settings’, *Nature Genetics*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 48, p. 1211. Available at: http://dx.doi.org/10.1038/ng.3644.

Feng, Y. *et al.* (2012) ‘Inheritance of the *Salmonella* virulence plasmids: Mostly vertical and rarely horizontal’, *Infection, Genetics and Evolution*, 12(5), pp. 1058–1063. doi:

https://doi.org/10.1016/j.meegid.2012.03.004.

Fierer, J. and Guiney, D. G. (2001) ‘Diverse virulence traits underlying different clinical outcomes of *Salmonella* infection’, *The Journal of Clinical Investigation*. The American Society for Clinical Investigation, 107(7), pp. 775–780. doi: 10.1172/JCI12561.

Fischbach, M. A. *et al.* (2006) ‘The pathogen-associated *iroA* gene cluster mediates bacterial evasion of lipocalin 2.’, *Proceedings of the National Academy of Sciences of the United States of America*, 103(44), pp. 16502–16507. doi: 10.1073/pnas.0604636103.

Fletcher, S. M., Stark, D. and Ellis, J. (2011) ‘Prevalence of gastrointestinal pathogens in Sub-Saharan Africa: systematic review and meta-analysis’, *Journal of public health in Africa*. PAGEPress Publications, 2(2). doi: 10.4081/jphia.2011.e30.

Foley, S. L. *et al.* (2011) ‘Population Dynamics of *Salmonella enterica* Serotypes in Commercial Egg and Poultry Production’, *Applied and Environmental Microbiology*, 77(13), pp. 4273–4279. doi: 10.1128/AEM.00598-11.

Food and Agriculture Organization of the United Nations (2019) *FAOSTAT Live Animals*. Available at: http://www.fao.org/faostat/en/#data/QA (Accessed: 19 November 2019).

Food and Drug Administration (FDA) (2019a) *2016-2017 NARMS Integrated Summary*. Available at: https://www.fda.gov/animal-veterinary/national-antimicrobial-resistance-monitoring-system/2016-2017-narms-integrated-summary-interactive (Accessed: 27 November 2019).

Food and Drug Administration (FDA) (2019b) *NARMS Now*, *Rockville, MD: U.S. Department of Health and Human Services*. Available at: https://www.fda.gov/animal-veterinary/national-antimicrobial-resistance-monitoring-system/narms-now-integrated-data (Accessed: 26 November 2019).

Food Standards Agency (2016) *The Second Study of Infectious Intestinal Disease in the Community (IID2 Study)*. Available at: https://www.food.gov.uk/research/research-projects/the-second-study-of-infectious-intestinal-disease-in-the-community-iid2-study.

Fookes, M. *et al.* (2011) ‘*Salmonella bongori* Provides Insights into the Evolution of the Salmonellae’, 7(8). doi: 10.1371/journal.ppat.1002191.

Francis, C. L., Starnbach, M. N. and Falkow, S. (1992) ‘Morphological and cytoskeletal changes in epithelial cells occur immediately upon interaction with *Salmonella* Typhimurium grown under low-oxygen conditions.’, *Molecular microbiology*. England, 6(21), pp. 3077–3087. doi: 10.1111/j.1365-2958.1992.tb01765.x.

Franco, A. *et al.* (2015) ‘Emergence of a Clonal Lineage of Multidrug- Resistant ESBL-Producing *Salmonella* Infantis Transmitted from Broilers and Broiler Meat to Humans in

Italy between 2011 and 2014’, pp. 1–15. doi: 10.1371/journal.pone.0144802.

Fu, S. *et al.* (2015) ‘Defining the Core Genome of *Salmonella enterica* Serovar Typhimurium for Genomic Surveillance and Epidemiological Typing’, *Journal of Clinical Microbiology*. Edited by D. J. Diekema, 53(8), pp. 2530– 2538. doi: 10.1128/JCM.03407-14.

Gal-Mor, O. *et al.* (2010) ‘Multidrug-Resistant *Salmonella enterica* Serovar Infantis, Israel’, *Emerg Infect Dis.*, 16(11), pp. 1754–1757. doi: https://dx.doi.org/10.3201/eid1611.100100.

Gal-Mor, O., Boyle, E. C. and Grassl, G. A. (2014) ‘Same species , different diseases : how and why typhoidal and non-typhoidal *Salmonella enterica* serovars differ’, *Frontiers in microbiology*, 5(August), pp. 1–10. doi: 10.3389/fmicb.2014.00391.

Galàn, J. E. (2001) ‘*Salmonella* Interactions With Host Cells : Type III Secretion at Work’, *Annu. Rev. Cell Dev. Biol.*, 17, pp. 53–86.

Garcia-Gutierrez, E. *et al.* (2016) ‘A comparison of the ATP generating pathways used by *S.* Typhimurium to fuel replication within human and murine macrophage and epithelial cell lines’, *PLoS ONE*, 11(3), pp. 1–15. doi: 10.1371/journal.pone.0150687.

Gast, R. K. and Holt, P. S. (1998) ‘Persistence of *Salmonella* Enteritidis from one day of age until maturity in experimentally infected layer chickens.’, *Poultry science*. England, 77(12), pp. 1759–1762. doi: 10.1093/ps/77.12.1759.

Gaze, W. H. *et al.* (2003) ‘Interactions between *Salmonella* Typhimurium and Acanthamoeba polyphaga, and observation of a new mode of intracellular growth within contractile vacuoles.’, *Microbial ecology*. United States, 46(3), pp. 358–369. doi: 10.1007/s00248-003-1001-3.

Giammanco, G. M. *et al.* (2002) ‘Persistent Endemicity of *Salmonella bongori* 48:z_{35}:– in Southern Italy: Molecular Characterization of Human, Animal, and Environmental Isolates’, *Journal of Clinical Microbiology*, 40(9), pp. 3502–3505. doi: 10.1128/JCM.40.9.3502-3505.2002.

Global Health Metrics (2017) ‘Global , regional , and national age-sex specific mortality for 264 causes of death , 1980 – 2016 : a systematic analysis for the Global Burden of Disease Study 2016’. doi: 10.1016/S0140-6736(17)32152-9.

Gomez-Eichelmann, M. C., Levy-Mustri, A. and Ramirez-Santos, J. (1991) ‘Presence of 5-methylcytosine in CC(A/T)GG sequences (Dcm methylation) in DNAs from different bacteria’, *Journal of bacteriology*, 173(23), pp. 7692–7694. doi: 10.1128/jb.173.23.7692-7694.1991.

Gong, H. *et al.* (2011) 'A *Salmonella* Small Non-Coding RNA Facilitates Bacterial Invasion and Intracellular Replication by Modulating the Expression of Virulence Factors', *PLOS Pathogens*. Public Library of Science, 7(9). Available at: https://doi.org/10.1371/journal.ppat.1002120.

Gordon, M. A. *et al.* (2008) 'Epidemics of Invasive *Salmonella enterica* Serovar Enteritidis and *S . enterica* Serovar Typhimurium Infection Associated with Multidrug Resistance among Adults and Children in Malawi', pp. 963–969. doi: 10.1086/529146.

Gordon, M. A. (2011) 'Invasive nontyphoidal *Salmonella* disease: Epidemiology, pathogenesis and diagnosis', *Current Opinion in Infectious Diseases*, 24(5), pp. 484–489. doi: 10.1097/QCO.0b013e32834a9980.

Granda, A. *et al.* (2019) 'Presence of Extended-Spectrum β -lactamase , CTX-M-65 in *Salmonella enterica* serovar Infantis Isolated from Children with Diarrhea in Lima , Peru', *JPediatr Infect Dis*.

Grimont, P. and Weill, F. (2007) 'Antigenic Formulae of the Salmonella serovars, (9th ed.) Paris: WHO Collaborating Centre for Reference and Research on Salmonella', *Institute Pasteur.*, pp. 1–166.

Group for Enteric Respiratory and Meningeal disease Surveillance in South Africa (2009) *GERMS-SA Annual Report 2009*. Available at: http://www.nicd.ac.za/assets/files/2009GERMS-SA_Annual_Report.pdf.

Group for Enteric Respiratory and Meningeal disease Surveillance in South Africa (2010) *GERMS-SA Annual Report 2010*. Available at: http://www.nicd.ac.za/assets/files/2010_GERMS-SA_Annual_report_Final.pdf.

Group for Enteric Respiratory and Meningeal disease Surveillance in South Africa (2013) *GERMS-SA Annual Report 2013*. Available at: http://www.nicd.ac.za/assets/files/GERMS-SA AR 2013(1).pdf.

Group for Enteric Respiratory and Meningeal disease Surveillance in South Africa (2016) *GERMS-SA Annual Report 2016*. Available at: http://www.nicd.ac.za/wp-content/uploads/2017/03/GERMS-SA-AR-2016-FINAL.pdf.

Guo, C. *et al.* (2011) 'Application of Bayesian techniques to model the burden of human salmonellosis attributable to U.S. food commodities at the point of processing: adaptation of a Danish model', *Foodborne pathogens and disease*. 2011/01/16. Mary Ann Liebert, Inc., 8(4), pp. 509–516. doi: 10.1089/fpd.2010.0714.

Gupta, S. M. *et al.* (2019) 'Genomic comparison of diverse *Salmonella* serovars isolated from swine', *PloS one*, 14(11).

Gurevich, A. *et al.* (2013) ‘QUAST: quality assessment tool for genome assemblies’, *Bioinformatics*, 29(8), pp. 1072–1075. doi: 10.1093/bioinformatics/btt086.

Gymoese, P. *et al.* (2019) ‘WGS based study of the population structure of *Salmonella enterica* serovar Infantis’, *BMC genomics*. BMC Genomics, 20(870), pp. 1–11.

Hadfield, J. *et al.* (2018) ‘Phandango: an interactive viewer for bacterial population genomics’, *Bioinformatics*, 34(2), pp. 292–293. Available at: http://dx.doi.org/10.1093/bioinformatics/btx610.

Hægland, H. (2013) *How to transpose a huge txt file with 1,743,680 columns and 2890 rows [duplicate]*. Available at: https://stackoverflow.com/questions/19843155/how-to-transpose-a-huge-txt-file-with-1-743-680-columns-and-2890-rows (Accessed: 17 November 2018).

Haider, G., Chowdhury, E. H. and Hossain, M. (2014) ‘Mode of vertical transmission of *Salmonella enterica* sub. enterica serovar Pullorum in chickens’, 8(12), pp. 1344–1351. doi: 10.5897/AJMR2013.6452.

Hald, T. *et al.* (2016) ‘World Health Organization Estimates of the Relative Contributions of Food to the Burden of Disease Due to Selected Foodborne Hazards: A Structured Expert Elicitation’, *PloS one*, 11. doi: 10.1371/journal.pone.0145839.

Hammarlöf, D. L. *et al.* (2018) ‘Role of a single noncoding nucleotide in the evolution of an epidemic African clade of *Salmonella*.’, *Proceedings of the National Academy of Sciences of the United States of America*, 115(11). doi: 10.1073/pnas.1714718115.

Han, J. *et al.* (2018) ‘International Journal of Food Microbiology Impact of co-carriage of IncA / C plasmids with additional plasmids on the transfer of antimicrobial resistance in *Salmonella enterica* isolates’, *International Journal of Food Microbiology*, 271, pp. 77–84. doi: 10.1016/j.ijfoodmicro.2018.01.018.

Haraga, A., Ohlson, M. B. and Miller, S. I. (2008) ‘Salmonellae interplay with host cells’, *Nature Reviews Microbiology*, 6(1), pp. 53–66. doi: 10.1038/nrmicro1788.

Hara-Kudo, Y. *et al.* (2013) ‘Prevalence of the main food-borne pathogens in retail food under the national food surveillance system in Japan.’, *Food additives & contaminants. Part A, Chemistry, analysis, control, exposure & risk assessment*, 30(8), pp. 1450–8. doi: 10.1080/19440049.2012.745097.

Harbottle, H. *et al.* (2006) ‘Genetics of Antimicrobial Resistance’, *Animal Biotechnology*. Taylor & Francis, 17(2), pp. 111–124. doi: 10.1080/10495390600957092.

Hauser, E. *et al.* (2012) ‘Clonal Dissemination of *Salmonella enterica* Serovar Infantis in Germany’, *Foodborne Pathogens and Disease*, 9(4), pp. 352–360. doi:

10.1089/fpd.2011.1038.

Havelaar, A. H. *et al.* (2015) 'World Health Organization Global Estimates and Regional Comparisons of the Burden of Foodborne Disease in 2010', *PLoS medicine*. Public Library of Science, 12(12). doi: 10.1371/journal.pmed.1001923.

Helaine, S. *et al.* (2014) 'Internalization of *Salmonella* by macrophages induces formation of nonreplicating persisters.', *Science (New York, N.Y.)*, 343(6167), pp. 204–208. doi: 10.1126/science.1244705.

Henderson, S. C., Bounous, D. I. and Lee, M. D. (1999) 'Early events in the pathogenesis of avian salmonellosis.', *Infection and immunity*. United States, 67(7), pp. 3580–3586.

Hendriksen, R. S. *et al.* (2011) 'Global Monitoring of *Salmonella* Serovar Distribution from the World Health Organization Global Foodborne Infections Network Country Data Bank: Results of Quality Assured Laboratories from 2001 to 2007', *Foodborne Pathogens and Disease*, 8(8), pp. 887–900. doi: 10.1089/fpd.2010.0787.

Hindermann, D. *et al.* (2017) '*Salmonella enterica* serovar Infantis from food and human infections, Switzerland, 2010-2015: Poultry-related multidrug resistant clones and an emerging ESBL producing clonal lineage', *Frontiers in Microbiology*, 8(JUL), pp. 1–9. doi: 10.3389/fmicb.2017.01322.

Holt, P. S. *et al.* (1999) 'Hyporesponsiveness of the systemic and mucosal humoral immune systems in chickens infected with *Salmonella enterica* serovar Enteritidis at one day of age.', *Poultry science*. England, 78(11), pp. 1510–1517. doi: 10.1093/ps/78.11.1510.

Huang, Y.-K. *et al.* (2019) 'Pathogenicity differences of *Salmonella enterica* serovars Typhimurium, Enteritidis, and Choleraesuis-specific virulence plasmids and clinical *S*. Choleraesuis strains with large plasmids to the human THP-1 cell death', *Microbial Pathogenesis*, 128, pp. 69–74. doi: https://doi.org/10.1016/j.micpath.2018.12.035.

Hunt, M. *et al.* (2015) 'Circlator: automated circularization of genome assemblies using long sequencing reads', *Genome Biology*, 16(1), p. 294. doi: 10.1186/s13059-015-0849-0.

Hunt, M. *et al.* (2017) 'ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads', *Microbial Genomics*, 3(10), p. Available at: http://mgen.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000131.

Hyatt, D. *et al.* (2010) 'Prodigal: prokaryotic gene recognition and translation initiation site identification.', *BMC bioinformatics*. England, 11, p. 119. doi: 10.1186/1471-2105-11-119.

Illumina (2015) *Nextera XT Library Prep : Tips and Troubleshooting Sample Input*.

Available at: https://www.illumina.com/content/dam/illumina-marketing/documents/products/technotes/nextera-xt-troubleshooting-technical-note.pdf (Accessed: 15 July 2019).

Illumina (2017) *MiSeq System Denature and Dilute Libraries Guide*. Available at: http://os.bio-protocol.org/attached/file/20171217/miseq denature dilute libraries guide 15039740 03.pdf (Accessed: 15 July 2019).

Illumina (2018a) *Nextera XT DNA Library Prep Kit Reference Guide*. Available at: https://genome.med.harvard.edu/documents/libraryPrep/IlluminaNexteraXTProtocol.pdf (Accessed: 20 April 2018).

Illumina (2018b) 'NextSeq System Denature and Dilute Libraries Guide', *Document # 15048776 v09*. Available at: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwjT29jBheXmAhUBQUEAHSd6AswQFjAAegQIAhAC&url=https%253A%252F%252Fsupport.illumina.com%252Fcontent%252Fdam%252Fillumina-support%252Fdocuments%252Fdocumentation%252Fsystem_documentation%252 (Accessed: 15 July 2019).

Illumina (2019a) *Access a wide range of BaseSpace Apps for simplified data analysis*. Available at: https://emea.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps.html (Accessed: 12 July 2019).

Illumina (2019b) *bcl2fastq Conversion Software*. Available at: http://emea.support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html (Accessed: 15 July 2019).

Ilyas, B., Tsai, C. N. and Coombes, B. K. (2017) 'Evolution of *Salmonella*-Host Cell Interactions through a Dynamic Bacterial Genome', *Frontiers in Cellular and Infection Microbiology*, 7, p. 428. doi: 10.3389/fcimb.2017.00428.

Imre, A. *et al.* (2013) 'Gene expression analysis of *Salmonella enterica* SPI in macrophages indicates differences between serovars that induce systemic disease from those normally causing enteritis', *Veterinary microbiology*. 2013/08/09. Elsevier Scientific Pub. Co, 167(3–4), pp. 675–679. doi: 10.1016/j.vetmic.2013.07.034.

Institute of Environmental Science and Research (2020) *Human Salmonella Isolates, 2019, Information for New Zealand Public Health Action*. Available at: https://surv.esr.cri.nz/enteric_reference/human_salmonella.php?we_objectID=5083 (Accessed: 4 February 2020).

Iriarte, A. *et al.* (2017) 'Draft Genome Sequence of *Salmonella enterica* subsp. *enterica*

Serovar Infantis Strain SPE101, Isolated from a Chronic Human Infection', *Genome Announcements*, 5(29). doi: 10.1128/genomeA.00679-17.

Issenhuth-Jeanjean, S. *et al.* (2014) 'Supplement 2008-2010 (no. 48) to the White-Kauffmann-Le Minor scheme', *Research in Microbiology*, 165(48), pp. 526–530. doi: 10.1016/j.resmic.2014.07.004.

Jacobsen, A. and Hendriksen, R. S. (2011) 'The *Salmonella enterica* Pan-genome', *Microbial Ecology*, 62(487), pp. 487–504. doi: 10.1007/s00248-011-9880-1.

Jaillard, M. *et al.* (2018) 'A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events', *PLOS Genetics*. Public Library of Science, 14(11), pp. 1–28. doi: 10.1371/journal.pgen.1007758.

Jajere, S. M. (2019) 'A review of *Salmonella enterica* with particular focus on the pathogenicity and virulence factors , host specificity and antimicrobial resistance including multidrug resistance', *Veterinary World*, 12, pp. 504–521. doi: 10.14202/vetworld.2019.504-521.

Jennings, E., Thurston, T. L. M. and Holden, D. W. (2017) '*Salmonella* SPI-2 Type III Secretion System Effectors: Molecular Mechanisms And Physiological Consequences', *Cell Host & Microbe*. Elsevier, 22(2), pp. 217–231. doi: 10.1016/j.chom.2017.07.009.

Jironkin, A. *et al.* (2017) *PHEnix*. Available at: https://github.com/phe-bioinformatics/PHEnix (Accessed: 9 February 2017).

Jones, T. F. *et al.* (2008) 'Salmonellosis Outcomes Differ Substantially by Serotype', *The Journal of Infectious Diseases*, 198, pp. 109–14. doi: 10.1086/588823.

Kaiser, P. *et al.* (2000) 'Differential cytokine expression in avian cells in response to invasion by *Salmonella* Typhimurium, *Salmonella* Enteritidis and *Salmonella* Gallinarum.', *Microbiology (Reading, England)*. England, 146 Pt 12, pp. 3217–3226. doi: 10.1099/00221287-146-12-3217.

Kans, J. (2019) *Entrez Direct: E-utilities on the UNIX Command Line*. National Center for Biotechnology Information (US). Available at: https://www.ncbi.nlm.nih.gov/books/NBK179288/ (Accessed: 12 July 2019).

KAPA Biosystems (2017) *KAPA Pure Beads*. Available at: http://netdocs.roche.com/DDM/Effective/0000000000001200000190108_000_03_005_Native.pdf (Accessed: 9 July 2019).

Kazemi, M., Gumpert, G. and Marks, M. I. (1974) 'Clinical spectrum and carrier state of nontyphoidal salmonella infections in infants and children', *Canadian Medical Association journal*, 110(11), pp. 1253–1257. Available at:

https://www.ncbi.nlm.nih.gov/pubmed/4857958.

Khaitsa, M. L. and Doetkott, D. (2009) ‘Antimicrobial Drug Resistance and Molecular Characterization of *Salmonella* Isolated from Domestic Animals , Humans and Meat Products’, *Foodborne Pathog Dis*, (May). doi: 10.1089/fpd.2008.0134.

Kingsbury, J. M. *et al.* (2019) ‘Prevalence and Genetic Analysis of *Salmonella enterica* from a Cross-Sectional Survey of the New Zealand Egg Production Environment’, *Journal of Food Protection*, 82(12), pp. 2201–2214.

Kingsley, R. A. and Bäumler, A. J. (2000) ‘Host adaptation and the emergence of infectious disease : the *Salmonella* paradigm’, *Molecular Microbiology*, 36(5), pp. 1006–1014.

Kirk, M. D. *et al.* (2015) ‘World Health Organization Estimates of the Global and Regional Disease Burden of 22 Foodborne Bacterial, Protozoal, and Viral Diseases, 2010 : A Data Synthesis’, pp. 1–21. doi: 10.1371/journal.pmed.1001921.

De Knegt, L. V, Pires, S. M. and Hald, T. (2015) ‘Attributing foodborne salmonellosis in humans to animal reservoirs in the European Union using a multi-country stochastic model’, *Epidemiol. Infect.*, 143, pp. 1175–1186. doi: 10.1017/S0950268814001903.

Kogut, H. M. *et al.* (2016) ‘Chicken-Specific Kinome Array Reveals that *Salmonella enterica* Serovar Enteritidis Modulates Host Immune Signaling Pathways in the Cecum to Establish a Persistence Infection’, *International Journal of Molecular Sciences* . doi: 10.3390/ijms17081207.

Kogut, M. H. and Arsenault, R. J. (2017) ‘Immunometabolic Phenotype Alterations Associated with the Induction of Disease Tolerance and Persistent Asymptomatic Infection of *Salmonella* in the Chicken Intestine’, *Frontiers in Immunology*, 8, p. 372. doi: 10.3389/fimmu.2017.00372.

Kohler, P. F. (1964) ‘Hospital Salmonellosis: A Report of 23 Cases of Gastroenteritis Caused by *Salmonella* Infantis’, *JAMA*, 189(1), pp. 6–10. doi: 10.1001/jama.1964.03070010012002.

Kongsoi, S., Nakajima, C. and Suzuki, Y. (2017) ‘Quinolone Resistance in Non-typhoidal *Salmonella*’, in Mares, M. (ed.) *Current Topics in Salmonella and Salmonellosis*. IntechOpen, pp. 115–135. doi: 10.5772/32009.

Koren, S. *et al.* (2017) ‘Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation.’, *Genome research*. United States, 27(5), pp. 722–736. doi: 10.1101/gr.215087.116.

Koziolek, M. *et al.* (2015) ‘Investigation of pH and Temperature Profiles in the GI Tract of Fasted Human Subjects Using the Intellicap® System’, *Journal of Pharmaceutical Sciences*, 104(9), pp. 2855–2863. doi: https://doi.org/10.1002/jps.24274.

Kruchten, T. (2002) 'U . S . Broiler Industry Structure', *National Agricultural Statistics Service*.

Krumm, A., Gröne, M. and Maurer, F. (2017) 'Quantifying double-stranded DNA with fluorescent dyes : Qubit^{TM} on BMG LABTECH instruments', *BMG LABTECH*.

Kumar, S., Stecher, G. and Tamura, K. (2016) 'MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets.', *Molecular biology and evolution*. United States, 33(7), pp. 1870–1874. doi: 10.1093/molbev/msw054.

Kurtz, S. *et al.* (2004) 'Versatile and open software for comparing large genomes.', *Genome biology*. England, 5(2). doi: 10.1186/gb-2004-5-2-r12.

Lai, J. *et al.* (2014) 'Serotype distribution and antibiotic resistance of *Salmonella* in food-producing animals in Shandong province of China , 2009 and 2012', *International Journal of Food Microbiology*. Elsevier B.V., 180, pp. 30–38. doi: 10.1016/j.ijfoodmicro.2014.03.030.

Laing, C. R., Whiteside, M. D. and Gannon, V. P. J. (2017) 'Pan-genome Analyses of the Species *Salmonella enterica*, and Identification of Genomic Markers Predictive for Species, Subspecies, and Serovar', *Frontiers in microbiology*, 8(1345), pp. 1–16. doi: 10.3389/fmicb.2017.01345.

Lamas, A. *et al.* (2018) 'A comprehensive review of non-*enterica* subspecies of *Salmonella enterica*', *Microbiological Research*. Elsevier, 206(October 2017), pp. 60–73. doi: 10.1016/j.micres.2017.09.010.

Lamont, S. J. (2010) '*Salmonella* in Chickens', in Bishop, S. et al. (eds) *Breeding For Disease Resistance In Farm Animals*. 3rd edn.

Langridge, G. C. *et al.* (2009) 'Simultaneous assay of every *Salmonella* Typhi gene using one million transposon mutants', *Genome Research*, 19(12), pp. 2308–2316. doi: 10.1101/gr.097097.109.

Langridge, G. C. *et al.* (2015) 'Patterns of genome evolution that have accompanied host adaptation in *Salmonella*.', *Proceedings of the National Academy of Sciences of the United States of America*, 112(3), pp. 863–8. doi: 10.1073/pnas.1416707112.

Langridge, G. C., Nair, S. and Wain, J. (2009) 'Nontyphoidal *Salmonella* Serovars Cause Different Degrees of Invasive Disease Globally', *The Journal of Infectious Diseases*, 199(4), pp. 602–603. doi: 10.1086/596208.

Langridge, G. C., Wain, J. and Nair, S. (2012) 'Invasive Salmonellosis in Humans', *EcoSal Plus*, 5(1). doi: 10.1128/ecosalplus.8.6.2.2.

Leekitcharoenphon, P. *et al.* (2012) ‘Genomic variation in *Salmonella enterica* core genes for epidemiological typing’, *BMC Genomics*, 13(1), p. 88. doi: 10.1186/1471-2164-13-88.

Lees, J. A. *et al.* (2018) ‘pyseer: a comprehensive tool for microbial pangenome-wide association studies’, *Bioinformatics*, 34(24), pp. 4310–4312. doi: 10.1093/bioinformatics/bty539.

Lees, J. A. (2019) *unitig-counter*. Available at: https://github.com/johnlees/unitig-counter/blob/master/README.md (Accessed: 10 June 2019).

Lees, J. and Galardini, M. (2018) *PySeer Documentation - Usage*. Available at: http://gensoft.pasteur.fr/docs/pyseer/1.0.2/usage.html (Accessed: 1 November 2019).

Leinonen, R., Akhtar, R., *et al.* (2011) ‘The European Nucleotide Archive’, *Nucleic acids research*. 2010/10/23. Oxford University Press, 39(Database issue). doi: 10.1093/nar/gkq967.

Leinonen, R., Sugawara, H., *et al.* (2011) ‘The sequence read archive’, *Nucleic acids research*. 2010/11/09. Oxford University Press, 39(Database issue). doi: 10.1093/nar/gkq1019.

Leonard, E. K. *et al.* (2011) ‘Evaluation of pet-related management factors and the risk of *Salmonella* spp. carriage in pet dogs from volunteer households in Ontario (2005-2006).’, *Zoonoses and public health*. Germany, 58(2), pp. 140–149. doi: 10.1111/j.1863-2378.2009.01320.x.

Letunic, I. and Bork, P. (2016) ‘Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees’, *Nucleic acids research*. doi: 10.1093/nar/gkw290.

Lewith, G. T., Jonas, W. B. and Walach, H. (2010) *Clinical Research in Complementary Therapies E-Book: Principles, Problems and Solutions*. Elsevier Health Sciences. Available at: https://books.google.co.uk/books?id=CSNw-spnFdkC.

Lex, A. *et al.* (2014) ‘UpSet : Visualization of Intersecting Sets’, *IEEE Transactions on Visualization and Computer Graphics*. IEEE, 20(12), pp. 1983–1992. doi: 10.1109/TVCG.2014.2346248.

Li, H. *et al.* (2009) ‘The Sequence Alignment/Map format and SAMtools.’, *Bioinformatics (Oxford, England)*, 25(16), pp. 2078–9. doi: 10.1093/bioinformatics/btp352.

Li, H. (2013) ‘Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM’, *arXiv*. Available at: http://arxiv.org/abs/1303.3997.

Li, Y. *et al.* (2014) ‘Nontyphoidal *Salmonella* Infection in Children with Acute Gastroenteritis : Prevalence, Serotypes, and Antimicrobial Resistance in Shanghai, China’, *Foodborne Pathogens and Disease*, 11(3). doi: 10.1089/fpd.2013.1629.

Liao, J. *et al.* (2019) 'Serotype-specific evolutionary patterns of antimicrobial-resistant *Salmonella enterica*', *BMC Evolutionary Biology*, 19(1). doi: 10.1186/s12862-019-1457-5.

Liaquat, S. *et al.* (2018) 'Virulotyping of *Salmonella enterica* serovar Typhi isolates from Pakistan: Absence of complete SPI-10 in Vi negative isolates', *PLOS Neglected Tropical Diseases*. Public Library of Science, 12(11). Available at: https://doi.org/10.1371/journal.pntd.0006839.

Libby, S. J. *et al.* (1997) 'The spv genes on the *Salmonella* Dublin virulence plasmid are required for severe enteritis and systemic infection in the natural host.', *Infection and Immunity*, 65(5), pp. 1786–1792. Available at: http://iai.asm.org/content/65/5/1786.abstract.

Litvak, Y. *et al.* (2019) 'Commensal *Enterobacteriaceae* Protect against *Salmonella* Colonization through Oxygen Competition', *Cell Host & Microbe*, 25, pp. 128–139. doi: 10.1016/j.chom.2018.12.003.

Liu, L. *et al.* (2012) 'Global , regional , and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000', *The Lancet*. Elsevier Ltd, 379(9832), pp. 2151–2161. doi: 10.1016/S0140-6736(12)60560-1.

Loman, N. J., Quick, J. and Simpson, J. T. (2015) 'A complete bacterial genome assembled de novo using only nanopore sequencing data', *Nature Methods*. Nature Publishing Group, 12, p. 733. Available at: https://doi.org/10.1038/nmeth.3444.

Lou, L. *et al.* (2019) '*Salmonella* Pathogenicity Island 1 (SPI-1) and Its Complex Regulatory Network', *Frontiers in cellular and infection microbiology*. Frontiers Media S.A., 9, p. 270. doi: 10.3389/fcimb.2019.00270.

De Lucia, A. *et al.* (2018) 'Role of wild birds and environmental contamination in the epidemiology of *Salmonella* infection in an outdoor pig farm', *Veterinary Microbiology*, 227, pp. 148–154. doi: https://doi.org/10.1016/j.vetmic.2018.11.003.

Luk, J. M. *et al.* (1993) 'Selective amplification of abequose and paratose synthase genes (rfb) by polymerase chain reaction for identification of *Salmonella* major serogroups (A, B, C2, and D).', *Journal of Clinical Microbiology*, 31(8), pp. 2118–2123. Available at: http://jcm.asm.org/content/31/8/2118.abstract.

Lunn, A. D. *et al.* (2010) 'Prevalence of mechanisms decreasing quinolone-susceptibility among *Salmonella* spp . clinical isolates', pp. 15–20. doi: 10.2436/20.1501.01.107.

Lupolova, N. *et al.* (2016) 'Support vector machine applied to predict the zoonotic potential of *E. coli* O157 cattle isolates', *Proceedings of the National Academy of Sciences*.

National Academy of Sciences, 113(40), pp. 11312–11317. doi: 10.1073/pnas.1606567113.

Lupolova, N. *et al.* (2017) ‘Patchy promiscuity: machine learning applied to predict the host specificity of Salmonella enterica and Escherichia coli’, *Microbial genomics*. Microbiology Society, 3(10). doi: 10.1099/mgen.0.000135.

MacLennan, C. *et al.* (2004) ‘Interleukin (IL)-12 and IL-23 are key cytokines for immunity against *Salmonella* in humans.’, *The Journal of infectious diseases*. United States, 190(10), pp. 1755–1757. doi: 10.1086/425021.

Makendi, C. *et al.* (2016) ‘A Phylogenetic and Phenotypic Analysis of *Salmonella enterica* Serovar Weltevreden , an Emerging Agent of Diarrheal Disease in Tropical Regions’, *PLOS Neglected Tropical Diseases*, 10(2). doi: 10.1371/journal.pntd.0004446.

Mandal, B. K. and Brennand, J. (1988) ‘Bacteraemia in salmonellosis: a 15 year retrospective study from a regional infectious diseases unit’, *BMJ*, 297, pp. 1242–1243.

Mandal, R. K. and Kwon, Y. M. (2017) ‘Global Screening of *Salmonella enterica* Serovar Typhimurium Genes for Desiccation Survival’, *Frontiers in Microbiology*, 8, p. 1723. doi: 10.3389/fmicb.2017.01723.

Marcus, S. L. *et al.* (2000) ‘*Salmonella* pathogenicity islands: big virulence in small packages’, *Microbes and Infection*, 2(2), pp. 145–156. doi: https://doi.org/10.1016/S1286-4579(00)00273-2.

Marin, C. *et al.* (2011) ‘Sources of *Salmonella* contamination during broiler production in Eastern Spain’, *Preventive Veterinary Medicine*. Elsevier B.V., 98(1), pp. 39–45. doi: 10.1016/j.prevetmed.2010.09.006.

Mashima, J. *et al.* (2017) ‘DNA Data Bank of Japan.’, *Nucleic acids research*. England, 45, pp. 25–31. doi: 10.1093/nar/gkw1001.

MathWorks (2014) ‘MATLAB Compiler’. Available at: https://uk.mathworks.com/products/compiler/matlab-runtime.html.

Matthews, T. D., Rabsch, W. and Maloy, S. (2011) ‘Chromosomal Rearrangements in *Salmonella enterica* Serovar Typhi Strains Isolated from Asymptomatic Human Carriers’, *mBio*. Edited by J. E. Davies. American Society for Microbiology, 2(3). doi: 10.1128/mBio.00060-11.

McEntire, J. *et al.* (2014) ‘The Public Health Value of Reducing Salmonella Levels in Raw Meat and Poultry’, *Food Protection Trends*, 34(6).

McGinnis, S. and Madden, T. L. (2004) ‘BLAST: at the core of a powerful and diverse set of sequence analysis tools’, *Nucleic Acids Research*, 32(suppl_2), pp. 20–25. doi:

10.1093/nar/gkh435.

McQuiston, J. R. *et al.* (2008) ‘Do *Salmonella* carry spare tyres?’, *Trends in Microbiology*, 16(4), pp. 142–148. doi: https://doi.org/10.1016/j.tim.2008.01.009.

Mead, P. S. *et al.* (1999) ‘Food-related illness and death in the United States’, *Emerging infectious diseases*. Centers for Disease Control, 5(5), pp. 607–625. doi: 10.3201/eid0505.990502.

Medalla, F. *et al.* (2017) ‘Estimated Incidence of Antimicrobial Drug–Resistant Nontyphoidal *Salmonella* Infections, United States, 2004–2012.’, *Emerg Infect Dis.*, 2017;23(1)(1), pp. 2004–2012. doi: https://dx.doi.org/10.3201/eid2301.160771.

Medema, G. J. and Schijven, J. F. (2001) ‘Modelling the sewage discharge and dispersion of cryptosporidium and giardia in surface water’, *Water Research*, 35(18), pp. 4307–4316. doi: https://doi.org/10.1016/S0043-1354(01)00161-0.

Meunier, R. A. and Latour, M. (2005) ‘Commercial Egg Production and Processing’, *Purdue Agriculture*. Available at: http://ag.ansc.purdue.edu/poultry/publication/commegg/.

Michael, G. *et al.* (2006) ‘Genes and mutations conferring antimicrobial resistance in *Salmonella*: an update’, 8, pp. 1898–1914. doi: 10.1016/j.micinf.2005.12.019.

Michonneau, F. *et al.* (2019) *phylobase: Base package for phylogenetic structures and comparative data*. Available at: https://www.rdocumentation.org/packages/phylobase/versions/0.8.4/topics/Import Nexus and Newick files (Accessed: 17 June 2019).

Millan, J. *et al.* (2004) ‘*Salmonella* isolates from wild birds and mammals in the Basque Country (Spain).’, *Revue scientifique et technique (International Office of Epizootics)*. France, 23(3), pp. 905–911. doi: 10.20506/rst.23.3.1529.

Miller, T. *et al.* (2010) ‘Epidemiological relationship between *Salmonella* Infantis isolates of human and broiler origin’, *Lohmann Information*, 45(2), pp. 27–31.

Miller, T. *et al.* (2018) ‘Recurrent outbreaks caused by the same *Salmonella enterica* serovar Infantis clone in a German rehabilitation oncology clinic from 2002 to 2009’, *Journal of Hospital Infection*. Elsevier Ltd, 100(4). doi: 10.1016/j.jhin.2018.03.035.

Milne, I. *et al.* (2013) ‘Using Tablet for visual exploration of second-generation sequencing data’, *Briefings in Bioinformatics*, 14(2), pp. 193–202. Available at: http://dx.doi.org/10.1093/bib/bbs012.

Miranda, A. L. *et al.* (2017) ‘Phenotypic and genotypic characterization of *Salmonella* spp. isolated from foods and clinical samples in Brazil’, *Annals of the Brazilian Academy of Sciences*, 89(2), pp. 1143–1153.

Misselwitz, B. *et al.* (2011) '*Salmonella enterica* serovar Typhimurium binds to HeLa cells via Fim-mediated reversible adhesion and irreversible type three secretion system 1-mediated docking.', *Infection and immunity*. United States, 79(1), pp. 330–341. doi: 10.1128/IAI.00581-10.

Mitchell, M. A. and Kettlewell, P. J. (2009) 'Welfare of poultry during transport – a review', (May), pp. 18–22.

Modena, B. D. *et al.* (2019) 'Leveraging genomics to uncover the genetic, environmental and age-related factors leading to asthma', in Ginsburg, G. S. et al. (eds) *Genomic and Precision Medicine*. 3rd edn. Boston: Academic Press, pp. 331–381. doi: https://doi.org/10.1016/B978-0-12-801496-7.00018-6.

Mollenhorst, H. *et al.* (2005) 'Risk factors for *Salmonella* Enteritidis infections in laying hens.', *Poultry science*, 84(8), pp. 1308–13. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16156216.

Monte, D. F. *et al.* (2019) 'Current insights on high priority antibiotic-resistant *Salmonella enterica* in food and foodstuffs : a review', *Current Opinion in Food Science*. Elsevier Ltd, 26, pp. 35–46. doi: 10.1016/j.cofs.2019.03.004.

Morgan, E. *et al.* (2004) 'Identification of host-specific colonization factors of *Salmonella enterica* serovar Typhimurium', *Molecular Microbiology*. John Wiley & Sons, Ltd, 54(4), pp. 994–1010. doi: 10.1111/j.1365-2958.2004.04323.x.

Morningstar-Shaw, B. R. *et al.* (2016) *Salmonella Serotypes Isolated from Animals and Related Sources*. Available at: https://www.cdc.gov/nationalsurveillance/pdfs/salmonella-serotypes-isolated-animals-and-related-sources-508.pdf.

Mortelmans, J., Huygelen, C. and Pinckers, F. (1958) 'Isolation of *Salmonella* Infantis from an aborted bovine foetus.', *Nature*. England, 181(4622), pp. 1539–1540. doi: 10.1038/1811539b0.

Mughini-Gras, L., Heck, M. and van Pelt, W. (2016) 'Increase in reptile-associated human salmonellosis and shift toward adulthood in the age groups at risk, the Netherlands, 1985 to 2014', *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*. European Centre for Disease Prevention and Control (ECDC), 21(34). doi: 10.2807/1560-7917.ES.2016.21.34.30324.

Mulder, R. W. A. W. (1995) 'Impact of transport and related stresses on the incidence and extent of human pathogens in pigmeat and poultry', *Journal of Food Safety*. John Wiley & Sons, Ltd, 15(3), pp. 239–246. doi: 10.1111/j.1745-4565.1995.tb00136.x.

Muranaka, K. *et al.* (2015) 'A Case of Osteomyelitis with a Granulomatous Lesion Caused by *Salmonella* Infantis', *Kansenshogaku zasshi. The Journal of the Japanese Association for Infectious Diseases*. Japan, 89(4), pp. 476–480. doi: 10.11150/kansenshogakuzasshi.89.476.

Müştak, Başak, İ. and Yardımcı, H. (2019) 'Construction and in vitro characterisation of aroA defective (aroA Δ) mutant *Salmonella* Infantis', *Archives of microbiology*. doi: 10.1007/s00203-019-01694-0.

Musto, J. *et al.* (2006) 'Multi-drug resistant *Salmonella* Java infections acquired from tropical fish aquariums, Australia, 2003-04.', *Communicable diseases intelligence quarterly report*. Australia, 30(2), pp. 222–227.

Nakadai, A. *et al.* (2005) 'Prevalence of *Salmonella* spp. in Pet Reptiles in Japan', *Journal of Veterinary Medical Science*, 67(1), pp. 97–101. doi: 10.1292/jvms.67.97.

Namata, H. *et al.* (2009) 'Identification of risk factors for the prevalence and persistence of *Salmonella* in Belgian broiler chicken flocks', *Preventive Veterinary Medicine*, 90(3), pp. 211–222. doi: https://doi.org/10.1016/j.prevetmed.2009.03.006.

Narzisi, G. and Mishra, B. (2011) 'Comparing De Novo Genome Assembly: The Long and Short of It', *PLOS ONE*. Public Library of Science, 6(4), pp. 1–17. doi: 10.1371/journal.pone.0019175.

National Center for Biotechnology Information (2014) *Using the SRA Toolkit to convert .sra files into other formats*. Available at: https://www.ncbi.nlm.nih.gov/books/NBK158900/ (Accessed: 26 September 2017).

National Center for Emerging and Zoonotic Infectious Diseases (2018) *National Enteric Disease Surveillance: Salmonella Annual Report, 2016*. Available at: https://www.cdc.gov/nationalsurveillance/pdfs/2016-Salmonella-report-508.pdf.

Nawrocki, E. P. and Eddy, S. R. (2013) 'Infernal 1.1: 100-fold faster RNA homology searches.', *Bioinformatics (Oxford, England)*. England, 29(22), pp. 2933–2935. doi: 10.1093/bioinformatics/btt509.

Nieto, P. A. *et al.* (2016) 'New insights about excisable pathogenicity islands in *Salmonella* and their contribution to virulence', *Microbes and Infection*, 18(5), pp. 302–309. doi: https://doi.org/10.1016/j.micinf.2016.02.001.

Northcutt, J. K. *et al.* (2003) 'Effect of broiler age, feed withdrawal, and transportation on levels of coliforms, *Campylobacter*, *Escherichia coli* and *Salmonella* on carcasses before and after immersion chilling.', *Poultry science*. England, 82(1), pp. 169–173. doi: 10.1093/ps/82.1.169.

O'Brien, S. J. (2013) 'The " Decline and Fall " of Nontyphoidal *Salmonella* in the United

Kingdom', *Clinical Infectious Diseases*, 56(5), pp. 705–710. doi: 10.1093/cid/cis967.

O'Brien, S. J. *et al.* (2016) 'Modelling study to estimate the health burden of foodborne diseases: cases, general practice consultations and hospitalisations in the UK, 2009.', *BMJ open*, 6(9). doi: 10.1136/bmjopen-2016-011119.

Ochman, H. and Groisman, E. A. (1996) 'Distribution of pathogenicity islands in *Salmonella* spp.', *Infection and Immunity*, 64(12). Available at: http://iai.asm.org/content/64/12/5410.abstract.

Olsen, J. E. *et al.* (1994) 'Stability of plasmids in five strains of *Salmonella* maintained in stab culture at different temperatures', *Journal of Applied Bacteriology*, 77(2), pp. 155–159. doi: 10.1111/j.1365-2672.1994.tb03059.x.

Ondov, B. D. *et al.* (2016) 'Mash: fast genome and metagenome distance estimation using MinHash', *Genome Biology*, 17(1). doi: 10.1186/s13059-016-0997-x.

Oren, Y. *et al.* (2014) 'Transfer of noncoding DNA drives regulatory rewiring in bacteria', *PNAS*, 111(45). doi: 10.1073/pnas.1413272111.

Oxford Nanopore Technologies (2018) *Rapid Barcoding Sequencing (SQK-RBK004)*. Available at: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=2ahUKEwjf7pbL8onjAhWRqHEKHUeSA7sQFjABegQIBRAC&url=https%3A%2F%2Fcommunity.nanoporetech.com%2Fprotocols%2Frapid-barcoding-sequencing-sqk-rbk004%2Fchecklist_example.pdf&usg=AOvVaw2b7NiOGClQ (Accessed: 9 July 2019).

Ozdemir, K. and Acar, S. (2014) 'Plasmid profile and pulsed-field gel electrophoresis analysis of *Salmonella enterica* isolates from humans in Turkey', *PLoS ONE*, 9(5), pp. 1–7. doi: 10.1371/journal.pone.0095976.

Page, A. J. *et al.* (2015) 'Sequence analysis Roary : rapid large-scale prokaryote pan genome analysis', *Bioinformatics*, 31(July), pp. 3691–3693. doi: 10.1093/bioinformatics/btv421.

Painter, J. A. *et al.* (2013) 'Attribution of Foodborne Illnesses, Hospitalizations, and Deaths to Food Commodities by using Outbreak Data, United States, 1998–2008', *Emerging Infectious Diseases*, 19(3), pp. 407–415. doi: 10.3201/eid1903.111866.

Palma, F. *et al.* (2018) 'Genome-wide identification of geographical segregated genetic markers in *Salmonella enterica* serovar Typhimurium variant 4,[5],12:i:-', *Scientific Reports*, 8(1), p. 15251. doi: 10.1038/s41598-018-33266-5.

Papadopoulos, T. *et al.* (2017) 'Multiple clones and low antimicrobial resistance rates for *Salmonella enterica* serovar Infantis populations in Greece', *Comparative Immunology,* *Microbiology and Infectious Diseases*. Elsevier Ltd, 51, pp. 54–58. doi: 10.1016/j.cimid.2017.02.002.

Paradis, E. and Schliep, K. (2018) ‘ape 5.0: an environment for modern phylogenetics and evolutionary analyses in {R}’, *Bioinformatics*.

Parry, C. M. *et al.* (2002) ‘Typhoid Fever’, *New England Journal of Medicine*. Massachusetts Medical Society, 347(22), pp. 1770–1782. doi: 10.1056/NEJMra020201.

Pate, M. *et al.* (2019) ‘*Salmonella* Infantis in Broiler Flocks in Slovenia : The Prevalence of Multidrug Resistant Strains with High Genetic Homogeneity and Low Biofilm-Forming Ability’, *BioMed Research International*, 2019.

Patil, S. R. *et al.* (2013) ‘Cellulitis Due to *Salmonella* Infantis’, *Online Journal of Health and Allied Sciences*, 11(4).

Pickering, A. J. *et al.* (2012) ‘Fecal Contamination and Diarrheal Pathogens on Surfaces and in Soils among Tanzanian Households with and without Improved Sanitation’, *Environmental Science & Technology*. American Chemical Society, 46(11), pp. 5736–5743. doi: 10.1021/es300022c.

Pires, S. M. *et al.* (2010) ‘Using Outbreak Data for Source Attribution of Human Salmonellosis and Campylobacteriosis in Europe’, *Foodborne Pathogens and Disease*, 7(11). doi: 10.1089/fpd.2010.0564.

Pires, S. M. *et al.* (2014) ‘Source attribution of human salmonellosis: an overview of methods and estimates.’, *Foodborne pathogens and disease*. United States, 11(9), pp. 667–676. doi: 10.1089/fpd.2014.1744.

Plaut, A. G. (2000) ‘Clinical Pathology of Foodborne Diseases: Notes on the Patient with Foodborne Gastrointestinal Illness’, *Journal of Food Protection*. International Association for Food Protection, 63(6), pp. 822–826. doi: 10.4315/0362-028X-63.6.822.

Ponstingl, H. and Ning, Z. (2014) *SMALT*. Available at: https://www.sanger.ac.uk/science/tools/smalt-0 (Accessed: 16 November 2018).

Porwollik, S. *et al.* (2004) ‘DNA Amplification and Rearrangements in Archival *Salmonella enterica* Serovar Typhimurium LT2 Cultures’, *Journal of Bacteriology*. American Society for Microbiology Journals, 186(6), pp. 1678–1682. doi: 10.1128/JB.186.6.1678-1682.2004.

Promega Corporation (2018) *QuantiFluor ® dsDNA System*. Available at: https://www.promega.co.uk/-/media/files/resources/protocols/technical-manuals/101/quantifluor-dsdna-system-protocol.pdf?la=en (Accessed: 9 July 2019).

Public Health England (2015) *Common gastrointestinal infections, England and Wales: laboratory reports weeks 49 to 52, 2014*. Available at:

https://www.gov.uk/government/publications/common-gastrointestinal-infections-in-england-and-wales-laboratory-reports-in-2014/common-gastrointestinal-infections-england-and-wales-laboratory-reports-weeks-49-to-52-2014 (Accessed: 11 June 2019).

PubMLST (2017) *Salmonella locus/sequence definitions database*. Available at: https://pubmlst.org/bigsdb?db=pubmlst_salmonella_seqdef (Accessed: 8 August 2017).

PubMLST (2018) *Salmonella locus/sequence definitions database*. Available at: https://pubmlst.org/bigsdb?db=pubmlst_salmonella_seqdef (Accessed: 31 October 2018).

Qiagen Sample & Assay Technologies (2016) 'QIAamp® DNA Mini and Blood Mini Handbook', (May). Available at: https://www.qiagen.com/gb/resources/download.aspx?id=62a200d6-faf4-469b-b50f-2b59cf738962&lang=en.

Qu, J., Huang, Y. and Lv, X. (2019) 'Crisis of Antimicrobial Resistance in China: Now and the Future', *Frontiers in Microbiology*, 10. doi: 10.3389/fmicb.2019.02240.

R Core Team (2018) 'R: A Language and Environment for Statistical Computing'. Available at: https://www.r-project.org/.

Raghavan, R. *et al.* (2015) 'Genome rearrangements can make and break small RNA genes', *Genome biology and evolution*. Oxford University Press, 7(2), pp. 557–566. doi: 10.1093/gbe/evv009.

Rahmani, M. *et al.* (2013) 'Molecular clonality and antimicrobial resistance in *Salmonella enterica* serovars Enteritidis and Infantis from broilers in three Northern regions of Iran.', *BMC veterinary research*, 9(1), p. 66. doi: 10.1186/1746-6148-9-66.

Ranjbar, R., Rahmati, H. and Shokoohizadeh, L. (2018) 'Detection of common clones of *Salmonella enterica* serotype infantis from human sources in tehran hospitals', *Gastroenterology and Hepatology from Bed to Bench*, 11(1), pp. 54–59. doi: 10.22037/ghfbb.v0i0.1202.

Rao, G. G. (1995) 'Control of outbreaks of viral diarrhoea in hospitals—a practical approach', *Journal of Hospital Infection*, 30(1), pp. 1–6. doi: https://doi.org/10.1016/0195-6701(95)90244-9.

Revell, L. J. (2012) 'phytools: an R package for phylogenetic comparative biology (and other things)', *Methods in Ecology and Evolution*. John Wiley & Sons, Ltd (10.1111), 3(2), pp. 217–223. doi: 10.1111/j.2041-210X.2011.00169.x.

Revolugen (2017) *PuriSpin Fire Monkey (100) kit*. Available at: http://revolugen.co.uk/wp-content/uploads/2018/02/Fire-Monkey-IFU-Feb18.pdf

(Accessed: 26 June 2019).

Ribeiro-Gonçalves, B. *et al.* (2016) 'PHYLOViZ Online: web-based tool for visualization, phylogenetic inference, analysis and sharing of minimum spanning trees', *Nucleic Acids Research*, 44(W1), pp. 246–251. doi: 10.1093/nar/gkw359.

Roth, N. *et al.* (2018) 'The application of antibiotics in broiler production and the resulting antibiotic resistance in Escherichia coli: A global overview', *Poultry Science*, 98(4), pp. 1791–1804. doi: 10.3382/ps/pey539.

Rowe, P. C. *et al.* (1993) 'Diarrhoea in close contacts as a risk factor for childhood haemolytic uraemic syndrome', *Epidemiology and Infection*. 2009/05/15. Cambridge University Press, 110(1), pp. 9–16. doi: 10.1017/S0950268800050627.

RStudio (2018) *RStudio*. Available at: https://www.rstudio.com (Accessed: 25 October 2018).

Rutherford, K. *et al.* (2000) 'Artemis: sequence visualization and annotation ', *Bioinformatics*, 16(10), pp. 944–945. doi: 10.1093/bioinformatics/16.10.944.

Ryan, M. P., Dwyer, J. O. and Adley, C. C. (2017) 'Evaluation of the Complex Nomenclature of the Clinically and Veterinary Significant Pathogen *Salmonella*'. Hindawi, 2017. doi:10.1155/2017/3782182.

Rychlik, I., Gregorova, D. and Hradecka, H. (2006) 'Distribution and function of plasmids in *Salmonella enterica*', *Veterinary Microbiology*, 112(1), pp. 1–10. doi: 10.1016/j.vetmic.2005.10.030.

Al Saif, N. and Brazier, J. S. (1996) 'The distribution of *Clostridium difficile* in the environment of South Wales', *Journal of Medical Microbiology*. Microbiology Society, 45(2), pp. 133–137. doi: https://doi.org/10.1099/00222615-45-2-133.

Salipante, S. J., Barlow, M. and Hall, B. G. (2003) 'GeneHunter , a Transposon Tool for Identification and Isolation of Cryptic Antibiotic Resistance Genes', *Antimicrobial Agents and Chemotherapy*, 47(12), pp. 3840–3845. doi: 10.1128/AAC.47.12.3840.

Sanderson, K. E. and Nair, S. (2013) 'Taxonomy and Species Concepts in the Genus *Salmonella*', in *Salmonella in Domestic Animals*, pp. 1–19.

Saroj, S. D. *et al.* (2008) 'Distribution of *Salmonella* pathogenicity island (SPI)-8 and SPI-10 among different serotypes of Salmonella.', *Journal of medical microbiology*. England, 57(Pt 4), pp. 424–427. doi: 10.1099/jmm.0.47630-0.

Sassetti, C. M., Boyd, D. H. and Rubin, E. J. (2001) 'Comprehensive identification of conditionally essential genes in mycobacteria', *Proceedings of the National Academy of Sciences*. National Academy of Sciences, 98(22), pp. 12712–12717. doi:

10.1073/pnas.231275498.

Scallan, E. *et al.* (2011) ‘Foodborne illness acquired in the United States--major pathogens’, *Emerging infectious diseases*. Centers for Disease Control and Prevention, 17(1), pp. 7–15. doi: 10.3201/eid1701.P11101.

Schmieger, H. and Schicklmaier, P. (1999) ‘Transduction of multiple drug resistance of *Salmonella enterica* serovar typhimurium DT104’, *FEMS Microbiology Letters*, 170(1), pp. 251–256. doi: 10.1111/j.1574-6968.1999.tb13381.x.

Schneitz, C. (2005) ‘Competitive exclusion in poultry - 30 years of research’, *Food Control*. Orion Corporation, 16(8 SPEC. ISS.), pp. 657–667. doi: 10.1016/j.foodcont.2004.06.002.

Schuster, C. J. *et al.* (2005) ‘Infectious Disease Outbreaks Related to Drinking Water in Canada, 1974–2001’, *Canadian Journal of Public Health*, 96(4), pp. 254–258. doi: 10.1007/BF03405157.

Seals, J. E. *et al.* (1983) ‘Nursery salmonellosis: delayed recognition due to unusually long incubation period.’, *Infection control : IC*. United States, 4(4), pp. 205–208. doi: 10.1017/s0195941700058239.

Seemann, T. (2014) ‘Prokka : rapid prokaryotic genome annotation’, *Bioinformatics*, 30(14), pp. 2068–2069. doi: 10.1093/bioinformatics/btu153.

Sehn, J. K. (2015) ‘Chapter 9 - Insertions and Deletions (Indels)’, in Kulkarni, S. and Pfeifer, J. (eds) *Clinical Genomics*. Boston: Academic Press, pp. 129–150. doi: https://doi.org/10.1016/B978-0-12-404748-8.00009-5.

Seth-Smith, H. M. B. *et al.* (2012) ‘Structure, Diversity, and Mobility of the *Salmonella* Pathogenicity Island 7 Family of Integrative and Conjugative Elements within Enterobacteriaceae’, *Journal of Bacteriology*. American Society for Microbiology Journals, 194(6), pp. 1494–1504. doi: 10.1128/JB.06403-11.

Setta, A. M. *et al.* (2012) ‘Early immune dynamics following infection with *Salmonella enterica* serovars Enteritidis, Infantis, Pullorum and Gallinarum: Cytokine and chemokine gene expression profile and cellular changes of chicken cecal tonsils’, *Comparative Immunology, Microbiology and Infectious Diseases*, 35(5), pp. 397–410. doi: https://doi.org/10.1016/j.cimid.2012.03.004.

Sever, N. K. and Akan, M. (2019) ‘Microbial Pathogenesis Molecular analysis of virulence genes of *Salmonella* Infantis isolated from chickens and turkeys’, *Microbial Pthogenesis*. Elsevier Ltd, 126(November 2018), pp. 199–204. doi: 10.1016/j.micpath.2018.11.006.

Shah, D. H. *et al.* (2005) ‘Identification of *Salmonella* Gallinarum virulence genes in a chicken infection model using PCR-based signature-tagged mutagenesis’, *Microbiology*.

Microbiology Society, 151(12), pp. 3957–3968. doi: https://doi.org/10.1099/mic.0.28126-0.

Shah, D. H. *et al.* (2012) ‘Transposon Mutagenesis of *Salmonella enterica* Serovar Enteritidis Identifies Genes That Contribute to Invasiveness in Human and Chicken Cells and Survival in Egg Albumen’, *Infection and Immunity*. Edited by A. J. Bäumler. American Society for Microbiology Journals, 80(12), pp. 4203–4215. doi: 10.1128/IAI.00790-12.

Shahada, F. *et al.* (2010) ‘Genetic analysis of multi-drug resistance and the clonal dissemination of β-lactam resistance in *Salmonella* Infantis isolated from broilers’, *Veterinary Microbiology*, 140(1–2), pp. 136–141. doi: 10.1016/j.vetmic.2009.07.007.

Shahada, F., Chuma, T. and Dahshan, H. (2010) ‘Detection and Characterization of Extended-Spectrum b-Lactamase (TEM-52)-Producing *Salmonella* Serotype Infantis from Broilers in Japan’, *Foodborne Pathogens and Disease*, 7(5).

Shimi, A. and Barin, A. (1977) ‘*Salmonella* in cats’, *Journal of Comparative Pathology*, 87(2), pp. 315–318. doi: https://doi.org/10.1016/0021-9975(77)90020-2.

Shintani, M., Sanchez, Z. K. and Kimbara, K. (2015) ‘Genomics of microbial plasmids : classification and identification based on replication and transfer systems and host taxonomy’, *Frontiers in Microbiology*, 6, pp. 1–16. doi: 10.3389/fmicb.2015.00242.

Shivaprasad, H. L. (2000) ‘Fowl typhoid and pullorum disease’, *Revue Scientifique et Technique-Office International des Epizooties*. OIE Office International Des Epizooties, 19(2), pp. 405–416.

Silva, C., Calva, E. and Maloy, S. (2014) ‘One Health and Food-Borne Disease: *Salmonella* Transmission between Humans, Animals, and Plants.’, *Microbiology spectrum*, 2(1). doi: 10.1128/microbiolspec.OH-0020-2013.

Silva, C., Puente, J. L. and Calva, E. (2017) ‘*Salmonella* virulence plasmid: pathogenesis and ecology.’, *Pathogens and disease*. United States. doi: 10.1093/femspd/ftx070.

Silva, C., Wiesner, M. and Calva, E. (2012) ‘The Importance of Mobile Genetic Elements in the Evolution of *Salmonella*: Pathogenesis, Antibiotic Resistance and Host Adaptation’, in Kumar, Y. (ed.) *Salmonella*. Rijeka: IntechOpen. doi: 10.5772/28024.

Smith, A. *et al.* (2006) ‘Outbreaks of waterborne infectious intestinal disease in England and Wales, 1992–2003’, *Epidemiology and Infection*. 2006/05/11. Cambridge University Press, 134(6), pp. 1141–1149. doi: DOI: 10.1017/S0950268806006406.

Snel, S. J. *et al.* (2009) ‘A tale of two parasites: the comparative epidemiology of cryptosporidiosis and giardiasis’, *Epidemiology and Infection*. 2009/04/27. Cambridge University Press, 137(11), pp. 1641–1650. doi: DOI: 10.1017/S0950268809002465.

Somvanshi, V. S. *et al.* (2012) ‘A single promoter inversion switches Photorhabdus between pathogenic and mutualistic states.’, *Science (New York, N.Y.)*. United States, 337(6090), pp. 88–93. doi: 10.1126/science.1216641.

Stamatakis, A. (2014) ‘RAxML version 8 : a tool for phylogenetic analysis and post-analysis of large phylogenies’, 30(9), pp. 1312–1313. doi: 10.1093/bioinformatics/btu033.

Stanaway, J. D., Parisi, A., *et al.* (2019) ‘The global burden of non-typhoidal *Salmonella* invasive disease: a systematic analysis for the Global Burden of Disease Study 2017’, *The Lancet Infectious Diseases*. doi: https://doi.org/10.1016/S1473-3099(19)30418-9.

Stanaway, J. D., Reiner, R. C., *et al.* (2019) ‘The global burden of typhoid and paratyphoid fevers: a systematic analysis for the Global Burden of Disease Study 2017’, *The Lancet Infectious Diseases*. Elsevier, 19(4), pp. 369–381. doi: 10.1016/S1473-3099(18)30685-6.

Stangroom, J. (2019) *Z Score Calculator for 2 Population Proportions*. Available at: https://www.socscistatistics.com/tests/ztest/ (Accessed: 1 November 2019).

Statistics South Africa (2018) *Statistical release- Mid-year population estimates July 2018*. Available at: http://www.statssa.gov.za/publications/P0302/P03022011.pdf (Accessed: 22 March 2019).

Stuart, B. M. and Pullen, R. L. (1946) ‘Typhoid; clinical analysis of 360 cases.’, *Archives of internal medicine (Chicago, Ill. : 1908)*. United States, 78(6), pp. 629–661. doi: 10.1001/archinte.1946.00220060002001.

Su, L. *et al.* (2004) ‘Antimicrobial Resistance in Nontyphoid Salmonella Serotypes : A Global Challenge’, *Clinical Infectious Diseases*, 39.

Svedhem and Kaijser, B. (1981) ‘Isolation of *Campylobacter jejuni* from domestic animals and pets: probable orgin of human infection’, *Journal of Infection*. Elsevier, 3(1), pp. 37–40. doi: 10.1016/S0163-4453(81)92261-1.

Szmolka, A. *et al.* (2018) ‘Molecular epidemiology of the endemic multiresistance plasmid pSI54/04 of *Salmonella* Infantis in broiler and human population in Hungary’, *Food Microbiology*. Elsevier Ltd, 71, pp. 25–31. doi: 10.1016/j.fm.2017.03.011.

Tacconelli, E. *et al.* (2018) ‘Articles Discovery, research, and development of new antibiotics : the WHO priority list of antibiotic-resistant bacteria and tuberculosis’, *The Lancet Infectious Diseases*, 3099(17), pp. 1–10. doi: 10.1016/S1473-3099(17)30753-3.

Tack, D. M. *et al.* (2019) ‘Preliminary Incidence and Trends of Infections with Pathogens Transmitted Commonly Through Food - Foodborne Diseases Active Surveillance Network, 10 U.S. Sites, 2015-2018’, *MMWR. Morbidity and mortality weekly report*. Centers for Disease Control and Prevention, 68(16), pp. 369–373. doi: 10.15585/mmwr.mm6816a2.

Tam, C. C. *et al.* (2012) ‘Longitudinal study of infectious intestinal disease in the UK (IID2 study): incidence in the community and presenting to general practice’, pp. 69–78. doi: 10.1136/gut.2011.238386.

Tanner, J. R. and Kingsley, R. A. (2018) ‘Evolution of *Salmonella* within Hosts’, *Trends in Microbiology*. Elsevier Ltd, 26(12), pp. 986–998. doi: 10.1016/j.tim.2018.06.001.

Tanny, R. (2014) *Quantification of DNA*. Available at: http://andersenlab.org/Protocols/96-wellDNAQuantification_Synergy4_Qubit_v2.pdf (Accessed: 20 August 2018).

Tate, H. *et al.* (2017) ‘Comparative Analysis of Extended-Spectrum-beta-Lactamase CTX-M-65-Producing *Salmonella enterica* Serovar Infantis Isolates from Humans, Food Animals, and Retail Chickens in the United States.’, *Antimicrobial agents and chemotherapy*. United States, 61(7). doi: 10.1128/AAC.00488-17.

Tewolde, R., Dallman, T., Schaefer, U., Sheppard, Carmen L., *et al.* (2016) ‘MOST: a modified MLST typing tool based on short read sequencing’, *PeerJ*, 4. doi:10.7717/peerj.2308.

Thankaswamy-Kosalai, S., Sen, P. and Nookaew, I. (2017) ‘Evaluation and assessment of read-mapping by multiple next-generation sequencing aligners based on genome-wide characteristics’, *Genomics*, 109(3), pp. 186–191. doi:https://doi.org/10.1016/j.ygeno.2017.03.001.

Thermo Fisher (2015a) *Quant-iT^{TM} dsDNA Broad-Range Assay Kit*. Available at: https://www.thermofisher.com/document-connect/document-connect.html?url=https%3A%2F%2Fassets.thermofisher.com%2FTFS-Assets%2FLSG%2Fmanuals%2FQuant_iT_dsDNA_BR_Assay_UG.pdf&title=VXNlciBHdWlkZTogUXVhbnQtaVQgZHNETkEgQnJvYWQtUmFuZ2UgQXNzYXkgS2l0 (Accessed: 26 June 2019).

Thermo Fisher (2015b) *Quant-iT^{TM} dsDNA High-Sensitivity Assay Kit*. Available at: Quant-iT^{TM} dsDNA High-Sensitivity Assay Kit (Accessed: 29 July 2019).

Thermo Fisher (2015c) *Qubit® dsDNA BR Assay Kits*. Available at: http://tools.thermofisher.com/content/sfs/manuals/Qubit_dsDNA_BR_Assay_UG.pdf (Accessed: 26 June 2019).

Thermo Fisher (2015d) *Qubit® dsDNA HS Assay Kits*. Available at: https://www.thermofisher.com/document-connect/document-connect.html?url=https%3A%2F%2Fassets.thermofisher.com%2FTFS-Assets%2FLSG%2Fmanuals%2FQubit_dsDNA_HS_Assay_UG.pdf&title=VXNlciBHdWlkZTog

UXViaXQgZHNETkEgSFMgQXNzYXkgS2l0cw== (Accessed: 26 June 2019).

Thomas, M. K. *et al.* (2015) 'Estimates of Foodborne Illness–Related Hospitalizations and Deaths in Canada for 30 Specified Pathogens and Unspecified Agents', *Foodborne Pathogens and Disease*. Mary Ann Liebert, Inc., publishers, 12(10), pp. 820–827. doi:10.1089/fpd.2015.1966.

Thomason, B. M., Biddle, J. W. and Cherry, W. B. (1975) 'Dection of salmonellae in the environment.', *Applied microbiology*. United States, 30(5), pp. 764–767.

Thomsen, L. E. *et al.* (2003) 'Reduced amounts of LPS affect both stress tolerance and virulence of *Salmonella enterica* serovar Dublin', *FEMS Microbiology Letters*, 228(2), pp. 225–231. doi: 10.1016/S0378-1097(03)00762-6.

Thong, K. L. *et al.* (2015) 'Quinolone Resistance Mechanisms Among *Salmonella enterica* in Malaysia', *Microbial Drug Resistance*. Mary Ann Liebert, Inc., publishers, 22(4), pp. 259–272. doi: 10.1089/mdr.2015.0158.

Thornbrough, J. M. and Worley, M. J. (2012) 'A Naturally Occurring Single Nucleotide Polymorphism in the *Salmonella* SPI-2 Type III Effector srfH/sseI Controls Early Extraintestinal Dissemination', *PLOS ONE*. Public Library of Science, 7(9). Available at: https://doi.org/10.1371/journal.pone.0045245.

Thorpe, H. A. *et al.* (2017) 'Comparative Analyses of Selection Operating on Nontranslated Intergenic Regions of Diverse Bacterial Species', *Genetics*, 206(1), pp. 363–376. doi: 10.1534/genetics.116.195784.

Thorpe, H. A. *et al.* (2018) 'Piggy: a rapid, large-scale pan-genome analysis tool for intergenic regions in bacteria', *GigaScience*, 7(4). doi: 10.1093/gigascience/giy015.

Tizard, I. (2004) 'Salmonellosis in Wild Birds', *Seminars in Avian and Exotic Pet Medicine*, 13(2), pp. 50–66. doi: 10.1053/j.saep.2004.01.008.

Tonkin-Hill, G. *et al.* (2018) 'RhierBAPS: An R implementation of the population clustering algorithm hierBAPS', *Wellcome open research*. F1000 Research Limited, 3, p. 93. doi: 10.12688/wellcomeopenres.14694.1.

Tonkin-Hill, G. *et al.* (2019) 'Fast hierarchical Bayesian analysis of population structure', *Nucleic acids research*, 47(11), pp. 5539–5549. doi: 10.1093/nar/gkz361.

Troeger, C. E. *et al.* (2018) 'Estimates of the global, regional, and national morbidity, mortality, and aetiologies of diarrhoea in 195 countries : a systematic analysis for the Global Burden of Disease Study 2016', pp. 1211–1228. doi: 10.1016/S1473-3099(18)30362-1.

U.S. Food & Drug Administration (2019) *Outbreak Investigation of Salmonella Infantis*

Linked to Del Monte Fresh Produce Vegetable Trays, Spring 2019. Available at: https://www.fda.gov/food/outbreaks-foodborne-illness/outbreak-investigation-salmonella-infantis-linked-del-monte-fresh-produce-vegetable-trays-spring (Accessed: 20 August 2019).

Uche, I. V., MacLennan, C. A. and Saul, A. (2017) ‘A Systematic Review of the Incidence, Risk Factors and Case Fatality Rates of Invasive Nontyphoidal *Salmonella* (iNTS) Disease in Africa (1966 to 2014)’, *PLOS Neglected Tropical Diseases*. Public Library of Science, 11(1). Available at: https://doi.org/10.1371/journal.pntd.0005118.

UniProt (2019a) *UniProtKB - A0A0D6IR70 (A0A0D6IR70_SALTM)*. Available at: https://www.uniprot.org/uniprot/A0A0D6IR70 (Accessed: 6 November 2019).

UniProt (2019b) *UniProtKB - P0A1H5 (EFTU_SALTY)*. Available at: https://www.uniprot.org/uniprot/P0A1H5 (Accessed: 5 November 2019).

UniProt (2019c) *UniProtKB - P67388 (UPPP_SALTY)*. Available at: https://www.uniprot.org/uniprot/P67388 (Accessed: 8 November 2019).

Vaser, R. *et al.* (2017) ‘Fast and accurate de novo genome assembly from long uncorrected reads.’, *Genome research*. United States, 27(5), pp. 737–746. doi: 10.1101/gr.214270.116.

Velhner, M. *et al.* (2014) ‘Clonal spread of *Salmonella enterica* serovar infantis in Serbia: Acquisition of mutations in the topoisomerase genes gyrA and parC leads to increased resistance to fluoroquinolones’, *Zoonoses and Public Health*, 61(5), pp. 364–370. doi: 10.1111/zph.12081.

Vetting, M, W. *et al.* (2004) ‘A Bacterial Acetyltransferase Capable of Regioselective N-Acetylation of Antibiotics and Histones’, *Chemistry & BIology*, 11, pp. 565–573. doi: 10.1016/j.

Vila Nova, M. *et al.* (2019) ‘Genetic and metabolic signatures of *Salmonella enterica* subsp. *enterica* associated with animal sources at the pangenomic scale’, *BMC Genomics*, 20(1), p. 814. doi: 10.1186/s12864-019-6188-x.

Vinueza-Burgos, C. *et al.* (2016) ‘Prevalence and Diversity of *Salmonella* Serotypes in Ecuadorian Broilers at Slaughter Age’, *PLOS ONE*. Public Library of Science, 11(7). Available at: https://doi.org/10.1371/journal.pone.0159567.

Vohra, P. *et al.* (2019) ‘Retrospective application of transposon-directed insertion-site sequencing to investigate niche-specific virulence of *Salmonella* Typhimurium in cattle’, *BMC Genomics*, 20(1), p. 20. doi: 10.1186/s12864-018-5319-0.

Vos, T. *et al.* (2016) ‘Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015', *The Lancet*, 388(10053), pp. 1545–1602. doi:https://doi.org/10.1016/S0140-6736(16)31678-6.

Voss-Rech, D. *et al.* (2015) 'A temporal study of *Salmonella enterica* serotypes from broiler farms in Brazil', *Poultry Science*, 94(3), pp. 433–441. doi: 10.3382/ps/peu081.

Wajid, M. *et al.* (2019) 'Detection and characterization of multidrug-resistant *Salmonella enterica* serovar Infantis as an emerging threat in poultry farms of Faisalabad, Pakistan', *Journal of Applied Microbiology*, 127(1). doi: 10.1111/jam.14282.

Wales, A. D. *et al.* (2010) 'Review of the Carriage of Zoonotic Bacteria by Arthropods, with Special Reference to *Salmonella* in Mites, Flies and Litter Beetles', *Zoonoses and Public Health*. John Wiley & Sons, Ltd (10.1111), 57(5), pp. 299–314. doi: 10.1111/j.1863-2378.2008.01222.x.

Walker, B. J. *et al.* (2014) 'Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement', *PLOS ONE*. Public Library of Science, 9(11), pp. 1–14. doi: 10.1371/journal.pone.0112963.

Walle, I. Van *et al.* (2019) 'EFSA and ECDC technical report on the collection and analysis of whole genome sequencing data from food-borne pathogens and other relevant microorganisms isolated from human , animal , food , feed and food / feed environmental samples in the joint ECDC – EFSA molecular typing database', (April). doi:10.2903/sp.efsa.2019.EN-1337.

Webb-Johnson, A. . (1917) 'Hunterian Lecture on the surgical compilcations of typhoid and paratyhpoid fevers', *The Lancet*, 190(4918), pp. 813–821. doi:https://doi.org/10.1016/S0140-6736(01)56859-2.

Webster, A. B. and Fletcher, D. L. (1996) 'Humane On-Farm Killing Of Spent Hens', *Applied Poultry Science*, (5), pp. 191–200.

Wheeler, N. E., Gardner, P. P. and Barquist, L. (2018) 'Machine learning identifies signatures of host adaptation in the bacterial pathogen Salmonella enterica', *PLOS Genetics*. Public Library of Science, 14(5). Available at: https://doi.org/10.1371/journal.pgen.1007333.

Wick, R. R. *et al.* (2015) 'Bandage: interactive visualization of de novo genome assemblies', *Bioinformatics*, 31(20), pp. 3350–3352. doi: 10.1093/bioinformatics/btv383.

Wick, R. R. *et al.* (2017) 'Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads', *PLOS Computational Biology*. Public Library of Science, 13(6). Available at: https://doi.org/10.1371/journal.pcbi.1005595.

Wickham, H. (2016) ‘ggplot2: Elegant Graphics for Data Analysis’. Springer-Verlag New York. Available at: https://ggplot2.tidyverse.org.

Wickham, H. *et al.* (2018) ‘devtools: Tools to Make Developing R Packages Easier’. Available at: https://cran.r-project.org/web/packages/devtools/index.html.

Williams, L. E. *et al.* (2013) ‘Large plasmids of *Escherichia coli* and *Salmonella* encode highly diverse arrays of accessory genes on common replicon families’, *Plasmid*. Elsevier Inc., 69(1), pp. 36–48. doi: 10.1016/j.plasmid.2012.08.002.

Williamson, D. A. *et al.* (2018) ‘Increasing antimicrobial resistance in nontyphoidal *Salmonella* isolates in Australia from 1979 to 2015’, *Antimicrobial Agents and Chemotherapy*, 62(2), pp. 1–9. doi: 10.1128/AAC.02012-17.

Winter, S. E. *et al.* (2010) ‘Gut inflammation provides a respiratory electron acceptor for *Salmonella*’, *Nature*, 467(7314), pp. 426–429. doi: 10.1038/nature09415.

Zhang, S. *et al.* (2019) ‘Zoonotic Source Attribution of *Salmonella enterica* Serotype Typhimurium Using Genomic Surveillance Data, United States’, *Emerging infectious diseases*. Centers for Disease Control and Prevention, 25(1), pp. 82–91. doi: 10.3201/eid2501.180835.

Withanage, G. S. K. *et al.* (2005) ‘Cytokine and Chemokine Responses Associated with Clearance of a Primary *Salmonella enterica* Serovar Typhimurium Infection in the Chicken and in Protective Immunity to Rechallenge’, *Infection and Immunity*, 73(8), pp. 5173–5182. doi: 10.1128/IAI.73.8.5173-5182.2005.

World Health Organization (2015) ‘Food safety’. Available at: https://www.who.int/en/news-room/fact-sheets/detail/food-safety (Accessed: 18 November 2019).

World Health Organization (2016) *Salmonella (non-typhoidal)*. Available at: http://www.who.int/mediacentre/factsheets/fs139/en/ (Accessed: 18 November 2019).

Worley, J. *et al.* (2018) ‘*Salmonella enterica* Phylogeny Based on Whole-Genome Sequencing Reveals Two New Clades and Novel Patterns of Horizontally Acquired Elements’, *Ecological and Evolutionary Science*, 9(6).

Yokoyama, E. *et al.* (2014) ‘Infection, Genetics and Evolution Phylogenetic and population genetic analysis of *Salmonella enterica* subsp. *enterica* serovar Infantis strains isolated in Japan using whole genome sequence data’, *Infection, Genetics and Evolution*. Elsevier B.V., 27, pp. 62–68. doi: 10.1016/j.meegid.2014.06.012.

Yokoyama, E. *et al.* (2015) ‘A novel subpopulation of *Salmonella enterica* serovar Infantis strains isolated from broiler chicken organs other than the gastrointestinal tract’, *Veterinary Microbiology*, 175(2), pp. 312–318. doi:https://doi.org/10.1016/j.vetmic.2014.11.024.

Zambrano, L. D. *et al.* (2014) ‘Human diarrhea infections associated with domestic animal husbandry: a systematic review and meta-analysis’, *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 108(6), pp. 313–325. doi: 10.1093/trstmh/tru056.

Zankari, E. *et al.* (2012) ‘Identification of acquired antimicrobial resistance genes’, *Journal of Antimicrobial Chemotherapy*, 67(11), pp. 2640–2644. doi: 10.1093/jac/dks261.

Zhang, L. *et al.* (2018) ‘Highly Prevalent Multidrug-Resistant *Salmonella* From Chicken and Pork Meat at Retail Markets in Guangdong, China’, *Frontiers in Microbiology*, p. 2104. Available at: https://www.frontiersin.org/article/10.3389/fmicb.2018.02104.

Zhang, S. *et al.* (2019) ‘Zoonotic Source Attribution of Salmonella enterica Serotype Typhimurium Using Genomic Surveillance Data, United States’, *Emerging infectious diseases*. Centers for Disease Control and Prevention, 25(1), pp. 82–91. doi: 10.3201/eid2501.180835.

Zollner-Schwetz, I. and Krause, R. (2015) ‘Therapy of acute gastroenteritis: role of antibiotics’, *Clinical Microbiology and Infection*, 21(8), pp. 744–749. doi:https://doi.org/10.1016/j.cmi.2015.03.002.